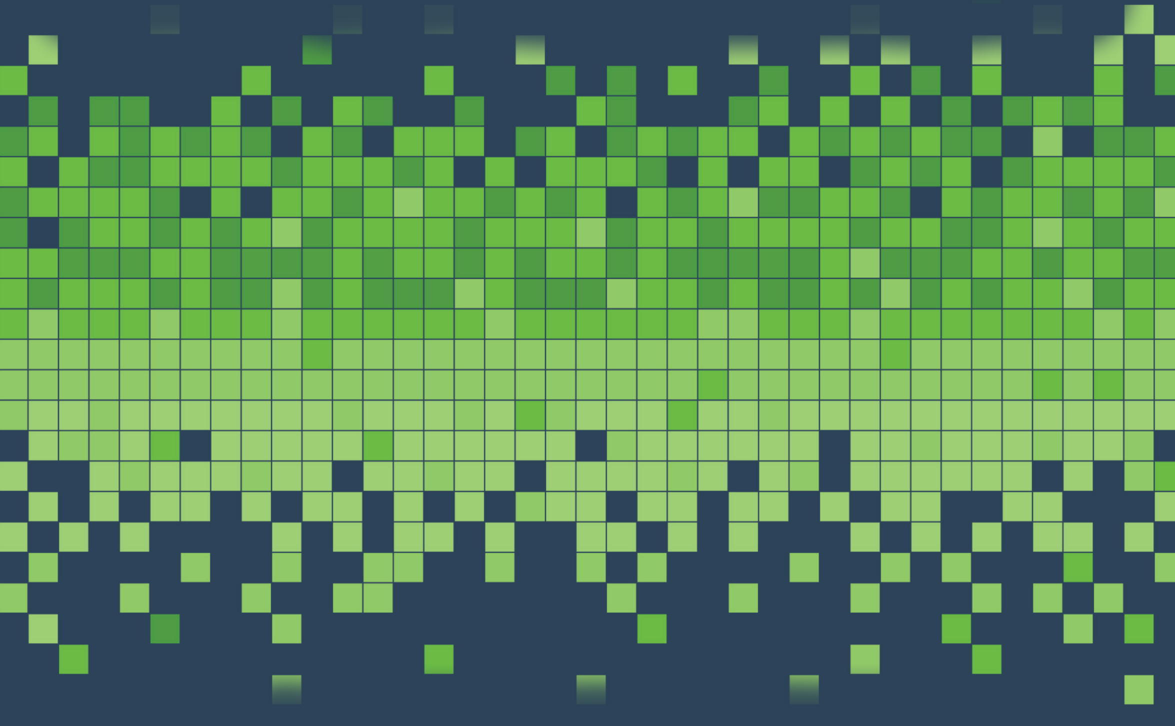
COMPUTATIONAL METHODS FOR DATA EVALUATION AND ASSIMILATION



DAN GABRIEL CACUCI
IONEL MICHAEL NAVON
MIHAELA IONESCU-BUJOR

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

COMPUTATIONAL METHODS
FOR
DATA EVALUATION AND ASSIMILATION

COMPUTATIONAL METHODS FOR DATA EVALUATION AND ASSIMILATION

DAN GABRIEL CACUCI
IONEL MICHAEL NAVON
MIHAELA IONESCU-BUJOR

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20130701

International Standard Book Number-13: 978-1-58488-736-2 (eBook - PDF)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contributors

Dan Gabriel Cacuci
University of South Carolina
Columbia, South Carolina

Ionel Michael Navon
Florida State University
Tallahassee, Florida

Mihaela Ionescu-Bujor
Karlsruher Institute of Technology
Karlsruhe, Campus North

Preface

This book is addressed to graduate and postgraduate students and researchers in the interdisciplinary methods of data assimilation, which refers to the integration of experimental and computational information. Since experiments and corresponding computations are encountered in many fields of scientific and engineering endeavors, the concepts presented in this book are illustrated using paradigm examples that range from the geophysical sciences to nuclear physics. In an attempt to keep the book as self-contained as possible, the mathematical concepts mostly from probability theory and functional analysis needed to follow the material presented in the book's five chapters, are summarized in the book's three appendices.

This book was finalized at the University of South Carolina. The authors wish to acknowledge the outstanding professional assistance of Dr. Madalina Corina Badea of the University of South Carolina, who has thoroughly reviewed the final version of the book, providing very valuable suggestions while improving its readability. Also acknowledged are the services of Dr. Erkan Arslan for his typing the word-version of this book into Latex. Last but not least, this book would have not have appeared without the continued patience, guidance, and understanding of Bob Stern (Executive Editor, Taylor and Francis Group), whom the authors appreciate immensely.

List of Figures

List of Tables

Contents

Introduction

Experience shows that it is practically impossible to measure exactly the true value of a physical quantity. This is because various imperfections occur at various stages involved in a measurement, including uncontrollable experimental errors, inaccurate standards, and other uncertainties arising in the data measurement and interpretation (reduction) process. Around any reported experimental value, therefore, there always exists a certain range of similar, more or less plausible, values that may also be true. In turn, this means that all inferences, predictions, engineering computations, and other applications of measured data are necessarily founded on weighted averages over all the possibly true values, with weights indicating the degree of plausibility of each value. These weights and weighted averages are what we call *probabilities* and *expectation values*. Consequently, the evaluation of scientific data is intrinsically intertwined with probability theory. The basic concepts underlying the evaluation of experimental data are presented in Chapter 1, which commences with a discussion, in Section 1.1, of the basic types of errors and probability distributions usually associated with them.

Section 1.2 presents the basic concepts of probability theory involved in the evaluation of uncertainty-afflicted scientific data. Since probabilities cannot be measured directly, they are either inferred from the results of observations or they are postulated and (partially) verified through accumulated experience. In scientific data evaluation, probabilities encode incomplete information. Persons possessing different information or knowledge assign different probabilities; furthermore, these probabilities are updated whenever new relevant information becomes available. It follows that probabilities cannot be considered as measurable physical properties. They are subjective in the sense that they depend on a person's knowledge. However, this does not mean that probabilities are arbitrary. They must obey the rules of logic, which demand, for instance, that rational persons with the same knowledge assign the same probabilities. The elementary conditions for logical consistency imply two fundamental rules from which all other mathematical relationships between probabilities can be derived. These two fundamental rules are the sum and the product rules, respectively. A powerful logical tool that follows immediately from the two forms of the product rule is Bayes' theorem (1763). The customary way to express Bayes' theorem in words is: "the posterior is proportional to the likelihood times the prior" where the "prior" distribution summarizes our knowledge extant prior to observing the (new) data, the "likelihood" distribution conveys the impact of the new information brought by the (new) data, while the "posterior" distribution contains the full information available for further inferences, predictions, and decision making. Thus, Bayes' theorem is a formal model for updating, or learning from observations.

A "measurable" or "physical" quantity is a property of phenomena, bodies, or substances that can be determined qualitatively and can be expressed quantitatively. Measurement is the process of experimentally finding the value of a physical quantity, with the help of devices called *measuring instruments*. It is important to note that: (1) the purpose of a measurement is to represent a property of an object by a number, so the result of a measurement must always be a number expressed in sanctioned units of measurements; (2) a measurement is always performed with the help of a measuring instrument; and (3) a measurement is always an experimental procedure. If it were known, the true value of a measurable quantity would ideally reflect, both qualitatively and quantitatively, the corresponding property of the object. The theory of measurement relies on the following postulates: (a) the true value of the measurable quantity exists; (b) the true value of the measurable quantity is constant relative to the conditions of the measurement; and (c) the true value cannot be found.

Since measuring instruments are imperfect, and since every measurement is an experimental procedure, the results of measurements cannot be accurate. This unavoidable imperfection of measurements is quantitatively characterized by measurement uncertainty or measurement error, which can be expressed in absolute or relative form. Consequently, repeated measurements of the same physical quantity can never yield identical results; even the most carefully measured scientific data in data banks will inevitably differ from the true values of the measured quantities. Consequently, nominal values for data, by themselves, are insufficient for applications. Quantitative uncertainties are also needed, along with the respective nominal values. Since the use of uncertain data may necessitate costly safety margins (in medicine, weather and climate prediction, or in the chemical, automotive, aerospace, or nuclear industries), working groups of the International Standards Organization have been developing uniform rules for reporting data uncertainties.

Combination of data from different sources involves a weighted propagation (via sensitivities, as will be seen subsequently) of all input uncertainties to uncertainties in the output values. Hence, data evaluation is intrinsically intertwined with uncertainty analysis, requiring reasoning from incomplete information, using probability theory for extracting "best estimate" values together with "best estimate" uncertainties from often sparse, incomplete, error-afflicted, and occasionally discrepant experimental data. A wide range of probability theory concepts and tools are employed in data evaluation and combination, from deductive statistics involving mainly frequencies and sample tallies to inductive inference for assimilating non-frequency data and a *priori knowledge.* Although grossly erroneous procedures and unintended mistakes (e.g., overlooking or miscalculating important corrections, equipment failure or improper calibration, bugs in computer codes, etc.) can produce defective data, such defective data will not be treated as "uncertainties" in this book. Nevertheless, data points that exhibit atypical behavior, which cannot be explained, need to be carefully scrutinized since outright rejection may not necessarily be appropriate. The terms "error" and "uncertainty" are interpreted in this book as being equivalent to the standard deviation of the probability distribution associated with the measurement process. This interpretation is consistent with the usual convention of considering "error" or "uncertainty" as an inherently positive number that quantifies the measurement dispersion of a specific observable parameter.

Legitimate errors are categorized either as random errors or as systematic errors. If the results of separate measurements of the same quantity differ from one another, and the respective differences cannot be predicted individually, then the error owing to this scatter of the results is called *random error.* Random errors can be identified by repeatedly measuring the same quantity under the same conditions. The scatter in results cannot be always tested in practice, particularly in large-scale modern experiments, where it may be impractical to provide sufficient repetition in order to satisfy the explicit needs for quantifying the random errors based on strict statistical requirements. Nevertheless, reasonable estimates of random errors can often be made, particularly when the nature of the underlying probability distribution can be inferred from previous experience. Furthermore, due to the influence of the central limit theorem, many sources of random error tend to be normally distributed. A significant feature of random errors is that repeated measurements (under fixed conditions) not only permit these errors to be better determined, but they also lead to error reduction, as assured by the law of large numbers. This feature is particularly important when high precision (i.e., small random errors) is required. A subtle issue regarding random errors stems, from the fact that such errors may contain correlated components: whether an error component is correlated or not within a particular data set depends upon the role that the associated random variable plays in the respective physical problem.

In contradistinction to a random error, a systematic error is defined as a measurement error that remains constant or changes in a regular fashion when the measurements of that quantity are repeated. Such errors arise because of inherent laws in the investigative process itself, and they lead to bias. Although systematic errors are difficult to distinguish from blunders, particularly when the impact of a blunder is small, the most reliable way to uncover systematic errors is by using a more accurate measuring instrument and/or by comparing a given result with a measurement of the same quantity, but performed by a different method. Each distinct approach leads to results that differ somewhat from those obtained in other ways. These differences exhibit a pattern (i.e., are "systematic") no matter how many times each approach is repeated, because the inherent systematic deficiencies of each method cannot be avoided by mere repetition. When the errors are truly systematic, statistical regularity will emerge from the ensemble of all measurements. Such a statistical regularity will not emerge when the data sets are afflicted with blunders, since blunders are generally one-time occurrences that can be detected if a particular procedure is repeated. Consequently, redundancy within a given investigative procedure is desirable not only to improve precision, but also to purge the results of blunders.

Probability theory is a branch of mathematical sciences that provides a model for describing the process of observation. The need for probability theory arises from the fact that most observations of natural phenomena do not lead to uniquely predictable results. Probability theory provides the tools for dealing with actual variations in the outcome of realistic observations and measurements. The challenging pursuit to develop a probability theory which is mathematically rigorous and also describes many phenomena observable in nature has generated over the years notable disputes over conceptual and logical issues. Modern probability theory is based on postulates constructed from three axioms attributed to Kolmogorov (1933), all of which are consistent with the notion of frequency of occurrence of events. The alternative approach, traceable to Laplace (1812), is based on the concept that probability is simply a way of providing a numerical scale to quantify our reasonable beliefs about a situation which we know only incompletely; this approach is consistent with Bayes' theorem, conditional probabilities, and inductive reasoning. Either approach to probability theory would completely describe a natural phenomenon if sufficient information were available to determine the underlying probability distribution exactly. In practice, though, such exact knowledge is seldom, if ever, available so the features of the probability distribution underlying the physical phenomenon under consideration must be estimated. Such estimations form the study object of statistics, which is defined as the branch of mathematical sciences that uses the results of observations and measurements to estimate, in a mathematically well-defined manner, the essential features of probability distributions. Both statistics and probability theory use certain generic terms for defining the objects or phenomena under study. A *system* is the object or phenomena under study. It represents the largest unit being considered. A system can refer to a nuclear reactor, corporation, chemical process, mechanical device, biological mechanism, society, economy, or any other conceivable object that is under study. The output or *response* of a system is a result that can be measured quantitatively or enumerated. The power of a nuclear reactor, the yield of a process, life span of cell, the atmospheric temperature and pressure, are all examples of system responses. A *model* is a mathematical idealization that is used as an approximation to represent the system and its output. Models can be quite simple or highly complex; regardless of its complexity, though, the model is an idealization of the system, so it cannot be exact. Usually, the more complex the system, the less exact the model, particularly since the ability to solve exactly mathematically highly complex expressions diminishes with increasing complexity. In other words, the simpler the model, the easier it is to analyze but the less precise the results.

Probabilities cannot be measured directly; they can be inferred from the results of observations or they can be postulated and (partially) verified through accumulated experience. In practice, though, certain random vectors tend to be more probable, so that most probability functions of practical interest tend to be localized. Therefore, the essential features regarding probability distributions of practical interest are measures of location and of dispersion of the observed results. Practice indicates that location is best described by the mean value, while dispersion of observed results appears to be best described by the variance, which is a second-order moment. In particular, the mean value can be interpreted as a locator of the center of gravity, while the variance is analogous to the moment of inertia (which linearly relates applied torque to induced angular acceleration in mechanics). For multivariate probability distributions, the collection of all second-order moments forms the so-called variance-covariance matrix, or, simply, the covariance matrix. If the probability function is known, then these moments can be calculated directly, through a process called *statistical deduction*. Otherwise, if the probability function is not known, then the respective moments must be estimated from experiments, through a process called *statistical inference*. The definitions, interpretations, and quantifications of the moments of a distribution, particularly the means and covariances, are discussed in Section 1.3. Particularly important is the method of propagation of errors or propagation of moments, which can be used to compute the error in a systems response (which can be either the result of an indirect measurement or the result of a computation), by propagating the uncertainties of the component system parameters using a form Taylor-series expansion of the response as a function of the underlying model parameters.

In practice, users of measured data seldom require knowledge of the complete posterior distribution, but usually request a “recommended value” for the respective quantity, accompanied by “error bars” or some suitably equivalent summary of the posterior distribution. Decision theory can provide such a summary, since it describes the penalty for bad estimates by a loss function. Since the true value is never known in practice, it is not possible to avoid a loss completely, but it is possible to minimize the expected loss, which is what an optimal estimate must accomplish. As will be shown in the first section of Chapter 2, in the practically most important case of "quadratic loss" involving a multivariate posterior distribution, the "recommended value" turns out to be the vector of mean values, while the "error bars" are provided by the corresponding covariance matrix. Thus, Section 2.1 also presents simple examples of estimating covariances and confidence intervals from experimental data.

Practical applications require not only mathematical relations between probabilities, but also rules for assigning numerical values to probabilities. As is well known, Bayesian statistics provides no fundamental rule for assigning the prior probability to a theory. The choice of the "most appropriate" prior distribution lies at the heart of applying Bayes' theorem to practical problems, and has caused considerable debates in the past, lasting over a century. Section 2.2 discusses the assignment of prior probability distributions under incomplete information. Of course, when complete prior information related to the problem under consideration is available and can be expressed in the form of a probability distribution, this information should certainly be used. In such cases, the repeated application of Bayes' theorem will serve to refine the knowledge about the respective problem. At the other extreme, when no specific information is available, it may be possible to construct prior distributions using concepts of group theory to reflect the possible invariance and/or symmetry properties of the problem under consideration, as discussed in Section 2.2.1. On the other hand, if repeatable trials are not feasible, but some information could nevertheless be inferred by some other means, information theory can be used in conjunction with the maximum entropy principle (the modern generalization of Bernoulli's principle of insufficient reason) to assign numerical values to probabilities, thus constructing a prior distribution, as will be shown in Section 2.2.2. The material presented in Section 2.2 is certainly not exhaustive regarding the use of group theory and symmetries for assigning priors (which continues to remain an area of active research), but is limited to presenting only the most commonly encountered priors in practice, and which are also encountered throughout this book.

Section 2.3 presents methods for evaluating unknown parameters from data which is consistent "within error bars," and is afflicted solely by random errors. Three common situations are considered, as follows: (i) evaluation of a location parameter when the scale parameters are known; (ii) both the scale and location parameters are unknown but need to be evaluated; and (iii) evaluation of a counting rate (a scale parameter) in the presence of background (noise). It is probably not too unfair to say that, although measurements without systematic errors are the exception rather than the rule, conventional (frequentist) sampling theory has not much to offer to practitioners in science and technology who are confronted with systematic errors and correlations. This is in marked contrast to the wealth of material on statistical errors, for which satisfactory techniques are available, based on counting (Poisson) statistics or on Gaussian models for the scatter of repeatedly measured data. Using Bayesian parameter estimation under quadratic loss, group-theoretical least informative priors, and probability assignment by entropy maximization, Section 2.4 addresses the practical situation of evaluating means and covariances from measurements affected by both random (uncorrelated) and systematic (correlated) errors issues. It is explained how common errors, the most frequent type of systematic error, invariably induce correlations, and how correlations are described by nondiagonal covariance matrices. As will be mathematically shown in this section, the random errors can be reduced by repeated measurements of the same quantity, but the systematic errors cannot be reduced this way. They remain as a "residual" uncertainty that could be reduced only by additional measurements, using different techniques and instrumentation, geometry, and so on. Eventually, additional measurements using different techniques would reduce the correlated (systematic) error just as repetitions of the same measurement using the same technique reduce the uncorrelated statistical uncertainty.

As discussed in Chapter 1, unrecognized or ill-corrected experimental effects, including background, dead time of the counting electronics, instrumental resolution, sample impurities, and calibration errors usually yield inconsistent experimental data. Although legitimately discrepant data may occur with a nonzero probability (e.g., for a Gaussian distribution, the probability that two equally precise measurements are outside of two standard deviations is about 15.7%), it is much more likely that apparently discrepant experiments actually indicate the presence of unrecognized errors. Section 2.5 illustrates the basic principles for evaluating discrepant data fraught by unrecognized, including systematic (common), errors. The marginal distributions for both recognized and unrecognized errors are obtained for both the Jeffreys least informative prior and for an exponential prior (when the unrecognized errors can be characterized by a known scale factor). This treatment of unrecognized systematic errors is an example of a two-stage "hierarchical" Bayesian method, involving a twofold application of Bayes' theorem to the sampling distribution that depends on parameters having a Gaussian prior, which in turn depended on a so-called "hyper-parameter," which had itself a "hyper-prior" distribution. Also the first-order expressions obtained for the various quantities are similar to the James-Stein estimators, which have sometimes lower risk than the estimates resulting from Bayesian estimation under quadratic loss (that minimize the square error averaged over all possible parameters, for the sample at hand). It is important to note, though, that the two-stage method Bayesian used in Section 2.5 yields results that are superior to the James-Stein estimators, especially for small samples. Moreover, the results presented in this section yield further improvements in a systematic and unambiguous way, without the discontinuities, questions of interpretation and restrictions associated with James-Stein estimators. This fact is particularly valuable and relevant for scientific data evaluation, where best values must often be inferred (for quadratic or any other loss) from just a single available sample.

Chapter 3 presents minimization algorithms, which are best suited for unconstrained and constrained minimization of large-scale systems such as time-dependent variational data assimilation in weather prediction and similar applications in the geophysical sciences. The operational implementation of "four-dimensional variational" data assimilation (customarily called 4–D VAR) hinges crucially upon the fast convergence of efficient gradient-based large-scale unconstrained minimization algorithms which are called to minimize a cost function that attempts to quantify the discrepancies between forecast and observations in a window of assimilation, subject to constraints imposed by the geophysical model. Data assimilation problems in oceanography and meteorology contain many typically of the order of ten million degrees of freedom. Consequently, conjugate-gradient (CG) methods and limited-memory quasi-Newton (LMQN) methods come into consideration since they typically require storage of only a few vectors, containing information from a few iterations, converge to local minima even from remote starting points, and can be efficiently implemented on multiprocessor machines.

Section 3.2 highlights the common as well as distinctive salient features of the minimization algorithms called (acronyms) CONMIN, E04DGF, and of the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method, and Bert-Buckley-Variable-Storage-Conjugate-Gradient (BBVSCG) method. These methods all fall under the category of Limited-Memory Quasi-Newton (LMNQ) methods, combining the advantages of CG-algorithms (low storage requirements) with the advantages of quasi-Newton of (QN) methods (computational efficiency stemming from their superlinear convergence). The LMQN algorithms build several rank one or rank two matrix updates to the Hessian matrix, thereby avoiding the need to store the approximate Hessian matrix, as required by full QN methods. Like the CG-methods, the LMQN methods require only modest amount storage for generating the search directions. Currently, the L-BFGS algorithm is the widest used minimization algorithm at the operational numerical weather prediction centers which rely on large-scale 4–D VAR assimilation and prediction methodologies.

Section 3.3 presents Truncated Newton (T-N) methods, which attempt to retain the rapid (quadratic) convergence rate of classic Newton methods, while requiring evaluations of functions and gradients only, thereby reducing the storage and computational requirements to sizes that are feasible for large-scale minimization applications. The T-N methods are also called *Hessian-free methods.* When used together with a finite difference approximation to the Hessian-vector products, the T-N methods achieve a quasi-quadratic convergence rate. In recent implementation, these Hessian-vector products are computed most efficiently using adjoint methods. Thus, the T-N methods require forward and backward adjoint inner iterations within a CG-formalism. Therefore, although the T-N methods offer competitive alternatives for two-dimensional problems, they are not yet competitive for 3-D operational problems. This is because the high cost of the CG inner iterations offsets the advantage of its almost quadratic convergence rate.

Section 3.4 discusses the use of information provided by the Hessian matrix for large-scale optimization, highlighting, in particular, the way in which the eigenvalues of the Hessian matrix determine the convergence rate for unconstrained minimization. Section 3.5 discusses issues related to nonsmooth and nondifferentiable optimization, in view of the fact that precipitation and radiation parameterizations involve on/off processes. Methods of nondifferentiable optimization are needed to minimize nonsmooth functionals. Nonsmooth optimization methods are based on the assumptions that: (i) the functional to be minimized is locally Lipschitz continuous, and (ii) the functional and its arbitrary subgradients can be evaluated at each point. Nonsmooth optimization methods can be divided into two main classes: subgradient methods and bundle methods. The guiding principle underlying bundle methods is to gather the subgradient information from previous iterations into a bundle of subgradients. Although the additional computational cost for building the sub-gradient is several times larger than that for L-BFGS, bundle nonsmooth optimization methods may work advantageously for problems with discontinuities, where L-BFGS methods usually fail. Bundle nonsmooth optimization methods have not been tested yet on operational 4–D VAR systems, but investigations are in progress to assess the applicability of such methods to realistic large-scale problems.

Section 3.6 addresses two fundamental issues related to step-size searches in conjugate-gradient type methods: (i) How good is the search direction?; and (ii) What is the best choice for the length of the step along the search direction? Section 3.7 highlights the salient features of trust-region methods, which also seek global convergence while retaining fast local convergence of optimization algorithms. It is noted that the trust region methods follow a reverse sequence of operations, by first choosing a trial step length, and subsequently using a quadratic model to select the best step length.

Section 3.8 discusses scaling and preconditioning for linear and nonlinear problems. The goal of preconditioning is to improve the performance of conjugate gradient–type minimization methods, by reducing the number of iterations required to achieve a prescribed accuracy. Scaling can substantially improve the performance of minimization algorithms. An effective automatic scaling could also improve the condition number of the Hessian matrix for well-scaled problems, thus facilitating their solution. Scaling by variable transformations converts the variables from units that reflect the physical nature of the problem to units that display desirable properties for improving the efficiency of the minimization algorithms. On the other hand, badly scaled nonlinear problems can become extremely difficult to solve.

Several popular methods for performing nonlinear constrained optimization are discussed in Section 3.9, namely: (i) the penalty method; (ii) barrier methods; (iii) augmented Lagrangian methods; and (iv) sequential quadratic programming (SQP) methods. The penalty method replaces a constrained optimization problem by a series of unconstrained problems whose solutions should converge to the solution of the original constrained problem. The unconstrained problems minimize an objective function which is constructed by adding to the original objective function a term that comprises a penalty parameter multiplying a measure of the violation of the constraints. The measure of violation is nonzero when the constraints are not satisfied, and is zero in the region where the constraints are satisfied. The original problem can thus be solved by formulating a sequence of unconstrained subproblems. Barrier methods are an alternative class of algorithms for constrained optimization. These methods also use a penalty-like term added to the objective function, but the results of iterations within the barrier methods are forced by the barrier to remain interior to and away from the boundary of the feasible solution domain. Augmented Lagrangian methods turn a constrained minimization problem into the unconstrained minimization The SQP solves a sequence of sub-problems designed to minimize a quadratic model of the objective functional subject to linearization of the constraints. The SQP method can be used within either a linesearch or a trust region framework, and is very efficient for solving both small and large problems. To be practical, a SQP method must be able to converge on nonconvex problems, starting from remote points.

Nonlinear optimization may involve cost functions characterized by the presence of multiple minima. The aim of global optimization is to determine all of the critical points of a function, particularly if several local optima exist where the corresponding function values differ substantially from one another. Global optimization methods can be classified into two major categories, namely deterministic and stochastic methods. Deterministic methods attempt to compute all of the critical points with probability one (i.e., with absolute success). On the other hand, stochastic methods sacrifice the possibility of an absolute guarantee of success, attempting to minimize the function under consideration in a random sample of points from a set, which is assumed to be convex, compact, and to contain the global minimum as an interior point. Section 3.10 briefly discusses two stochastic global minimization methods: simulated annealing and genetic algorithms, which have recently been implemented in variational data assimilation in geophysical sciences applications. The simulated annealing algorithm exploits the analogy between the search for a global minimum and the annealing process (i.e., the way in which a metal cools and freezes) into a minimum energy crystalline structure. The genetic algorithms attempt to simulate the phenomenon of natural evolution, as each species searches for beneficial adaptations in an ever-changing environment. As species evolve, new attributes are encoded in the chromosomes of individual members. This information changes by random mutation, but the actual driving force behind evolutionary development is the combination and exchange of chromosomal material during breeding.

Simulated annealing algorithms are intrinsically sequential. On the other hand, genetic algorithms are particularly well suited for implementation on parallel computers. Evaluation of the objective function and constraints can be done simultaneously for the entire population; the production of the new population by mutation and crossover can also be parallelized. On highly parallel machines, therefore, a genetical algorithm (GA) can be expected to run nearly N times faster than on non-parallel machines, where N is the population size. Currently, however, the convergence rate of these global minimization methods does not outperform the convergence rate of LMQN methods for large-scale operational 4–D VAR models.

Chapter 4 discusses several basic principles of four-dimensional variational assimilation (4–D VAR). Initially, data assimilation methods were referred to as "objective analyses," in contradistinction to "subjective analyses," in which numerical weather predictions (NWP) forecasts were adjusted "by hand" by meteorologists, using their professional expertise. Subsequently, methods called "nudging" were introduced based on the simple idea of Newtonian relaxation. In nudging, the rightside of the model's dynamical equations is augmented with a term which is proportional to the difference between the calculated meteorological variable and the observation value. This term keeps the calculated state vector closer to the observations. Nudging can be interpreted as a simplified Kalman-Bucy filter with the gain matrix being prescribed rather than obtained from covariances. The nudging method is used in simple operational global-scale and meso-scale models for assimilating small-scale observations when lacking statistical data. The recent advances in nudging methods are briefly presented in Section 4.1.

Section 4.2 briefly mentions the “optimal interpolation” (OI) method, “three-dimensional variational data assimilation” (3–D VAR), and the physical space statistical analysis (PSAS) methods. These methods were introduced independently, but were shown to be formally equivalent; in particular, PSAS is a dual formulation of 3–D VAR.

Data assimilation requires the explicit specification of the error statistics for model forecast and the current observations, which are the primary quantities needed for producing an analysis. A correct specification of observation and background error covariances are essential for ensuring the quality of the analysis, because these covariances determine to what extent background fields will be corrected to match the observations. The essential parameters are the variances, but the correlations are also very important because they specify the manner in which the observed information will be smoothed in the model space if the resolution of the model does not match the density of the observations. Section 4.3 briefly outlines the prevailing operational practices employed for the practical estimation of observation error covariance matrices and background error covariance matrices.

The goal of the 4–D VAR formalism is to find the solution of a numerical forecast or numerical weather prediction (NWP) model that best fits sequences of observational fields distributed in space over a finite time interval. Section 4.4 discusses the basic framework of “four-dimensional variational” data assimilation (4–D VAR) methods utilizing optimal control theory (variational approach). The advance brought by the variational approaches is that the meteorological fields satisfy the dynamical equations of the forecast model while simultaneously minimizing a cost functional, which measures the differences between the computed and the observed fields, by solving a constrained minimization problem. The 4–D VAR formalism is first presented without taking the modeling errors into account; subsequently, the functional to be minimized is extended to include model errors. This section concludes with a discussion of the consistent optimality and transferable optimality properties of the 4–D VAR procedure.

Section 4.5 presents results of numerical experiments with unconstrained minimization methods for 4–D VAR using the shallow water equations, which are widely used in meteorology and oceanography for testing new algorithms since they contain most of the physical degrees of freedom (including gravity waves) present in the more sophisticated operational models. The numerical experiments were performed with four limited-memory quasi-Newton (LMQN) methods (CONMIN-CG, E04DGF, L-BFGS, and BBVSCG) and two truncated Newton (T-N) methods. The CONMIN-CG and BBVSCG algorithms failed after the first iteration, even when both gradient scaling and non-dimensional scaling were applied. The L-BFGS algorithm was successful only with gradient scaling. On the other hand, the E04DGF algorithm worked only with the non-dimensional shallow water equations model. This indicates that using additional scaling is essential for the success of LMQN minimization algorithms when applied to large-scale minimization problems. On the other hand, T-N methods appear to perform best for large-scale minimization problems, especially in conjunction with a suitable preconditioner. The importance of preconditioning increases with increasing dimensionality of the minimization problem under consideration. Furthermore, for the Shallow-Water Equations (SWE) numerical experiments, the T-N methods required far fewer iterations and function calls than the LMQN methods.

In the so-called strong constraint Variational Data Assimilation (VDA), or classical VDA, it is assumed that the forecast model perfectly represents the evolution of the actual atmosphere. The best fit model trajectory is obtained by adjusting only the initial conditions via the minimization of a cost functional that is subject to the model equations as strong constrains. However, numerical weather prediction (NWP) models are imperfect since subgrid processes are not included. Furthermore numerical discretizations produce additional dissipative and dispersion errors. Modeling errors also arise from the incomplete mathematical modeling of the boundary conditions and forcing terms, and from the simplified representation of physical processes and their interactions in the atmosphere. Usually, all of these modeling imperfections are collectively called *model error* (ME). Model error is formally introduced as a correction to the time derivatives of model variables. Section 4.6 highlights the treatment of MEs in VDA, taking into account numerical errors explicitly as additional terms in the cost functional to be minimized. However, taking MEs into account doubles (and can even triple) the size of the system to be optimized by comparison to minimizing the cost functional when the model errors are neglected.

Chapter 5 highlights specific difficulties in applying 4–D VAR to large-scale operational numerical weather prediction models. Recall that the objective of 4–D VAR is to find the optimal set of control variables, usually the initial conditions and/or boundary conditions, such that a cost function, comprising a weighted least square norm that quantifies the misfit between model forecast and observations, is minimized subject to the constraints of satisfying the geophysical model. In order to minimize this cost function, we need to know the gradient of this function with respect to the control variables. A straightforward way of computing this gradient is to perturb each control variable in turn and estimate the change in the cost function. But this method is impractical when the number of control variables is large a typical meteorological model comprises $O(10^7)$ control variables. Furthermore, the iterative minimization of the cost function requires several gradient estimations on the way to finding a local minimum. Often, these gradient estimations are insufficiently accurate to guarantee convergence of the minimization process. As discussed in Section 5.1, the convergence of the minimization process in 4–D VAR is particularly affected (negatively) by strong nonlinearities and on/off physical processes such as precipitation and clouds.

Section 5.2 highlights the highly efficient adjoint method for computing exactly the gradient of the cost function with respect to the control variables by integrating once the adjoint model backwards in time. Such a backward integration is of similar complexity to a single integration of the forward model. Another key advantage of adjoint variational data assimilation is the possibility to minimize the cost function using standard unconstrained minimization algorithms (usually iterative descent methods). However, for precipitation observations, highly nonlinear parameterization schemes must be linearized for developing the adjoint version of the model required by the minimization procedure for the cost function. Alternatively, simpler physics schemes, which are not a direct linearization of the full model physics (i.e., the "linear model" is not tangent linear to the nonlinear full model), can be coded for the linear model. Of course, it is desirable to have a linearized model that approximates as closely as possible the sensitivity of the full nonlinear model; otherwise the forecast model may not be in balance with its own analysis, producing so-called “model spin-up.” Furthermore, multi-incremental approaches can exhibit discrete transitions affecting the stability of the overall minimization process.

On the one hand, nonlinear models have steadily evolved in complexity in order to improve forecast skill. For example, the prognostic cloud scheme introduced into the European Centre for Medium Range Weather Forecasts (ECMWF) model includes many highly nonlinear processes that are often controlled by threshold switches. On the other hand, even if it were possible to construct the tangent linear and, respectively, adjoint models corresponding to this complex cloud scheme, the validity of these models would be restricted due to these thresholds and their value would be questionable. Issues related to using non-smooth optimization methods to address discontinuities is an ongoing research topic. Standard solvers of elliptic equation often perform “fast Fourier transform” (FFT) and inverse FFT operations. Section 5.3 summarizes the adjoint coding of the FFT and of the inverse FFT, showing that the adjoint of the FFT is obtained by calling the inverse FFT routine and multiplying the output by the number of data sample points. Conversely, the adjoint of the inverse FFT is obtained by calling the FFT routine and dividing the output by the number of data sample points.

Section 5.4 indicates that the correctness of the adjoint code for interpolations and on/off processes can be verified by performing an additional integration of the nonlinear model with added bit vectors in order to determine the routes for the IF statements included in the physical processes. This additional integration of the nonlinear model with added bit vectors is needed for the verification of both the tangent linear model and the adjoint model. Section 5.5 discusses the construction of background covariance matrices, which play the following important roles: (i) spreading the information from the observations to neighboring domains; (ii) providing statistically consistent increments at the neighboring grid points and levels of the model; and (iii) ensuring that observations of one model variable (e.g., temperature) produce dynamically consistent increments in the other model variables (e.g., vorticity and divergence).

Section 5.6 revisits the characterization of model errors specifically for the 4–D VAR data assimilation procedure. Such errors are attributable to the dynamical model (e.g., poor representation of processes, omissions, or incorrect formulations of key processes, numerical approximations) and observations or measurements (e.g., sensor design, performance, noise, sample averaging, aliasing). For dynamically evolving systems, the model errors are expected to depend on time and, possibly, on the model state variables. Controlling modeling errors in addition to the model's initial conditions in the weak constraint 4–D VAR doubles the size of the optimization problem by comparison to the strong constraint 4–D VAR. Furthermore, if the stochastic component is included in the model error formulation, then the random realization would need to be saved at each model time step. Consequently, the size of the optimization problem would be tripled. The size of the model error control vector can be reduced by projecting it onto the subspace of eigenvectors corresponding to the leading eigenvalues of the adjoint-tangent linear operators.

Most data assimilation systems are not equipped to handle large, systematic corrections; they were designed to make small adjustments to the background fields that are consistent with the presumed multivariate and spatial structures of random errors. Statistics of "observed-minus-background" residuals provide a different, sometimes more informative, view on systematic errors afflicting the model or observations. Operational NWP centers routinely monitor time- and space-averaged background residuals associated with different components of the observing system, providing information on the quality of the input data as well as on the performance of the assimilation system. In general, small root-mean-square residuals imply that the system is able to accurately predict future observations. Nonzero mean residuals, however, indicate the presence of biases in the observations and/or their model-predicted equivalents. It is paramount to develop physically meaningful representations of model errors that can be clearly distinguished from possible observation errors. This issue is the subject of intensive ongoing research.

Section 5.7 discusses the incremental 4–D VAR algorithm, which was formulated in the mid-1990s, and decisively facilitated the adoption, application, and implementation of 4–D VAR data assimilation at major operational centers, thereby advancing the timely state of weather prediction. Prior to the development of the "incremental 4–D VAR algorithm," implementation of the full 4–D VAR algorithm in operational models was impractical, since a typical minimization requires between 10 and 100 evaluations of the gradient. The cost of the adjoint model is typically 3 times that of the forward model, and the analysis window in a typical operational model such as the ECMWF system is 12-hours. Thus, the cost of a 12 hour analysis was roughly equivalent to between 20 and 200 days of model integration (with 108 variables), making it computationally prohibitive for NWP centers which had to deliver timely forecasts to the public. In addition, the nonlinearity of the model and/or of the observation operator could produce multiple minima in the cost function, which impacted the convergence of the minimization algorithm.

The incremental 4–D VAR algorithm reduces the resolution of the model and eliminates most of the time-consuming physical packages, thereby enabling the 4–D VAR method to become computationally feasible. Furthermore, the incremental 4–D VAR algorithm removes the nonlinearities in the cost minimization by using a forward integration of the linear model instead of a nonlinear one. The minimization procedure is identical to the usual 4–D VAR algorithm except that the increment trajectory is obtained by integration of the linear model. The reference trajectory (which is needed for integrating the linear and adjoint models and which starts from the background integration) is not updated at every iteration. This simplified iterative procedure for minimizing the incremental cost function is called the *inner loop*, and is much cheaper computationally to implement by comparison to the full 4–D VAR algorithm. However, when the quadratic cost function is approximated in this way, the incremental 4–D VAR algorithm no longer converges to the solution of the original problem. Furthermore, the analysis increments are calculated at reduced resolution and must be interpolated to conform to the high-resolution model's grid. Consequently, after performing a user-defined number of inner loops, one outer loop is performed to update the high-resolution reference trajectory and the observation departures. After each outer loop update, it is possible to use progressively higher resolutions for the inner loop.

However, experiments show that the current implementations of the incremental 4–D VAR algorithm lead to divergent computational results after four outer loop iterations (e.g., this was the case when the incremental 4–D VAR was initially implemented into the ECMWF weather prediction system). Various numerical experiments indicate that convergence can be attained when the inner and outer loops use the same resolution and/or use the same time step. This feature is explained by the presence of gravity waves which propagate at different speeds in the linear and nonlinear models. These gravity waves are related to the shape of the leading eigenvector of the Hessian of the 4–D VAR cost function; this eigenvector is determined by the surface pressure observation and controls the convergence of the minimization algorithm. Chapter 5 concludes with a short discussion, in Section 5.8, of current research issues.

This book also comprises three appendices. Appendix A (Chapter 6) is intended to provide a quick reference to selected properties of distributions commonly used for data analysis, evaluation, and assimilation. Appendix B (Chapter 7) introduces and summarizes the most important properties of adjoint operators in conjunction with differential calculus in vector spaces, as used for data assimilation. Appendix C (Chapter 8) highlights the main issues arising when identifying and estimating model parameters from experimental data. The problem of parameter identification can be formulated mathematically as follows: "an unknown parameter is *identifiable* if it can be determined uniquely at all points of its domain by using the input-output relation of the system and the input-output data." Such a mapping is generally known as an "inverse problem," in contradistinction with the forward mapping, which maps the space of parameters to the space of outputs. The uniqueness of inverse mappings is difficult to establish. Appendix C also discusses briefly three methods for estimating parameters: the maximum likelihood method, the maximum total variation L_1-regularization method for estimating parameters with discontinuities, and the "extended Kalman filter" method.