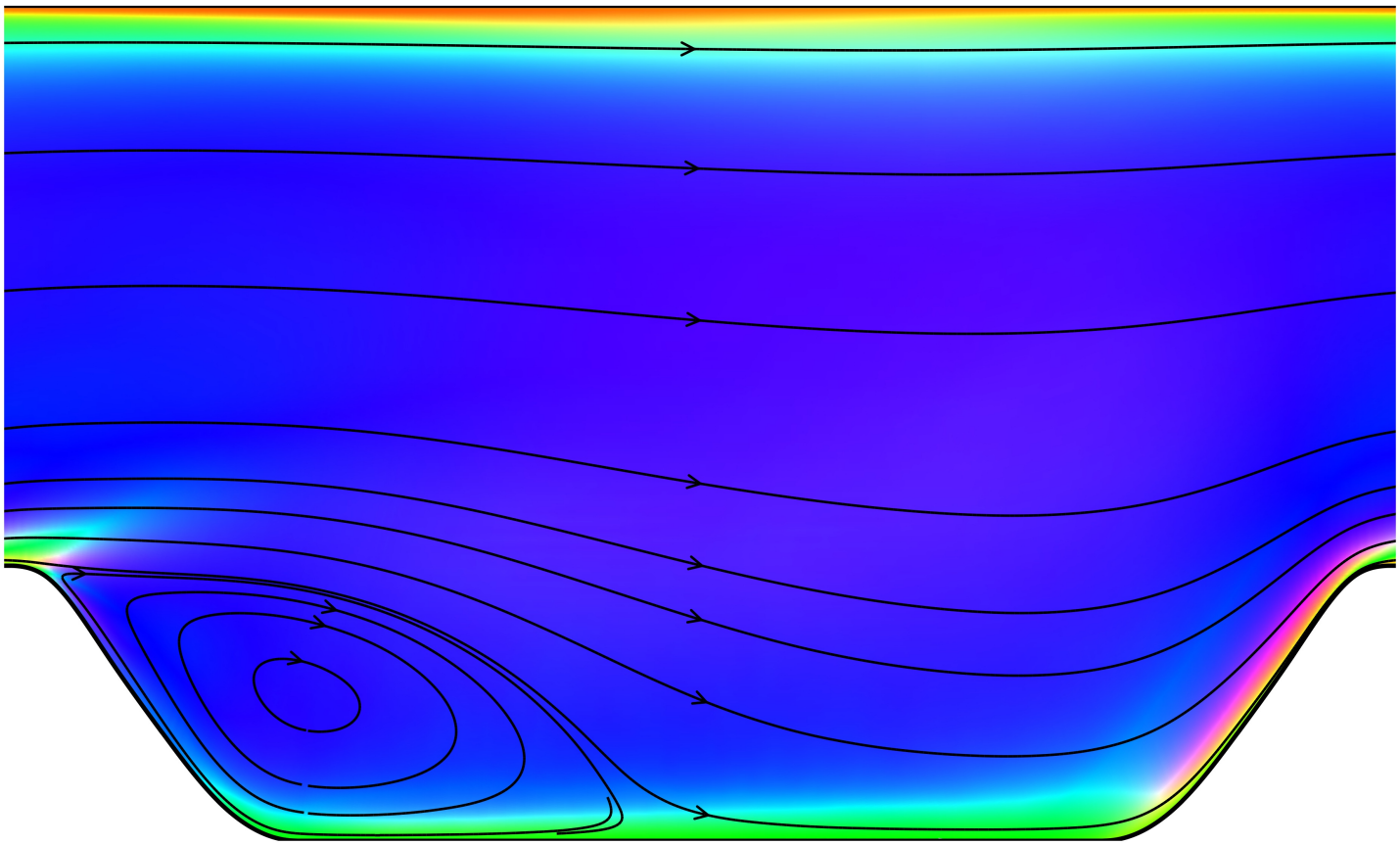
Technische Universität Braunschweig

TU Delft Delft University of Technology



Classification/Regression Hybrid Approaches for Data-Driven RANS Modelling

Henrik Wüstenberg

October 14, 2020

Technische Universität Braunschweig

TU Delft Delft University of Technology

Master's thesis

Classification/Regression hybrid approaches for data-driven RANS modelling

B. Eng. Henrik Wüstenberg
Matriculation number 4911788

For obtaining the degree Master of Science in Computational Sciences in Engineering at the Technische Universität Braunschweig

First examiner: Prof. Dr.-Ing. R. Radespiel
Institute of Fluid Mechanics
Technische Universität Braunschweig

Second examiner: Dr. Richard P. Dwight
Faculty of Aerospace Engineering
Technische Universiteit Delft

The thesis is published on the 14th of October 2020.

Declaration of truthfulness

I hereby confirm that I am the sole author of the present master's thesis. I only used the specified resources and all formulations and concepts taken verbatim or in substance from printed or unprinted material or from the Internet have been cited according to the rules of good scientific practice and indicated by exact references to the original source. The present thesis has not been submitted to another university for the award of an academic degree in this form. This thesis has been submitted in printed and electronic form. The content of the digital version is identical to the printed version. I understand that the provision of incorrect information may have legal consequences.

Place, date

Signature

Task description

Project title:

Classification / regression hybrid approaches for data-driven RANS modelling
Klassifizierung / Regression Hybridansätze für datengesteuerte RANS-Modellierung

Supervisor:

Dr. Richard Dwight

Faculty and section of supervisor(s):

TU Delft, Faculty of Aerospace Engineering, Aerodynamics

Main focus of project:

Modelling and algorithmic development

Difficulty level of the project:

Master level

Problem description

Turbulence is a multi-scale phenomenon, which presents a challenge for simulation. Resolving all scales of a flow is computationally intractable, necessitating modelling of effects of turbulence on the mean-flow. Reynolds-averaged Navier-Stokes (RANS) approaches split the flow into time-averaged and fluctuating parts, and the fluctuating parts are modelled. Modelling has historically been performed with a combination of physical insight and empirical data-fitting, and has resulted in RANS becoming the dominant fluid-modelling paradigm in engineering. However, progress in RANS modelling has stalled, existing models can be inaccurate in many specific flows of interest, and no reliable error estimates exist.

In this project we investigate data-driven approaches to designing new RANS models. With the proliferation of high-resolution data-sets from Large Eddy and Direct Numerical Simulation (LES/DNS) and advances in supervised machine-learning, we have the opportunity to design turbulence models based primarily on data [Duraisamy2017]. These promise new improved models, optimized for specific applications, and techniques for estimating uncertainty in existing models [Edeling2014].

The main problem – that this thesis should address – is that, while training models for specific, narrow classes of flows can be done quite successfully [Schmelzer2019], the resulting models tend to worsen predictions (compared to baseline $k-\varepsilon$ or $k-\omega$ models) for fundamental flows like channels, zero-pressure-gradient BLs and even isotropic decaying turbulence outside the training-data. They are hence very far from general-purpose models, and of limited utility.

One possible solution involves activating data-trained modelling terms only when the Boussinesq approximation is poor. So far, there has been only one attempt to automatically classify regions of flows in this way, by [Ling2015] who used decision trees as a classifier, and three different metrics to assess the quality of Boussinesq locally. The result for each point in the flow is a binary assessment, and Ling et al. showed relatively good results in terms of false-negatives and false-positives in prediction cases. However, this is difficult to use as a starting point for a new model, as the resulting decision tree (a) is

essentially a black-box, not amenable to inspection, (b) would be expensive to evaluate (at every mesh-cell at every step of a solver), and (c) returns a binary result, rather than a smooth prediction, and (d) is considered to predict non-connected spots of poor performance.

Using this as a starting point, we will develop a composite classifier/model method. The classifier must return a "Boussinesq confidence" prediction, capturing similar information as the decision tree, but with a smooth prediction (e.g. on the interval $[0,1]$). This classifier should ideally be an open- or gray-box, such as models based on symbolic regression (SR) or gene-expression programming (GEP). Prediction performance will be compared with Ling's classifier on a variety of fundamental and complex 2d test-cases. If work on the classification problem is successful, it will be employed as the constituent part of a model.

Tasks

- Literature study
  - Turbulence modelling basics (Pope, Leschziner)
  - Machine-learning basics, focus on classification (Bishop)
  - Data-driven turbulence modelling (see below)
  - Selection of promising algorithms.
- Derivation and implementation of a new classifier based on, for example, SR or GEP.
- Training/testing of hybrid models with combined classification/regression.
- Document results.

Qualifications

Knowledge in OpenFOAM, Python, C++, Fluid Mechanics and Statistics.

Literature

[Duraisamy2017] K. Duraisamy, Data-enabled, Physics-constrained Predictive Modeling of Complex Systems, SIAM News, 19 July 2017.

[Edeling2014] W. N. Edeling, P. Cinnella, R. P. Dwight, Predictive RANS simulations via Bayesian Model-Scenario Averaging, Journal of Computational Physics 275, pp. 65–91, 2014.

[Ling2016] J. Ling, A. Kurzawski, J. Templeton, Reynolds averaged turbulence modelling using deep neural networks with embedded invariance, Journal of Fluid Mechanics 807, pp. 155-166, 2016.

[Ling2015] J. Ling, J. Templeton, Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty, Physics of Fluids 27 (8), 085103, 2015.

[Parish2016] E. J. Parish and K. Duraisamy, A paradigm for data-driven predictive modeling using field inversion and machine learning, Journal of Computational Physics, 305, 758–774, 2016.

[Schmelzer2019] M. Schmelzer, R.P. Dwight, P. Cinnella, Machine Learning of Algebraic Stress Models using Deterministic Symbolic Regression, Flow, Turbulence & Combustion, 2019. (arXiv preprint arXiv:1905.07510)

[Wang2017] J.-X. Wang, J.-L. Wu and H. Xiao, Physics-informed machine learning approach for reconstructing Reynolds stress modeling discrepancies based on DNS data, Physical Review Fluids 2, 034603, 2017.

[Weatheritt2016] J. Weatheritt, R. Sandberg, A novel evolutionary algorithm applied to algebraic modifications of the RANS stress–strain relationship, Journal of Computational Physics 325: 22-37, 2016.

Preface

This is my master's thesis the final piece to complete my studies in Computational Sciences in Engineering at the TU Braunschweig. The thesis allowed me to dive into two fascinating fields of research: turbulence and machine learning. Turbulence describes a chaotic phenomena which combines the natural beauty of fluid flow with an intricate complexity that keeps stimulating my curiosity. In machine learning, computers search for patterns in data and learn to predict these. In combining both fields, I had the chance to conduct captivating research and gather a rich learning experience.

Throughout this thesis I have received much support and assistance. I would first like to thank my supervisor Dr. Richard Dwight for giving me the opportunity to join his research group at the TU Delft and for providing me with the freedom to creatively explore my research topic, thank you for numerous meetings and inspiring discussions.

I would like to acknowledge Prof. Dr.-Ing. Rolf Radespiel for his support throughout the project from Braunschweig and for sharing his extensive experience in the field of turbulence modelling.

In addition, I want to thank the students of the high-speed lab at TU Delft for their open and welcoming attitude. The many tea and coffee breaks broadened my perspective of research in Aerodynamics.

Finally, I want to thank my family and friends for wise counsel and an open ear whenever I needed it. You have helped me to disconnect from the work for a while. I am very grateful for your support throughout the last years of my studies and, especially, the last months.

Henrik Wüstenberg
Delft, October 2020

Abstract

Reynolds-averaged Navier-Stokes (RANS) turbulence models have a narrow range of applicability and low predictive accuracy. The SST model improves the range and overall accuracy with a smooth boundary layer identification. The identification allows blending of two turbulence models. The present work explores the possibility of more versatile identifications based on machine learning algorithms. The identification locally detects either high or low uncertainty in RANS predictions. High uncertainty is presumed when assumptions in RANS eddy-viscosity models are violated due to either anisotropic turbulence or negative eddy viscosity. The machine learning algorithms train identifier models to predict high uncertainty using local mean flow quantities based on high-fidelity simulation data. To provide challenging flow physics, the high-fidelity data includes a variety of flows with strong separation and adverse pressure gradients. The selected algorithms train algebraic identifier models with a simple mathematical form that enables direct insight into the identification. It is shown that algebraic models of higher complexity show significant improvements. They correctly identify most violations with mean true-positive rate of 88 % for anisotropy and 77 % for negative eddy viscosity. Qualitatively, the identifier models detect wall-blocking induced anisotropy and negative eddy viscosity in strongly accelerated flows. However, the effect of complexer flow physics on the identification including the convection of turbulence and influence of strong adverse pressure gradients are not adequately identified by the models.

Übersicht

Reynolds-averaged Navier-Stokes (RANS) Turbulenzmodelle haben einen eingeschränkten Anwendungsbereich und geringe Vorhersagegenauigkeit. Das SST Modell verbessert den Anwendunsbereich und die Genaugikeit, indem es zwei Turbulenzmodelle kombiniert. Diese Kombination basiert auf einer Grenzschichterkennung, welche einen glatten Übergang zwischen den Modellen ermöglicht. Die vorliegende Arbeit untersucht die Möglichkeit einer umfassenderen Erkennung und verwendet dafür Algorithmen für maschinelles Lernen. Die Erkennung sagt lokal eine entweder hohe oder niedrige Unsicherheit in RANS Modellen vorher. Dabei wird hohe Unsicherheit erkannt, wenn die Annahmen von Wirbelviskositätsmodellen durch entweder anisotrope Turbulenz oder negative Wirbelviskosität verletzt werden. Die Algorithmen trainieren Erkennungsmodelle, um hohe Unsicherheit auf Grundlage von mittleren Strömungsgrößen vorherzusagen. Zum Training werden Strömungsdaten aus skalenauflösenden Simulationen mit starker Ablösung oder positiven Druckgradienten verwendet. Da die gewählten Algorithmen Modelle mit einfachen mathematischen Termen trainieren, gewähren diese Einblick in die Erkennung. Die Auswertung zeigt, dass interpretierbare Modelle mit höherer Komplexität deutlich bessere Erkennungen durchführen. Sie detektieren den Großteil der verletzten Modellannahmen mit mittlerer Richtig-Positiv Rate von 88 % für Anisotropie und 77 % für negative Wirbelviskosität. Dabei identifizieren die Modelle Anisotropie in Wandnähe und negative Wirbelviskositäten in stark beschleunigten Strömungen. Jedoch wird der Einfluss komplexer Strömungsphänomene auf die Erkennung, einschließlich des Transports von Turbulenz und dem Effekt positiver Druckgradienten, nicht verlässlich identifiziert.

Contents

Nomenclature

Cartesian tensor notation is used throughout the document. The summation convention does not apply for Greek indexes.

Acronyms

Acronym	Full name
AoA	Angle of attack
APG	Adverse pressure gradient
CFD	Computational fluid dynamics
DNS	Direct numerical simulation
EASM	Explicit algebraic stress model
EVM	Eddy-viscosity model
FFX	Fast function extraction
FN	False-negative
FP	False-positive
GEP	Gene-expression programming
KDE	Kernel density estimate
kNN	k-Nearest-Neighbour
LDA	Laser doppler annemometry
LES	Large eddy simulation
LEVM	Linear eddy-viscosity model
MI	Mutual information
NACA	National Advisory Committee for Aeronautics
NLEVM	Non-linear eddy-viscosity model
PIV	Particle image velocimetry
RANS	Reynolds-averaged Navier-Stokes
RGB	Red-green-blue
ROC	Receiver operating characteristic
RSM	Reynolds stress transport model
SpaRTA	Sparse regression of turbulent stress anisotropy
TBL	Turbulent boundary layer
TN	True-negative
TP	True-positive

Indexes

Index	Name
$\overline{\{\cdot\}}$	Mean
$\{\cdot\}'$	Fluctuation
$\{\cdot\}^{(n)}$	nth realisation
$\widehat{\{\cdot\}}$	Normalisation
$\{\cdot\}^{\star}$	Normalisation factor
$\widetilde{\{\cdot\}}$	Approximation
$\{\cdot\}_{1C}$	One-component limiting turbulence state
$\{\cdot\}_{2C}$	Two-component limiting turbulence state
$\{\cdot\}_{3C}$	Three-component limiting turbulence state
$\{\cdot\}_{II}$	Anisotropy metric
$\{\cdot\}_{\nu_t}$	Non-negativity metric
$\{\cdot\}_{c}$	Chord length
$\{\cdot\}_{d}$	Wall distance
$\{\cdot\}_{H}$	Step height
$\{\cdot\}_{L_1}$	L1-norm
$\{\cdot\}_{L_2}$	L2-norm
$\{\cdot\}_{q}$	Features

Latin symbols

Symbol	Name
a_{ij}	Anisotropy tensor
b_{ij}	Normalised anisotropy tensor
$\mathcal{B}$	Candidate library
c	Scalar argument
C	Barycentric coordinate
d	Distance to the nearest wall
f	Basis function
H	Entropy
II	Second invariant of b_{ij}
III	Third invariant of b_{ij}
k	Turbulent kinetic energy
K_{ij}	Turbulent kinetic energy pseudo tensor
L	Loss function
$\mathcal{L}$	Length scale
L_c	Characteristic length
L_1	L1-regularisation
L_2	L2-regularisation
M	Data-driven model
MI	Mutual information function
N	Number of grid points
p	Pressure
$\check{p}$	Modified pressure
P_{ij}	Pressure gradient pseudo tensor
Pr	Probability function
Pr_m	Probability mass function
q	Feature
$\mathbf{Q}$	Orthogonal matrix
R	Loss weight
Re	Reynolds number
s	Sample size
S_{ij}	Mean-strain-rate tensor
t	Time
$\mathbf{T}$	Tensor argument
$\mathcal{T}$	Tensorial set
u	Velocity
$\mathcal{U}$	Velocity scale
v	Vector argument
w	Model coefficients

W	Arbitrary random variable
x	Space coordinate
X	Arbitrary random variable
y	Binary error metric
z	Sample point
Z	Group of sample points

Greek symbols

Symbol	Name
Γ	Unit velocity vector
δ_{ij}	Kronecker delta
Δ	Euclidean distance
ϵ	Dissipation rate
ϵ_{MI}	Error in mutual information
η	Barycentric map ordinate
θ	Reynolds stress tensor eigenvalue
λ	Anisotropy tensor eigenvalue
μ	Dynamic viscosity
ν	Kinematic viscosity
ν_t	Eddy viscosity
ξ	Barycentric map abscissa
ρ	Density
σ	Logistic function
σ_{ij}	Cauchy stress tensor
τ_{ij}	Specific Reynolds stress tensor
ϕ	Elastic net ratio
ψ	Digamma function
ω	Dissipation per unit k
Ω_{ij}	Vorticity tensor

1. Introduction

1.1. Motivation

The Navier-Stokes equations describe the evolution of turbulent flows. Computational fluid dynamics (CFD) approximates solutions to the equations. The approximating methods from high to low fidelity are direct numerical simulations (DNS), large eddy simulations (LES) and Reynolds-averaged Navier-Stokes (RANS). DNS solve the complete Navier-Stokes equations resolving all scales of turbulence. Their computational costs scale with the Reynolds number as Re^3 [1] which effectively limits their application to simple geometries and low Reynolds numbers. An alternative are LES which resolve only the larger scales and model the dissipating scales of turbulence. While computational costs are decreased compared to DNS, the turn-over time of LES is too large for industrial design. RANS methods provide a time-averaged solution where all scales of turbulence are modelled. However, the complexity of turbulence and empiricism in RANS models limit the accuracy of RANS predictions. The CFD vision 2030 study [2] concludes that, even with increasing computational power, RANS and hybrid RANS/LES methods continue to be the workhorse for the industry. They recommend further research into improving RANS turbulence models.

Well-established turbulence models for RANS are linear eddy-viscosity models (LEVM), for example the k-ϵ [3], k-ω [4] and Spalart-Allmaras [5] model. LEVMs use Boussinesq's hypothesis to estimate the effect of turbulence. This relationship is known to be invalid in many engineering flows including regions with curvature, separation, pressure gradients, swirl or secondary motion [6]. The limitations of LEVM are partially circumvented by models with non-linear relationships or models based on the Reynolds stress transport equations. Related types of models are non-linear eddy-viscosity models (NLEVM), explicit algebraic stress models (EASM) and Reynolds stress transport models (RSM) which improve predictions of turbulence, especially, for flows with curvature and secondary motion. Nonetheless, they are less robust than LEVM, which limits their range of application, and not guaranteed to be more accurate [7]. Recently, data-driven modelling became another possibility in improving RANS models [8]. Data-driven models learn to predict the discrepancy in RANS models from data of higher fidelity. These models show substantial improvements for anisotropy predictions, but their range of application is limited to flows similar to the training data [9]. In summary, specific turbulence models improve predictions for a narrow class of flows, but cannot generalise well.

This shortcoming of RANS models suggests an holistic approach to RANS turbulence modelling which identifies when a specific model should be used. One approach is the k-ω SST model [10] which identifies the boundary layer. The model applies a k-ω model close to the wall and smoothly transitions to a k-ϵ model in the free-stream. However, this approach blends only LEVMs and, thus, has the aforementioned limitations. Instead, a holistic approach requires an universal identification which allows switching, or blending, any turbulence model. For example, an identifier model predicts regions of high uncertainty in the solution of RANS using a probability. The probability allows to locally switch or blend turbulence models and improve predictions. Consequently, the identification potentially enables a general approach to RANS turbulence modelling which combines the advantages of a variety of application-specific models.

1.2. Literature review

The identification in the k-ω SST model [10] uses an analytical blending function to identify boundary layers. To the authors knowledge, no analytical blending function that universally identifies regions of high uncertainty in RANS has been found. Such a function is likely to reach high complexity, because of the complexity of the physics of turbulence. Instead, machine learning techniques enable a versatile usage of large data sets. Machine learning finds patterns in data, for example high-fidelity simulations of turbulent flow. The review [8] discusses novel data-driven techniques in the field of turbulence modelling. The methods include supervised machine learning approaches to RANS turbulence models. Supervised machine learning extracts relationships in data and embeds the relationships into models. These models learn to map from input features to outputs. For continuous outputs, the task is called regression and, for discrete outputs, classification. The review includes regression tasks where models learn to improve predictions of the anisotropy tensor $a_{ij} = \tau_{ij} - \frac{2}{3}k\delta_{ij}$. For these regressions, models are given inputs from RANS data and outputs from high-fidelity data to effectively learn to predict the discrepancy between high- and low-fidelity data. A similar approach is possible for universal identifier models.

The identification of high uncertainty in a turbulence model is a classification task. The review in [8] includes two methods for the identification of high uncertainty in RANS turbulence models. An analytical marker function is proposed in [11]. It assumes inaccurate LEVM predictions when the flow deviates from parallel shear flow. In a comparison of high-fidelity and RANS data, the method identifies regions where the divergence of the Reynolds stress tensor is inaccurate reasonably well. In [12], machine learning algorithms construct identifier models. The models learn to predict an error metric, which potentially indicates high uncertainty in RANS predictions. Error metrics give a binary response that identifies either anisotropic turbulence, negative or non-linear eddy viscosity. The training data provides a challenging set of complex and three-dimensional flows over a range of geometries and Reynolds numbers. Still, the identifier models achieve high true-positive rates and small false-positive rates when predicting on unseen data. True-positive rates reach up to 80 % and False-positive rates are below 10 % suggesting a promising approach.

The analytical marker in [11] is incorporated into the data-driven identifier models as a feature. Thus, the information of the marker prevails in the data-driven identifiers. Comparing both approaches, the data-driven models appear superior in predictive performance and, thus, for a universal identification.

The identifier models ideally provide a smooth blending between turbulence models. Blending functions are used in well-established turbulence models, for example, in the k-ω SST model [10]. The SST applies the k-ω model inside the boundary layer, but smoothly transitions towards the k-ϵ model in the freestream. A similarly smooth blending functionality is required for the identifier models.

The model structure and its capability to learn complex relationships depends on the choice of algorithm. The identifiier models in [12] are constructed with three distinct algorithms: support vector machines, Adaboost decision trees and random forests. They found that the random forest algorithm is robust to noisy data and, hence, achieved higher performance. Later, [13] showed that invariance properties embedded into data-driven models increase performance while decreasing computational costs. In [14], a deep neural network with embedded Galilean invariance learns to predict the anisotropy tensor a_{ij} directly and bypass a turbulence model. The predicted anisotropy tensor significantly improved over a LEVM and NLEVM in predicting the secondary motion in square ducts as well as separation regions on a wavy wall. [15] and [9] used a Galilean invariant random forest to predict six correction terms for the magnitude, shape and orientation of the anisotropy tensor a_{ij}. Corrections for streamwise velocity and shear stresses oscillate in regions

of low strain rate. Nonetheless, compared to a baseline turbulence model, the random forest model improved predictions in the recirculation and post-reattachment regions on a periodic hills geometry.

Clearly, the data-driven models outperform the baseline turbulence models. However, research [15, 9] has shown that data-driven models perform worse when tested on geometries different to the training data. Therefore, a database with varying flow conditions and complex phenomena, for example separation and curvature, is required. [9] discusses this issue of universality, and the interpretability, that is the possibility to examine a data-driven model for insights into the physics. The aforementioned algorithms, deep neural networks and random forests, train models with a black-box character which impedes physical insight. In contrast to black-box models, symbolic regression algorithms aim to construct interpretable models of low complexity, cf. [16]. An interpretable identifier model allows insights into the physical process of the identification. Moreover, the implementation of an algebraic model into a CFD solver is straightforward and low in computational costs.

A standard algorithm for classification tasks is Logistic Regression, cf. [17, 18]. The algorithm constructs symbolic models from linear combinations of input features. These models allow straightforward interpretation, but achieve only low complexity. Therefore, Logistic Regression models are a good choice to investigate the performance for minimum model complexity.

Higher model complexity is possible with the Sparse Regression of Turbulent Stress Anisotropy (SpaRTA) algorithm proposed in [19]. SpaRTA constructs a large library of candidate functions which is subsequently evaluated on training data to discover models. The algorithm uses sparsity constraints to ensure interpretable algebraic models. In [19], a database of separated flows serves as training data for models that learn the discrepancy in the anisotropy tensor a_{ij} when comparing RANS and DNS or LES data. The baseline k-ω model underpredicts the production of turbulence aft of the separation region. Applying the data-driven correction models, streamwise velocity and turbulent kinetic energy are substantially improved, while the models sustain a low complexity of at most two terms. Improving this complex process with rather simple correction terms makes SpaRTA a promising algorithm.

In contrast, [20] uses gene-expression programming (GEP) for complex but interpretable models. GEP is a stochastic algorithm for the discovery of models using evolutionary processes. The algorithm uses mutation and recombination operators on a set of candidates to construct and evolve model structures depending on a fitness function. Using a fitness function that compares the true and RANS anisotropy tensor a_{ij}, GEP is applied to separated [20] and duct flows [21] to learn correction models. The prediction of streamwise velocities and normal shear stresses strongly improves in the recirculation region compared to a baseline LEVM. Additionally, the correction in duct flow allows a LEVM to predict secondary motion.

Both, SpaRTA and GEP, seem a viable choice to construct universal identifier models. In comparison, GEP discovers, for each run, another model form with varying coefficients and complexity due to the stochastic operators. The GEP models for separated flows [20] and duct flows [21] consistently result in higher complexity models. They incorporate multiple terms including logarithmic and exponential functions while SpaRTA models achieve good performance with low-order polynomials of features. Consequently, the SpaRTA algorithm provides a promising approach for interpretable and, possibly, universal identifier models.

1.3. Approach

This thesis constructs algebraic identifier models for the identification of high uncertainty in RANS turbulence models. High uncertainty is identified with error metrics that detect violations of model assumptions in Boussinesq's hypothesis. The model construction uses supervised machine learning algorithms to find interpretable and universal models. For interpretability, Symbolic Regression algorithms are applied to achieve physical insight into the identification process. The two algorithms Logistic Regression and SpaRTA are tested individually. Logistic Regression builds low complexity models which offer direct insight and provide a lower limit to the interpretability of universal identifier models. SpaRTA, on the other hand, allows higher model complexity and, possibly, better performance, but enforces sparsity constraints to promote algebraic models. The universality of models is assessed on a database with challenging flow conditions for RANS models. The database includes a range of Reynolds numbers, geometries and complex physical phenomena.

The thesis is structured as follows. The background for RANS turbulence modelling and supervised Machine learning is introduced in chapter 2. Chapter 3 describes the methodology to construct interpretable identifier models which includes the error metrics, algorithms and model evaluation routines. The database for training and testing cases is presented in chapter 4. In chapter 6, interpretable models for the Logistic Regression and SpaRTA algorithms are selected and analysed. Furthermore, the quantitative and qualitative model performance on the database is discussed. The work is summarised and conclusions drawn in chapter 7.

2. Theory

The present work discusses the application of Machine Learning algorithms to turbulence modelling. The theory includes the fundamental equations for computational fluid dynamics (CFD) and introduces methods from high to low fidelity. Further, turbulence modelling for the RANS equations with the eddy-viscosity approximation is introduced. Additionally, the principles of Machine Learning for data-driven modelling are presented.

2.1. The Navier-Stokes equations

The motion of a viscous Newtonian fluid with constant properties is governed by the Navier-Stokes (NS) equations. The incompressible formulation without body forces is

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{2.1}$$

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial p}{\partial x_i} + \nu \frac{\partial}{\partial x_j}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{2.2}$$

where u is the velocity, p the pressure, ν the kinematic viscosity and ρ the fluid's density. Analytical solutions to the NS equations are only known for simple flows. Instead, the equations are solved with numerical approximations which is referred to as CFD in the literature [22].

The physics of fluid flow are characterised by the Reynolds number Re. The Reynolds number describes the ratio of inertia to viscous forces with

$$Re = \mathcal{U}\mathcal{L}/\nu, \tag{2.3}$$

where $\mathcal{U}$ is a characteristic velocity and $\mathcal{L}$ a characteristic length. When inertia forces are dominant, the flow becomes *turbulent*. Turbulent flows are three-dimensional, time-dependent and chaotic. Consequently, turbulent motion appears on various time and length scales.

For inertia-dominant flows, turbulence is produced on the largest scales. Eddy structures form which hold kinetic energy. The eddies transport their energy to smaller scales through a cascade process. At the smallest scales, viscosity becomes dominant and the kinetic energy of eddies is dissipated. An accurate prediction of all scales of turbulence is challenging for numerical approximations. Either an approximation resolves all scales with a fine discretization or it uses a model to describe part of the scales. Methods with fine discretization and little to no modelling are high-fidelity methods.

High-fidelity methods

A direct solution of the NS equations is called Direct numerical simulation (DNS). A DNS does not involve any modelling, but resolves the dissipating scales of turbulence. As a consequence of fine discretization, the computational costs for DNS are large. The number of grid points N_{DNS} scales with the Reynolds number

approximately as $N \sim Re^3$ [1]. Although computational power increases, direct solutions are limited to low-to-moderate Reynolds numbers.

The computational costs are reduced by modelling the smallest scales and resolving only larger scales. These methods are called Large Eddy simulation (LES). A LES applies a filter to the velocity which separates it into a resolved and unresolved component. The unresolved velocities are modelled with a sub-grid scale model. If a LES resolves most of the small scales, it approaches the accuracy of a DNS [23]. The database for the Machine Learning is based on such well-resolved LES and DNS solutions.

Reynolds-averaged Navier-Stokes

The computational costs of approximations are further reduced by modelling all scales of turbulence. These methods are based on averaging the NS equations in time. The averaging is based on the Reynolds decomposition:

$$u = \overline{u} + u', \qquad p = \overline{p} + p' \tag{2.4}$$

where $\overline{\{\cdot\}}$ indicates the mean and $\{\cdot\}'$ the fluctuating component for both velocity and pressure. Mean values are computed with an ensemble average of the form

$$\overline{u} = \lim_{N\to\infty} \sum_{n=1}^{N} u^{(n)} \tag{2.5}$$

where N is the number of realisations of a quantity that is used for averaging. Substituting the Reynolds decomposition into equation (2.1), the Reynolds-averaged Naiver-Stokes (RANS) equations result

$$\frac{\partial \overline{u}_i}{\partial x_i} = 0 \tag{2.6}$$

$$\frac{\partial \overline{u}_i}{\partial t} + \overline{u}_j \frac{\partial \overline{u}_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \nu \frac{\partial}{\partial x_j}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right) - \frac{\partial \tau_{ij}}{\partial x_j} \tag{2.7}$$

where $\tau_{ij} = \overline{u_i' u_j'}$ is the Reynolds stress tensor.

The Reynolds stress tensor appears when the non-linear convection term in equation (2.1) is averaged. The tensor τ_{ij} describes the influence of turbulent fluctuations on the mean flow. Each component of the tensor is the covariance of two velocity components. Since each of the six components is unknown, the RANS equations are not closed. A turbulence model that describes τ_{ij} is required for closure. A brief overview of turbulence modelling for the Reynolds stress tensor is given in the next section.

2.2. Turbulence modelling for RANS

Several approaches for closure of the RANS equations have been developed [7]. One class of models solves an additional transport equation for each component in the Reynolds stress tensor. These models are called Reynolds stress transport models (RSM). The transport equations allow RSMs to predict challenging effects like streamline curvature, sudden changes in strain rate and body forces. Still, the exact transport equation requires modelling to become solvable. In practise, RSMs have high computational costs compared to other RANS models and lack numerical stability in complex flows [2].

In contrast to RSMs, Eddy-viscosity models (EVM) approximate the complete Reynolds stress tensor. EVMs are based on the eddy-viscosity approximation which constructs a turbulent viscosity in analogy to molecular viscosity. The eddy viscosity ν_t represents averaged turbulence diffusion at a local point in the flow. The viscosity is modelled with a velocity $\mathcal{U}$ and length scale $\mathcal{L}$ which are provided by additional scalar (transport) equations. Well-established models are based on one or two additional equations [2]. For example, the Spalart-Allmaras one-equation [5] or k-ϵ two-equation model [3]. EVMs however do not predict complex flows reliable, because of their underlying model assumptions [1]. This section investigates the assumptions and limitations of EVMs further.

Eddy-viscosity approximation

The following paragraphs are generally based on [1] unless stated otherwise. The eddy-viscosity approximation, or Boussinesq's hypothesis, is used to derive models for the Reynolds stress tensor. The approximation assumes an analogy between molecular and turbulent stresses. The molecular stresses σ_{ij} inside a fluid are decomposed into an isotropic and deviatoric term:

$$\sigma_{ij} = \overline{p}\delta_{ij} - 2\nu S_{ij}, \qquad S_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right), \tag{2.8}$$

where δ_{ij} is the Kronecker delta, μ the dynamic viscosity and S_{ij} the mean-strain-rate tensor. Similarly, the Reynolds stress tensor for a EVM is decomposed:

$$\tau_{ij} = \frac{2}{3}k\delta_{ij} - 2\nu_t S_{ij}, \tag{2.9}$$

where the isotropic term comprises the turbulent kinetic energy

$$k = \tau_{ii}/2. \tag{2.10}$$

The deviatoric term is called the anisotropy tensor a_{ij}. The Boussinesq approximation defines it as

$$a_{ij} = -2\nu_t S_{ij}. \tag{2.11}$$

Hence, the approximation implies two assumptions for a_{ij}. First, it assumes a linear relationship between a_{ij} and S_{ij} and, second, that turbulence only acts as increased diffusion by depending on the local mean velocity gradient $\frac{\partial \overline{u}_i}{\partial x_j}$.

Both assumptions have implications on the validity of the eddy-viscosity approximation. The proportionality constant ν_t in equation (2.11) is a scalar. So, the anisotropy tensor is a scaled mean-strain-rate tensor. Consequently, the principal axes of the a_{ij} and S_{ij} need to be aligned for correct predictions. In practical flows, both principal axes are rarely aligned [6]. Misalignment occurs even in simple shear flows where turbulence prevails while $S_{11} = S_{22} = S_{33} = 0$.

Furthermore, a_{ij} is not solely determined by the local mean velocity gradient $\frac{\partial \overline{u}_i}{\partial x_j}$. This is true for rapid distortions when the turbulent time scale k/ϵ is much larger than the mean shear time scale $\mathcal{S} = \sqrt{2S_{ij}S_{ij}}$. Turbulence behaves like an elastic solid that slowly adapts to straining. As a consequence, anisotropy prevails and slowly decays after rapid changes in S_{ij}, even if $S_{ij} = 0$. This decay cannot be predicted by the approximation in 2.11. Hence, the second assumption and, in general, the Boussinesq eddy-viscosity approximation

is not valid.

Models based on the linear relationship in 2.11 are known as linear eddy viscosity models (LEVM). LEVM are based on the assumption of similarity to molecular stresses. This similarity becomes clear upon substitution of the decomposed Reynolds stress tensor in equation (2.9) and the strain tensor S_{ij} into the RANS momentum equation (2.7):

$$\frac{\partial \overline{u}_i}{\partial t} + \overline{u}_j \frac{\partial \overline{u}_i}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \breve{p}}{\partial x_i} + 2(\nu + \nu_t)\frac{\partial}{\partial x_j}(S_{ij}) \tag{2.12}$$

where the isotropic stress is absorbed into a modified pressure term $\breve{p} = (p - 2/3k)\delta_{ij}$. Thus, the eddy viscosity serves as additional diffusion due to turbulence.

Despite the validity of the eddy-viscosity approximation, LEVMs are used in commercial CFD solvers due to their simplicity and robustness. They achieve reasonable predictions for many attached flows by introducing corrections and tuning the model's coefficients to high-fidelity and experimental data. Still, LEVMs fail to predict important practical flows with separation, recirculation, rotation, curvature and body-forces [6].

Anisotropy

The characteristics of turbulence are described by the properties of the Reynolds stress tensor τ_{ij} and the anisotropy tensor a_{ij}. This section reviews properties of both tensors to introduce and visualise states of turbulence. The Reynolds stress tensor is a second-order symmetric tensor of the mean velocity fluctuations, that is

$$\tau_{ij} = \overline{u_i' u_j'}. \tag{2.13}$$

Physical realizability requires the tensor to have non-negative diagonal components and a non-zero determinant. It obeys

$$\tau_{\alpha\alpha} \geq 0, \quad \alpha = \{1,2,3\}, \qquad \det \tau_{ij} \geq 0, \tag{2.14}$$

where no summation convention applies for Greek indices. Additionally, Schumann [24] shows that the Cauchy-Schwartz inequality must hold

$$\tau_{\alpha\alpha} + \tau_{\beta\beta} = 2|\tau_{\alpha\beta}|, \quad \alpha, \beta = 1,2,3, \quad \alpha \neq \beta. \tag{2.15}$$

Therefore, the Reynolds stress tensor is a symmetric and positive semi-definite rank 2 tensor.

The eigendecomposition of the Reynolds stress tensor is given by

$$\tau_{ij} = 2k\left(\frac{1}{3}\delta_{ij} + V_{ik}\Theta_{kl}V_{jl}\right). \tag{2.16}$$

All eigenvalues θ_i are non-negative by definition. The physical realizability conditions limit the Reynolds stresses. The first condition in 2.14 and the definition of the turbulent kinetic energy in equation (2.10) show that the diagonal terms must lie within $[0, 2k]$. Further, the Cauchy-Schwartz inequality in equation (2.15) limits the off-diagonal components to $[-k, k]$. Accordingly, the eigenvalues θ_i are limited to $[0, 2k]$.

The normalised anisotropy tensor b_{ij} is related to the anisotropy and Reynolds stress tensor by

$$b_{ij} = \frac{a_{ij}}{2k} = \frac{\tau_{ij}}{2k} - \frac{1}{3}\delta_{ij}. \tag{2.17}$$

Turbulence state	Eigenvalues λ_i
One-component (1C)	$\left[\frac{2}{3}, -\frac{1}{3}, -\frac{1}{3}\right]^T$
Axisym. two-component (2C)	$\left[\frac{1}{6}, \frac{1}{6}, -\frac{1}{3}\right]^T$
Three-component (3C)	$[0, 0, 0]^T$
Axisym. expansion	$\lambda_1 \in \left(0, \frac{1}{3}\right), \quad \lambda_2 = \lambda_3 \in \left(-\frac{1}{6}, 0\right)$
Axisym. contraction	$\lambda_1 \in \left(-\frac{1}{3}, 0\right), \quad \lambda_2 = \lambda_3 \in \left(0, \frac{1}{6}\right)$
Two-component limit	$\lambda_1 + \lambda_3 = \frac{1}{3}, \quad \lambda_2 = -\frac{1}{3}$

Table 2.1.: The limiting states of turbulence with corresponding eigenvalues and shape of the turbulence. Axisymmetric is abbreviated with *Axisym.*.

Accordingly, the eigenvalues are related with

$$\lambda_i = \frac{\theta_i}{2k} - \frac{1}{3}\delta_{ij}. \tag{2.18}$$

It is straightforward to show the corresponding limits for the anisotropy tensor and its eigenvalues. The diagonal components of b_{ij} are limited to $\left[-\frac{1}{3}, \frac{2}{3}\right]$, off-diagonals to $\left[-\frac{1}{2}, \frac{1}{2}\right]$ and eigenvalues $\lambda_i \in \left[-\frac{1}{3}, \frac{2}{3}\right]$. Furthermore, the conditions $\lambda_1 \geq \lambda_2 \geq \lambda_3$ and $\sum_i \lambda_i = 0$ must hold [25].

The eigenvalues describe the relative strength of velocity fluctuations, that is the componentality of turbulence. They are used to distinguish states of turbulence. Table 2.1 gives an overview of limiting and intermediate states. Each state is characterised with the eigenvalues of the anisotropy tensor. One-component (1C) turbulence is confined to a line with one active fluctuating component. In axisymmetric two-component (2C) turbulence only one eigenvalue is zero. Turbulent fluctuations exist in two directions and restrict the turbulence to a plane. Three-component (3C) turbulence is isotropic with fluctuations in all directions and all λ_i are equal to zero.

Moreover, intermediate states of turbulence identify the regions between limiting states. The axisymmetric expansion describes the state between 1C and 3C turbulence. Similarly, axisymmetric contractions connect 2C and 3C turbulence. The two-component limit corresponds to the intermediate state of 1C and 2C. Finally, plane-strain appears when at least one λ_i is zero.

The anisotropy tensor and states of turbulence are visualised in two dimensions. Lumley and Newman [26] proposed a non-linear mapping based on the invariants of b_{ij}. The invariants $II = b_{ij}b_{ji}$ and $III = b_{ij}b_{jl}b_{li}$ span a plane where all realizable states are limited to a triangular shape. In contrast, a linear mapping onto the λ_1-λ_2 plane is known as Lumley triangle [27]. Banerjee et al. [25] introduced a barycentric mapping with the eigenvalues λ_i. The eigenvalues are mapped onto a equilateral triangle where each vertex corresponds to one-, two- or three-component turbulence, see figure 2.1a.

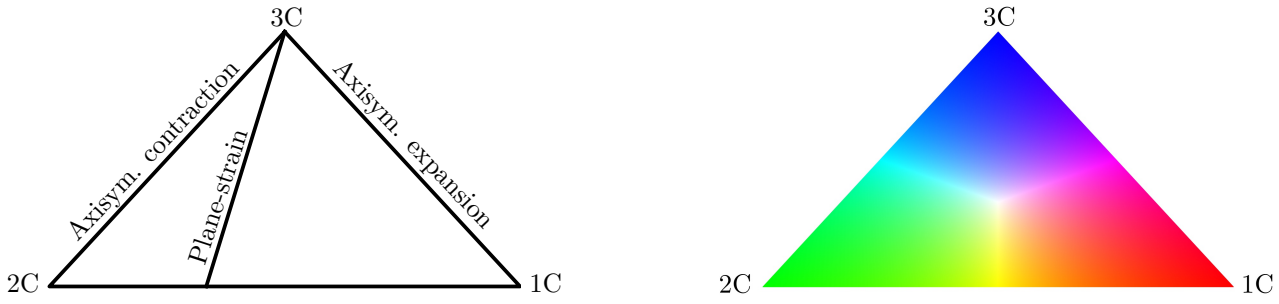
The barycentric map introduces three coordinates C_{1C}, C_{2C}, C_{3C} for a given anisotropy tensor. The coordinates are determined by the eigenvalues of b_{ij}

$$C_{1C} = \lambda_1 - \lambda_2, \tag{2.19}$$

$$C_{2C} = 2(\lambda_2 - \lambda_3), \tag{2.20}$$

$$C_{3C} = 3\lambda_3 + 1. \tag{2.21}$$

Uniqueness of the coordinates is ensured with $\sum_i C_{iC} = 1$. The maximum value of each coordinate is one at



(a) The barycentric map for states of turbulence.

(b) The RGB colormap in barycentric coordinates.

Figure 2.1.: The barycentric map for anisotropic turbulence as proposed by Banerjee et al. [25] with RGB colormap by Emory and Iaccarino [28].

the corresponding limiting state.

The equilateral triangle is visualised on the ξ-η space. Basis points for each vertex are chosen in 2D space as (ξ_{iC}, η_{iC}). Given an anisotropy tensor, the coordinates in the barycentric map are computed with:

$$\xi = C_{1C}\xi_{1C} + C_{2C}\xi_{2C} + C_{3C}\xi_{3C} \tag{2.22}$$

$$\eta = C_{1C}\eta_{1C} + C_{2C}\eta_{2C} + C_{3C}\eta_{3C} \tag{2.23}$$

Figure 2.1a shows the barycentric map with each limiting state of turbulence. Intermediate states lie between the corresponding vertices and are marked at the realizability limit. Further, the plane-strain indicates the path for a two-dimensional RANS solution with a LEVM.

The barycentric map enables an undistorted visualisation of the anisotropy. However, it lacks a straightforward correlation with the physical domain. Moreover, large amounts of data are difficult to analyse in the triangle. Emory and Iaccarino [28] converted the coordinates C_{1C}, C_{2C}, C_{3C} into a Red-Green-Blue (RGB) colormap. The RGB color vector results from a linear combination of the coordinates:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = C_{1C} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + C_{2C} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + C_{3C} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{2.24}$$

Each color vector is normalised with the largest coefficient to increase the brightness compared to the original colormap.

The RGB colormap is projected onto the equilateral triangle in figure 2.1b. Each limiting state is associated with one colour and intermediate states represent blended colours. Figure 2.2 shows LES data for the flow over periodic hills. The RGB colormap highlights 2C turbulence in the proximity of upper and lower walls. Furthermore, the contraction on the leeward hill becomes visible as 1C turbulence.

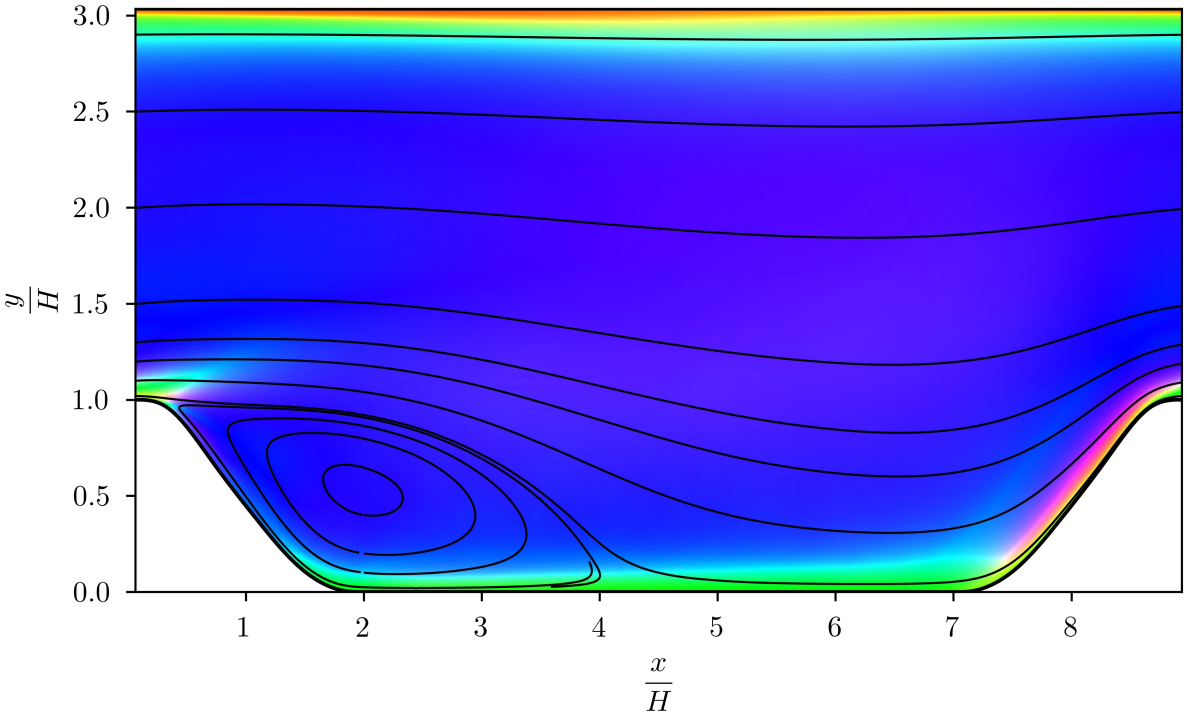


Figure 2.2.: Turbulence states visualised with an RGB colormap on periodic hills with $Re = 10\,595$.

2.3. Machine learning for turbulence modelling

Machine learning techniques find patterns in large amounts of data. The techniques include algorithms that construct models from a sample of the data, referred to as *training* data in the literature [18]. A data-driven model is trained to map from inputs to outputs. The inputs are referred to as *features* which convey information of the data to the model. In the context of turbulence modelling, an algorithm constructs models, for example, high-fidelity simulation data. The models are trained to predict the anisotropy tensor a_{ij} using inputs in the form of mean flow quantities or turbulence statistics. This section introduces distinct types of Machine Learning, methods to train and select models and as well as feature selection.

Machine Learning algorithms derive models from data. The models learn to predict the data from input features q. The input features ideally extract all information about the data set and make it available for a model. The learning process is an optimisation of model coefficients to adapt it to given data. Computational costs are reduced by selecting only the most-informative features out of a set. This section introduces the types of Machine Learning, methods to train and select models and feature selection.

Types of machine learning

Machine Learning approaches are distinguished based on the type of available data and the method by which training data is received.

- *Unsupervised Learning* identifies patterns in input features q from a given data set. It aims to reduce the complexity of data or recognise correlations between inputs.

- *Supervised Learning* derives models which map from inputs q to outputs y. The prediction of discrete categories is a classification task. Whereas, a regression task learns a model for continuous output variables. An example is curve fitting where a polynomial is fitted to given data points by adapting the coefficients.
- *Reinforcement Learning* uses explorative algorithms which discover decision rules for a given set of (subsequent) actions. The trial and error characteristic is used, for example, to optimise decisions in chess.

The present work concentrates on supervised learning with a hybrid classification and regression task. The underlying problem is a binary classification where a class y_0 or class y_1 is predicted. Model predictions for a class y however are continuous on the interval $[0, 1]$.

Model selection

The best model for given data and a Machine Learning algorithm reaches high generalisation. Generalisation is the ability of a model to predict independent test data. During model selection, the data is split into subsets to assess the generalisation. The training subset is given to the algorithm which learns a model. Algorithm hyper-parameters are tuned with a validation subset. Finally, the generalisation error is assessed on a test subset.

A number of models is trained with distinct hyper-parameters. Their *training error* is evaluated on the validation data. The best hyper-parameters are chosen by the lowest training error. The training error however cannot predict the generalisation [29]. Rather, it indicates the required model complexity. Model complexity is evaluated by comparing bias and variance, that is the average training error and average variance, respectively. Models of high complexity reach low bias and high variance which is known as *overfitting*. An overfitted model predicts even the noise in the training data and, thus, has high variance on test data. The opposite situation is an underfitted model with high bias and low variance. *Underfitting* is resolved with higher model complexity. Since data inevitably includes noise, the model complexity needs to be adapted to the data set by trading off bias and variance.

The training, validation and test data sets are 50 %, 25 %, 25 % of the full data set, respectively [18]. A validation with three distinct data sets is known as *holdout cross-validation*. Holdout validation fails to use all available data for training. It uses only half of the training data which can result in poor models [30]. An efficient data utilisation is possible with *K-fold Cross-Validation*.

K-fold cross-validation combines training and validation data. The combined data is split into K-folds or subsets. K-validations are performed where a model is trained on $K-1$ folds and tested on the remaining fold. The cross-validation error is determined as the average of K training errors. The most efficient data utilisation is achieved with *leave-one-out cross-validation* (LOOCV) where $K =$ number of data sets. Adversely, the rich data utilisation in LOOCV increases the computational costs [30].

The best hyper-parameters for a given algorithm are evaluated with either a random or deterministic search on the parameter space. The deterministic approach searches along a defined parameter grid which is costly for large numbers of parameters. In contrast, random search samples parameters from distributions to identify relevant dimensions in parameter space. The random approach is hence more efficient in computational resources than a grid search [31].

Feature selection

The features are distinguished as either informative, non-informative or redundant. For a given problem, features are often chosen with domain knowledge to exclude non-informative and redundant features. In complex problems, the types of features are hard to distinguish. Instead, *feature selection* methods are used to identify the value of each feature for a classification task. Tang et al. [32] shows, that removing redundant and non-informative features avoids overfitting and reduces computational costs.

Feature selection evaluates the correlation of a feature q with respect to a target y. Subsequently, low correlation features are excluded from the feature set. The evaluation is based on either distance or information measures [33]. Distance measures determine the separability between q and y. Separability describes whether q and y are easily distinguished in the q-y plane. An example is the euclidean distance $\Delta = \|q - y\|_2$. On the other hand, an information measure is the entropy H. The entropy for a continuous random variable X is defined as

$$H(x) = -\int \Pr(x_i) \log \Pr(x_i). \tag{2.25}$$

It determines the expected information in X. In feature selection, the information required for q and y is determined with the joint entropy

$$H(q, y) = H(y \mid q) + H(q), \tag{2.26}$$

where $H(y \mid q)$ is the conditional entropy, that is the information about the target y by knowing the feature q [17].

Two types of feature selection are filter and wrapper methods. Filter methods use a uni- or multivariate measure to reduce the set of features. These methods cannot detect dependencies between features, but are scalable to large data sets. Wrapper methods train models with all or a subset of features. The correlation for each feature is evaluated with the magnitude of model coefficients. Compared to filter methods, wrapper selection includes feature and model dependencies. Their computational costs, however, require smaller sets of features [32].

2.4. Symbolic Regression algorithms

Symbolic Regression (SR) attempts to discover simple algebraic models. Algorithms search a model space which is constructed from basis functions f and inputs q. There are stochastic and deterministic approaches for SR. Stochastic approaches generate a population of models where each model is iteratively evolved with mutation and recombination operators. The operators originate from Darwin's theory of biological evolution, thus, the name genetic algorithms [16]. Differently, a deterministic algorithm uses a defined library of models which is generated from the model space. Although the library limits exploration of the model space, the cost of deterministic algorithms are lower compared to genetic algorithms [34].

The present work searches models for a binary classification problem. Two deterministic algorithms are compared, because of their ability to scale to a large database. Logistic Regression is a standard algorithm for classification [17]. It learns linear models from features q and uses regularisation methods to adapt model complexity. The second algorithm is Sparse Regression of Turbulent Stress Anisotropy (SpaRTA). SpaRTA is developed by Schmelzer et al. [19] for applications in data-driven turbulence modelling. The algorithm identifies model structures in a first step and, subsequently, infers model coefficients.

3. Methodology

The present work investigates the identification of high uncertainty in Reynolds-averaged Navier-Stokes (RANS) solutions using data-driven models. The identifier model M is constructed with supervised Machine Learning algorithms. The algorithms train a model to map from input features q to targets y. The input features form the structure of the model and convey information about the physics of the flow. Targets are binary marker which identify high uncertainty using error metrics.

This chapter introduces the error metrics y, input features q and the machine learning algorithms. Furthermore, feature selection, undersampling and performance metrics are discussed. Feature selection is a filter method that reduces the set of features based on their amount of information. Undersampling allows an effective utilisation of the large database. The performance metrics serve for model evaluation.

3.1. Error metrics

This section relies on the theory of RANS turbulence modelling introduced in section 2.2. The objective of the identifier is to predict violations of the eddy-viscosity approximation. Ling and Templeton [12] presented three error metrics which detect either a negative eddy-viscosity, anisotropic turbulence and non-linearity in ν_t. Each metric is derived from local mean flow quantities and gives a binary response for each grid point.

The metric for non-linearity is excluded in this work. It determines the difference between the linear and a non-linear relationship for a_{ij}. The non-linear eddy-viscosity model (NLEVM) by Craft et al. is used for the comparison [12]. The model is designed and calibrated to estimate a_{ij} on flows with swirl and strong streamline curvature [35]. Since the database does not include corresponding flows, see chapter 4, the metric is not used.

Ling and Templeton set the error metrics inactive for points with low turbulence intensity k, because the flow becomes laminar. Since the eddy-viscosity ν_t, in an eddy-viscosity model (EVM), depends on the turbulent kinetic energy, ν_t becomes negligible for small values of k. Therefore, grid points with $k < 0.1\,\%$ are not informative for the identifier model. Consequently, error metrics are deactivated for low k.

Non-negativity metric

The eddy-viscosity approximation defines ν_t by analogy to the molecular viscosity ν which is a physical property. Since a physical property is non-negative, the approximation assumes ν_t to be non-negative. Pope [1] describes this assumption as invalid for complex flows, for example in separated flows, cf. section 2.2. Therefore, the non-negativity error metric y_{ν_t} detects violations with

$$y_{\nu_t} = \begin{cases} 1 & \text{if } \nu_t < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

The eddy-viscosity is extracted from high-fidelity data by contracting equation (2.11) to

$$\nu_t = \frac{-\tau_{ij}S_{ij} + \frac{2}{3}k\delta_{ij}}{2S_{kl}S_{kl}}. \tag{3.2}$$

Anisotropy metric

Linear eddy-viscosity models (LEVM) cannot correctly predict anisotropic turbulence, because of the assumptions in *Boussinesq's hypothesis*, see section 2.2. Ling and Templeton proposed a metric to detect anisotropic turbulence using the second invariant of the normalised anisotropy tensor II. The invariant is defined by the contraction $II = b_{ij}b_{ji}$. It ranges from 0 in isotropic turbulence to $2/3$ in the one-component limiting state of turbulence. The error metric is activated at the two-component limiting state of turbulence which corresponds to $II > \frac{1}{6}$, that is

$$y_{II} = \begin{cases} 1 & \text{if } II < \frac{1}{6}, \\ 0 & \text{otherwise.} \end{cases} \tag{3.3}$$

A straightforward interpretation of the anisotropy metric is possible by projecting it onto the barycentric map, see section 2.2. The barycentric coordinates C_{1C}, C_{2C} and C_{3C} are related to the tensor invariants II and III via the eigenvalues λ_i of b_{ij}. In principal components, the tensor invariants and its eigenvalues are related by

$$II = 2\left(\lambda_1^2 + \lambda_1\lambda_2 + \lambda_2^2\right), \tag{3.4}$$

$$III = -\lambda_1\lambda_2(\lambda_1 + \lambda_2). \tag{3.5}$$

The invariants are expressed in barycentric coordinates with equations (2.19) to (2.21):

$$II = \frac{2}{3}C_{1C}^2 + \frac{1}{3}C_{1C}C_{2C} + \frac{1}{6}C_{2C}^2, \tag{3.6}$$

$$III = \frac{2}{9}C_{1C}^3 + \frac{1}{6}C_{1C}^2C_{2C} - \frac{1}{12}C_{1C}C_{2C}^2 - \frac{1}{36}C_{2C}^3. \tag{3.7}$$

The transformed error metric is depicted in figure 3.1. The RGB coloured area corresponds to an inactive metric. Whereas, the white area is detected. The detection covers the one-component limiting state which often occurs in plane channel flow in the near-wall region, because of wall blocking, cf. [1] or figure 4.1b. Further, the error metric partially detects intermediate states of turbulence along the axisymmetric expansion and the two-component limiting state. Still, the anisotropy identification misses the axisymmetric two-component state which is found in the post-reattachment region for separated flows, see figures 4.2 to 4.4. Compared to two-dimensional RANS, it cannot predict all deviations from the plane-strain state.

3.2. Features

The features convey information of the mean flow to the machine learning algorithms. The algorithms construct a model from these features. This work uses two sets of features.

- Features based on physical intuition [12], [15].

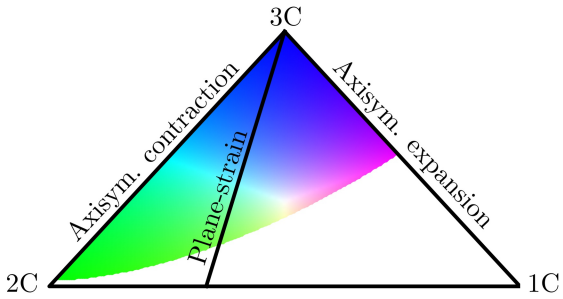


Figure 3.1.: The anisotropy metric visualised in barycentric coordinates and with RGB colormap, see section 2.2. The anisotropy metric y_{II} is active for the states of turbulence in the white area and inactive for the coloured area in the triangle.

- Invariants based on a set of tensors [9].

Each feature is constructed from local and mean flow quantities which are available from RANS solutions. Although high-fidelity data is provided, the models are designed to work within computational fluid dynamics (CFD) solvers where all inputs are based on RANS data.

The inputs for features are mean velocity $\overline{u}$ and pressure $\overline{p}$, molecular viscosity ν, turbulent kinetic energy k, dissipation rate ϵ, eddy viscosity ν_t and distance to the nearest wall d. Although the distance to the nearest wall is not a local quantity, it is often used for RANS turbulence models, cf. [7].

Properties of the features

The features should be non-dimensional, rotationally invariant and Galilean invariant. Non-dimensionality allows for extrapolation of the model to different flows. It mitigates the effect of varying magnitudes of values on the model outcome. Additionally, Wiesler and Ney [36] found that normalisation preconditions the optimisation problem for log-linear loss functions and improves its performance. The following normalisation procedure from [12] is adopted

$$\hat{q} = \frac{q}{|q| + |q^\star|} \tag{3.8}$$

where $\hat{q}$ is the normalised feature and $q^\star$ and q are the normalisation factor and raw feature, respectively. The procedure normalises features onto the range $[-1, 1]$. Raw features q and normalisation factors $q^\star$ are given in table 3.1.

In [13], features that obey rotational and Galilean invariance show improved model performance and generalisation for data-driven models. Rotational invariance requires each feature to be invariant under transformations with the special orthogonal group $SO(3)$. Given orthogonal matrices $\mathbf{Q} \in SO(3)$, rotational invariance requires

$$q(\mathbf{T}, \mathbf{v}, c) = q(\mathbf{Q}\mathbf{T}\mathbf{Q}^T, \mathbf{Q}\mathbf{v}, c) \tag{3.9}$$

for a feature based on an arbitrary tensor $\mathbf{T}$, vector $\mathbf{v}$ and scalar c argument. Rotational invariance is achieved by using only scalars or invariants of tensors and vectors as features.

Galilean invariance requires a system to be invariant under constant velocity offsets. Traditional and data-driven models for turbulence require Galilean invariance to achieve the invariance properties of the NS

equations. Otherwise, models are qualitatively incorrect [1]. The velocity $\mathbf{u}$ is not Galilean invariant, since it depends on the total value. However, the gradients $\nabla\mathbf{u}$, ∇p and ∇k obey the invariance, because they are independent of the reference value. Similarly, the derived strain $\mathbf{S}$ and vorticity tensor $\mathbf{\Omega}$ are Galilean invariant. A subset of the features uses the velocity $\mathbf{u}$ and do not obey the invariance, see table 3.1. It should be considered, that these features limit the generalisability of the model.

Physical features

The first set of features is based on domain knowledge in turbulence modelling. These features are adopted from [12] and [15]. Ling and Templeton [12] first introduced 12 non-dimensional features based on physical intuition.

Wang et al. [15] used a feature for streamline curvature which describes the misalignment of the velocity vector. The feature becomes informative when the flow deviates from parallel shear flow, for example inside the recirculation region. Its normalisation factor is based on the characteristic length L_c of the given flow, see chapter 4 for definitions.

In the present work, the vortex stretching feature and a feature for the comparison of Reynolds stress tensors for a LEVM and NLEVM are excluded. The vortex stretching requires three-dimensional data to determine the vorticity. The database however includes only two-dimensional data. The comparison of Reynolds stress tensors τ_{ij} is designed for the non-linearity metric which is not investigated, see section 3.1.

Additional features have to be excluded, because models are trained and tested on high-fidelity data without using RANS data. If RANS data is included, the identifier learns to map from the erroneous mean flow and turbulence statistics to the true error metric. Since only high-fidelity data is included, a model easily learns a metric if the true data is provided. For example, given the true ν_t, the model must only predict the sign correctly. As a consequence, the viscosity ratio feature q_5 is excluded for the non-negativity metric y_{ν_t}.

Further, features using the Reynolds stress tensor τ_{ij} are excluded for both error metrics. The tensor is unknown in a RANS solver before a turbulence model is used. Since the identifier influences the choice of turbulence model, features using the tensor are excluded. Additionally, τ_{ij} is directly related to the anisotropy tensor a_{ij}, cf. equation (2.11). Thus, a prediction of the anisotropy metric is straightforward when high-fidelity data is provided.

The turbulent kinetic energy k and dissipation rate ϵ are directly related to the eddy-viscosity ν_t used in EVM suggesting the non-negativity metric is easy to learn with high-fidelity data. Wilcox [7] states, that the relationship is based on dimensional analysis and assumes that ν_t depends only on turbulence quantities. However, the true eddy viscosity, which is extracted from high-fidelity data with equation (3.2), does not strictly depend on k and ϵ. Also, both quantities are, approximately, known in a CFD solver. Therefore, the features derived from k and ϵ are included for the non-negativity error metric.

The complete list of eleven features is presented in table 3.1. The features include the Q-criterion which is defined with the strain tensor $\mathbf{S}$ and vorticity tensor $\mathbf{\Omega}$. It is used in CFD applications for post-processing to visualise vortex structures [37]. The wall-distance Reynolds number Re_d identifies the near-wall region where a traditional turbulence model uses damping functions [7]. The feature is assumed to be informative for anisotropic turbulence which appears in the near-wall region due to wall blocking. Further, [11] proposed a marker function, that identifies derivations from parallel shear flow. It is defined as non-orthogonality between the velocity vector and its gradient.

Symbol	Description	Raw feature q	Normalisation factor $q^\star$
q_1	Q-criterion	$\frac{1}{2}\left(\|\mathbf{\Omega}\|^2 - \|\mathbf{S}\|^2\right)$	$\|\mathbf{S}\|^2$
q_2	Turbulence intensity	k	$\nu\|\mathbf{S}\|$
q_3	Wall-distance Reynolds number	$\min\left(\frac{\sqrt{k}d}{50\nu}, 2\right)$	-
q_4	Ratio of turbulence time scale to strain	k/ϵ	$1/\|\mathbf{S}\|$
q_5	Viscosity ratio	ν_t	100ν
q_6	Deviation from parallel shear flow [11]	$\left\|\overline{u}_i\overline{u}_j\frac{\partial \overline{u}_i}{\partial x_j}\right\|$	$\sqrt{\overline{u}_l\overline{u}_l\overline{u}_i\frac{\partial \overline{u}_i}{\partial x_j}\overline{u}_k\frac{\partial \overline{u}_k}{\partial x_j}}$
q_7	Streamline curvature	$\|D\mathbf{\Gamma}/Ds\|$	$1/L_c$
q_8	Pressure gradient along streamline	$\overline{u}_k\frac{\partial \overline{p}}{\partial x_k}$	$\sqrt{\frac{\partial \overline{p}}{\partial x_j}\frac{\partial \overline{p}}{\partial x_j}\overline{u}_i\overline{u}_i}$
q_9	Ratio of pressure normal to shear stress	$\sqrt{\frac{\partial \overline{p}}{\partial x_i}\frac{\partial \overline{p}}{\partial x_i}}$	$\frac{1}{2}\rho\frac{\partial \overline{u}_k^2}{\partial x_k}$

Table 3.1.: Physical features based on work from Ling and Templeton [12] and Wang et al. [15]. The raw feature for streamline curvature q_9 is defined with $\mathbf{\Gamma} = \overline{\mathbf{u}}/|\overline{\mathbf{u}}|$ and $Ds = |\overline{\mathbf{u}}|Dt$.

Description	Raw tensor $\mathcal{T}_i$	Normalisation factor $\mathcal{T}_i^\star$
Strain tensor	$\mathbf{S}$	ϵ/k
Vorticity tensor	$\mathbf{\Omega}$	$\|\mathbf{\Omega}\|$
Pressure gradient	∇p	$\rho\|D\mathbf{u}/Dt\|$
Turbulence intensity gradient	∇k	$\epsilon/\sqrt{k}$

Table 3.2.: Raw mean flow tensors $\mathcal{T}_i$ for the invariant features with normalisation factors $\mathcal{T}_i^\star$. Normalisation according to equation (3.8).

Invariant features

The second set of features comprises tensor invariants. The invariants are constructed from a finite tensorial set $\mathcal{T}$ and a required symmetry property, here rotational invariance. Ling et al. [13] introduced a systematic procedure to construct invariants based on the set $\mathcal{T} = \{\mathbf{S}, \mathbf{\Omega}\}$. This initial set of tensors assumes that the turbulence is described by the strain tensor $\mathbf{S}$ and vorticity tensor $\mathbf{\Omega}$. In [9] the pressure gradient $\nabla\overline{p}$ and turbulence intensity gradient ∇k are added. The additional tensors should include additional physics into the invariant features. On the one hand, the strain and vorticity tensor do not account for pressure gradients. Strong pressure gradients however stabilise (favourable) or destabilise (adverse) turbulence, cf. [6]. On the other hand, [9] includes ∇k to account for non-local processes. For example, the effects of strong convection and diffusion in separated flows. Including both gradients leads to the tensorial set $\mathcal{T} = \{\mathbf{S}, \mathbf{\Omega}, \nabla p, \nabla k\}$.

Each tensor is normalised with the factors given in table 3.2 and with the procedure used for the physical features, see equation (3.8). The normalisation is adopted from Wu et al. They showed that each factor is Galilean invariant [9].

(n_S, n_A)	Feature index	Invariant bases
$(1,0)$	q_{10}-q_{11}	$\hat{\mathbf{S}}^2$, $\hat{\mathbf{S}}^3$
$(0,1)$	q_{12}-q_{14}	$\hat{\mathbf{\Omega}}^2$, $\hat{\mathbf{P}}^2$, $\hat{\mathbf{K}}^2$
$(1,1)$	q_{15}-q_{23}	$\hat{\mathbf{\Omega}}^2\hat{\mathbf{S}}$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{S}}^2$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{S}}\hat{\mathbf{\Omega}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{P}}^2\hat{\mathbf{S}}$, $\hat{\mathbf{P}}^2\hat{\mathbf{S}}^2$, $\hat{\mathbf{P}}^2\hat{\mathbf{S}}\hat{\mathbf{P}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{K}}^2\hat{\mathbf{S}}$, $\hat{\mathbf{K}}^2\hat{\mathbf{S}}^2$, $\hat{\mathbf{K}}^2\hat{\mathbf{S}}\hat{\mathbf{K}}\hat{\mathbf{S}}^2$
$(0,2)$	q_{24}-q_{26}	$\hat{\mathbf{\Omega}}\hat{\mathbf{P}}$, $\hat{\mathbf{P}}\hat{\mathbf{K}}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{K}}$
$(1,2)$	q_{27}-q_{50}	$\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{S}}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{P}}\hat{\mathbf{S}}^\star$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{P}}\hat{\mathbf{S}}^{2\star}$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{S}}\hat{\mathbf{P}}\hat{\mathbf{S}}^{2\star}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{K}}\hat{\mathbf{S}}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{K}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{K}}\hat{\mathbf{S}}^\star$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{K}}\hat{\mathbf{S}}^{2\star}$, $\hat{\mathbf{\Omega}}^2\hat{\mathbf{S}}\hat{\mathbf{K}}\hat{\mathbf{S}}^{2\star}$, $\hat{\mathbf{P}}\hat{\mathbf{K}}\hat{\mathbf{S}}$, $\hat{\mathbf{P}}\hat{\mathbf{K}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{P}}^2\hat{\mathbf{K}}\hat{\mathbf{S}}^\star$, $\hat{\mathbf{P}}^2\hat{\mathbf{K}}\hat{\mathbf{S}}^{2\star}$, $\hat{\mathbf{P}}^2\hat{\mathbf{S}}\hat{\mathbf{K}}\hat{\mathbf{S}}^{2\star}$
$(0,3)$	q_{51}	$\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{K}}$
$(1,3)$	q_{52}-q_{56}	$\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{K}}\hat{\mathbf{S}}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{K}}\hat{\mathbf{P}}\hat{\mathbf{S}}$, $\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{K}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{\Omega}}\hat{\mathbf{K}}\hat{\mathbf{P}}\hat{\mathbf{S}}^2$, $\hat{\mathbf{\Omega}}\hat{\mathbf{P}}\hat{\mathbf{S}}\hat{\mathbf{K}}\hat{\mathbf{S}}^3$

Table 3.3.: The minimum integrity basis for the tensorial set $\mathcal{T}$ with symmetric tensor $\mathbf{S}$ and antisymmetric tensors $\mathbf{\Omega}$, $\mathbf{P}$ and $\mathbf{K}$. The invariant bases are the trace of matrix products of each tensor. Traces are not indicated. The number of symmetric tensors n_S and the number of asymmetric tensors n_A indicates the possible combinations for each base. The $\{\hat{\cdot}\}$ indicates a normalised tensor and the asterisk (*) indicates cyclic permutation of the anisotropic tensor.

The construction of invariants requires the tensorial set $\mathcal{T}$ to contain only second-order tensors. The gradients of pressure and turbulence intensity are transformed into anti-symmetric tensors with

$$\mathbf{P} = -\mathbf{I} \times \nabla p, \qquad \mathbf{K} = -\mathbf{I} \times \nabla k \tag{3.10}$$

Both tensors $\mathbf{P}$ and $\mathbf{K}$ are pseudo-tensors. Thus, they do not obey reflection invariance and reduce the generalisability of the identifier model.

The minimal integrity basis for $\mathcal{T}$ and the symmetry group $SO(3)$ (rotational invariance) includes all polynomial invariants. The polynomial invariants are the traces of all independent matrix products formed from the tensorial set $\mathcal{T}$. The number of these invariants is finite according to Hilbert's basis theorem. Further, the Cayley-Hamilton theorem shows that 47 independent invariants exist for $\mathcal{T}$ [9]. The bases of the polynomial invariants originate from [38] and are listed in table 3.3. They are referred to as invariant features in the present work.

3.3. Machine Learning algorithms

The models M map from the features q to the error metrics y. This mapping classifies each point in the feature space as either active $y = 1$ or inactive $y = 0$. The decision between active and inactive is described by the model. The model describes a decision boundary, that is a N_q-dimensional hyper-plane, in the feature space, see [18]. Therefore, models have the structure

$$M(\mathbf{q}) = \mathbf{w}^T\mathbf{q}, \tag{3.11}$$

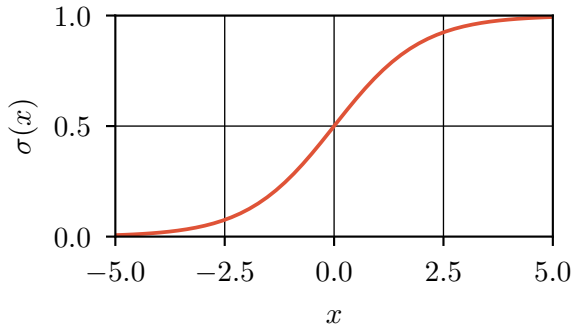


Figure 3.2.: The logistic function models the probability of a class for given inputs $\mathbf{w}^T\mathbf{q}$. The prediction is smooth and limited within the range $[0, 1]$.

where $\mathbf{q}$ is the vector of all features, $\mathbf{w}$ are the corresponding model coefficients. The decision boundary is limited to discrete predictions of $y = 0$ or $y = 1$. An equivalent representation is the probability of an active error metric given a feature vector, that is $\Pr(y = 1|\mathbf{q})$, see [17]. The probability of $\Pr(y = 1)$ is modelled using the logistic function

$$\sigma(\mathbf{w}^T\mathbf{q}) = \frac{1}{1 + e^{-\mathbf{w}^T\mathbf{q}}} = \Pr(y = 1|\mathbf{q}). \tag{3.12}$$

The S-shaped logistic function is depicted in figure 3.2. It models the probability of the active class by mapping from $\mathbf{q}$ onto the range $[0, 1]$. The σ function ensures a smooth prediction of the error metrics which aids convergence in a CFD solver.

The models M, however, give only predictions $\tilde{y}$ of the true state of an error metric y. Model predictions are optimised by adapting the model coefficients $\mathbf{w}$ to samples. The loss function L quantifies erroneous predictions, that is the misfit of the model. This section introduces the loss function that is optimised to find model coefficients. Next, the algorithms that solve the optimisation problem for model coefficients are introduced.

Loss function

The loss function is based on the Logistic Regression algorithm [17]. Logistic Regression fits the model coefficients by maximising the likelihood. For a binary metric, the likelihood is based on a Bernoulli distribution. Logistic Regression optimises the log-likelihood defined with

$$\begin{aligned} \log L(\mathbf{w}) &= \sum_{i=1}^{N} \{y_i \log \Pr(\mathbf{q}_i; \mathbf{w}) + (1 - y_i) \log(1 - \Pr(\mathbf{q}_i; \mathbf{w}))\}, \\ &= \sum_{i=1}^{N} \left\{ y_i \mathbf{w}^T\mathbf{q} - \log(\exp(\mathbf{w}^T\mathbf{q}_i) + 1) \right\}. \end{aligned} \tag{3.13}$$

The function is minimised with the stochastic gradient algorithm SAGA [39]. A stochastic gradient algorithm updates the gradient in each iteration based on a uniformly sampled point from all N data points. The SAGA applies the same concept, but uses an averaged gradient for the descent. The gradient is averaged over previously determined gradients. After computing initial gradients for each direction, gradient updates are performed for sampled data point. SAGA outperforms deterministic methods for large N, because it does not

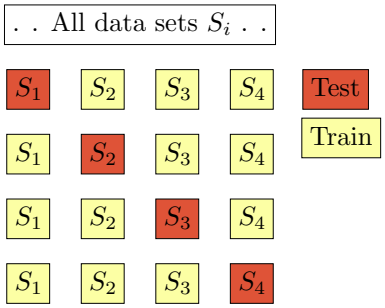


Figure 3.3.: Leave-one-out cross-validation for a total of four datasets.

evaluate the gradient for all samples at each step. Additionally, SAGA allows to use elastic net regularisation.

Elastic net is a hybrid regularisation method that combines Lasso and Ridge regularisation [40]. Lasso regularisation penalises the loss function with $L_1 = \alpha_{L_1}\|\mathbf{w}\|_1$. The L1 norm promotes sparser model structures and automatically selects features. Lasso, however, cannot handle correlated features and will choose one of a group of correlated features at random. Ridge methods penalise with $L_2 = \alpha_{L_2}\|\mathbf{w}\|_2$. This penalty allows to keep correlated features and shrink their coefficients towards each other with increasing α_{L_2} [41].

Elastic net regularisation combines both penalties, L_1 and L_2, and weights them with the elastic net ratio ϕ. Therefore, elastic net selects features automatically and shrinks coefficients to promote parsimonious models, but encourages correlated features [40]. The hyper-parameter $\phi = 1$ corresponds to L_1 regularisation and $0 < \phi <= 1$ increases the weight of L_2 regularisation while reducing Lasso. Adding the penalties to equation (3.13), the penalised loss function becomes

$$\mathbf{w} = \arg\min_{\mathbf{w}} \frac{1-\phi}{2}\|\mathbf{w}\|_2 + \phi\|\mathbf{w}\|_1 - R\sum_{i=1}^{N}\left\{y_i\mathbf{w}^T\mathbf{q} - \log(\exp(\mathbf{w}^T\mathbf{q}_i)+1)\right\}, \tag{3.14}$$

where R controls the regularisation strength. Therefore, both hyper-parameters, R and ϕ, influence the model complexity and performance.

Logistic regression

The algorithm constructs algebraic models with linear features q. The basis of 59 physical and invariant features are supplied, see section 3.2. Using elastic net regularisation, the complexity and performance of models is adjusted. The best hyper-parameters for the loss function are searched on a grid of (R, ϕ) where an optimisation is solved for each grid point (R_i, ϕ_j). The grid search subsequently performs cross-validations for the expected performance of a single grid point.

Leave-one-out cross-validation finds the expected performance by training and testing a model on a subset of the data. figure 3.3 shows the process for four data sets $\mathcal{S}_i$. The available data is split into three sets for training and one for testing. Each data set serves once for testing and else for training. Given four test performances, the expected performance of a model is the mean performance of the tests for given hyper-parameters (R_i, ϕ_j).

The parameter grid is defined with:

$$\phi = [0.01, 0.1, 0.2, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0]^T, \tag{3.15}$$
$$R = [R_{\text{min}}, \dots, R_{\text{upper}}]^T, \tag{3.16}$$

where $R_{\text{min}} = 2/\max(|\mathcal{B}^T\mathbf{y}|)$. The scaling of R ensures equal weight of misfit and regularisation in the loss function for $R = R_{\text{min}}$. All coefficients $\mathbf{w}$ will be equal to zero for $R < R_{\text{min}}$, because the regularisation terms would be dominant in the loss function. The upper end is $R_{\text{upper}} = 10\,000 R_{\text{min}}$ with log-scaled spacing.

The evaluation of each model is based on expected performance, model complexity and generalisation. After cross-validation, one model is fitted to all training data for each pair (R_i, ϕ_j). Therefore, a number of models is derived which are compared against one another to evaluate the best balance of interpretability and performance, see evaluation criteria in section 3.6.

SpaRTA

The discovery of non-linear models uses a modified version of the Sparse Regression of Turbulent Stress Anisotropy (SpaRTA) algorithm from [19]. SpaRTA is originally designed to derive regressors which act as correction terms for the k-ω SST model. It constructs a large library of candidate functions for the model structure and, subsequently, identifies models with sparsity constraint. In the present work the methodology is adapted to derive classifiers instead.

[34] describes the algorithm which constructs a library of candidate functions. Each candidate is a polynomial of features q_i which is combined in two steps.

- Monomials of each feature q_i up to a given order are constructed: $\mathcal{B}_1 = [\sqrt{q_i}, q_i^1, q_i^2, \dots]^T$
- Candidates are formed from interactions of monomials of the form: $\mathcal{B}_2 = [q_1 q_2, q_1 q_2^2, q_1^2 q_2^2, \dots]^T$
- Both sets are combined to yield the library: $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$.

The maximum order for the monomials is chosen to 3. Since each candidate serves as a feature in the loss function, the SpaRTA algorithm constructs models of higher complexity compared to logistic regression.

The discovery of identifier models is performed a model selection and, subsequently, model inference. The model selection identifies active candidates and excludes inactive candidates from the model structure. The selection follows the idea of parsimonious models which avoid overfitting and are interpretable for a modeller [42].

The process of model selection uses the loss function with elastic net regularisation, that is equation (3.14). The elastic net promotes sparsity using the L_1 norm and shrinks correlated coefficients through the L_2 norm. SpaRTA uses a grid of hyper-parameters (R, ϕ) to identify coefficients for each candidate. The range for both parameters is defined analogous to logistic regression with

$$\phi = [0.01, 0.1, 0.2, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0]^T, \tag{3.17}$$
$$R = [R_{\text{min}}, \dots, R_{\text{upper}}]^T, \tag{3.18}$$

see section 3.3.

The grid search yields a model with zero and non-zero coefficients for each pair (R_i, ϕ_j). Candidates with zero coefficients are removed to reduce the model complexity. Therefore, the model selection discovers a number of model structures with varying complexity.

The second step infers model coefficients for each model structure. Since the active candidates are already selected, the L_1 regularisation is not required in the model inference. The L_2 norm however shrinks correlated model coefficients without removing candidates. Hence, the loss function in equation (3.14) is modified to include only the ridge regularisation:

$$min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2 - R\sum_{i=1}^{N}\left\{y_i \mathbf{w}^T\mathbf{q} - \log(\exp(\mathbf{w}^T\mathbf{q}) + 1)\right\}. \tag{3.19}$$

The elastic net ratio ϕ is not required. Consequently, the hyper-parameters reduce to the loss weight R. R is chosen to within $[10, 100]$.

In [34], the computational complexity of Fast function extraction (FFX), the base algorithm for SpaRTA, is analysed. The overall complexity is largest in the model selection step with $\mathcal{O}(NN_q^4)$ where N is the number of samples and N_q the number of features. The number of features is $N_q = 56$, thus, $N_q^4 \approx 9\,840\,000$ and the number of samples is $N \sim 1\,000\,000$. Thus, the algorithm is infeasible for this number of features and a feature selection is necessary to reduce N_q. The feature selection is introduced in the following section.

3.4. Feature selection

The present work uses mutual information to filter the feature set a priori. Mutual information MI measures the dependency between two random variables, in this context, feature q and target y. Features that show little to no dependency to the y are discarded to reduce computational costs and reduce model complexity.

The dependency is measured by comparing the joint distribution $\Pr(q, y)$ to the product of the marginal distributions $\Pr(q)\Pr(y)$. Both are equivalent for strictly independent variables which translates to zero mutual information. For related variables q and y, mutual information is greater than zero, that is:

$$MI(q, y) = \begin{cases} 0 & \text{if independent,} \\ \geq 0 & \text{otherwise.} \end{cases} \tag{3.20}$$

Therefore, mutual information is defined with:

$$MI(q, y) = -\int_q \int_y \Pr(q, y) \ln \frac{\Pr(q, y)}{\Pr(q)\Pr(y)} dy dq. \tag{3.21}$$

The information measure requires joint and marginal distributions to evaluate feature and error metric dependencies.

Density approximation are possible with histogram methods, kernel density estimates (KDE) or k-Nearest-Neighbours (kNN).

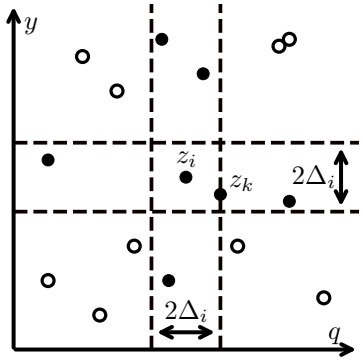


Figure 3.4.: Determination of nearest neighbour with k=1. Points within distance Δ_i, that is $n_x = 3$ in x and $n_y = 3$ in y, are marked as full points. Figure adapted from [44].

- **Histogram**-based methods collect samples z_i within bins of given width. The probability for each bin is determined with the relative frequency for the occupancy of each bin.
- **KDE** methods approximate the distribution with sample-centred kernels which estimate weighted distances. Kernels weight each surrounding sample based on its distance to the reference sample. The weighting depends on the chosen kernel, typically Gaussian, and a bandwidth parameter [43].
- **kNN** determines the distance to the kth nearest sample from a reference sample. Subsequently, joint distributions are estimated from the points within half the distance to the nearest neighbour. Estimations of marginal distributions are equivalently made from points within the one-dimensional distance [44].

Papana and Kugiumtzis [45] compare the three methods for non-linear, chaotic and noisy data. While all methods require proper configuration of the free parameter, kNN shows better computational performance for fine partitioned data. The distance measure in kNN adapts locally to the data structure. Hence, the local data structure is used effectively and computational costs are reduced. Additionally, the k parameter in kNN is not significantly effected by noise. Whereas, the bin size in histograms requires small bins and, thus, more computational resources, for noisy data. Also, Walters-Williams and Li [46] find that kNN performs better for Gaussian distributions when compared to KDE. Therefore, the kNN method is chosen for the present work.

The kNN algorithms follows the proposal in [44]. The algorithms estimates the joint and marginal distributions from a neighbourhood of points $Z(q, y)$. Given a point z_i in the q-y space, kNN estimates the distance $\Delta_i = \|z_i - z_k\|$ to the kth nearest point z_k. figure 3.4 depicts the q-y space (with continuous y). The distance from z_i to the $k = 1$ nearest point is determined with the maximum norm, that is

$$\Delta_i = \max\left(\|q_i - q_k\|, \|y_i - y_k\|\right). \tag{3.22}$$

Consequently, the area including all nearest neighbours is a square and not a disc (L2 norm). The joint distribution is approximated from all points within the square with length $2\Delta_i$. Marginals are based on all points within $\pm\Delta_i$ for q_i and y_i, respectively.

Knowing the k nearest neighbour, kNN approximates probability distributions with the entropy $H(q)$, see section 2.3 for a definition of H. The mutual information definition from equation (3.21) is restated in terms

of the entropy with

$$MI(q,y) = H(q) - H(y) + H(q,y), \tag{3.23}$$

where $H(q,y)$ is the joint entropy [17]. The definition suggests that mutual information measures the reduction of uncertainty in the error metric y by knowing the feature q, that is it indicates how much information q conveys about y (or vice versa) [44].

The entropy for a marginal distribution is approximated with

$$\hat{H}(X) = -N^{-1} \sum_{i}^{N} \log \Pr(x_i), \tag{3.24}$$

where $\hat{H}$ is the approximated entropy, X is a random variable, N is the number of data points and $\Pr(x_i)$ is the marginal probability distribution for a realisation x_i. In the context of this work, X is either q or y. The approximation uses the square with length Δ_i and assumes a uniform probability mass function $\Pr_m(\Delta_i)$ across the square. The expectation of the mass is described with the digamma function $\psi(x)$ by

$$E[log \Pr_m] = \psi(k) - \psi(N). \tag{3.25}$$

Assuming a constant distribution $\Pr(x_i)$ over Δ_i, $\Pr_m$ is related to the distribution with

$$\Pr_m(\Delta_i) \approx \Delta_i^d \Pr(x_i). \tag{3.26}$$

The dimension of x is d. Therefore, the log distribution of $\Pr(x_i)$ is constructed from the square to find the approximate entropy

$$\hat{H}(X) = -\psi(k) + \psi(N) + \frac{d}{N} \sum_{i=1}^{N} \log \Delta_i \tag{3.27}$$

The derivation of the joint entropy $H(X,W)$ is straightforward and leads to

$$\hat{H}(X,W) = -\psi(k) + \psi(N) + \frac{d_X + d_W}{N} \sum_{i=1}^{N} \log \Delta_i \tag{3.28}$$

The definition of the approximated entropy in equations (3.27) to (3.28) does not consider the difference in scales between joint and marginal spaces. The distance to the kth neighbour however will be different in joint space compared to the marginal space, see figure 3.4. Instead, the distance is adapted to each marginal distribution by using the point $n_x(i) + 1$ which always lies on the edge of the square [44]. Substituting these results into equation (3.23), the mutual information is estimated with

$$MI(q,y) = \psi(k) + \psi(N) - \frac{1}{N} \sum_{i=1}^{N} \left[\psi(n_x(i) + 1) + \psi(n_y(i) + 1) \right]. \tag{3.29}$$

Parameter dependencies

The estimation of mutual information depends on the amount of data points N. Further, [45] points out, that the configuration of the k parameter is crucial for kNN density estimates. The change in MI for increasing N is investigated using the ratio of total to normal Reynolds stresses q_8 and the anisotropy error metric, that is $MI(q_8; y_{II})$. Figure 3.5 presents the error in MI compared to the complete-data MI, that is

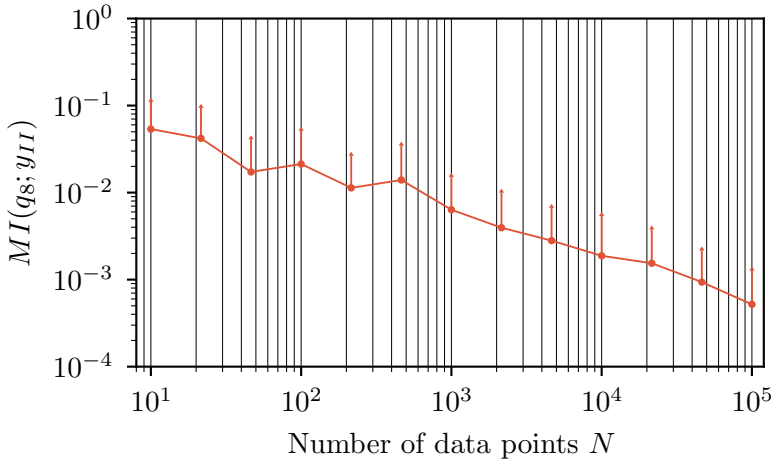


Figure 3.5.: The dependence of mutual information on the amount of data points N for q_8 and y_{II}.

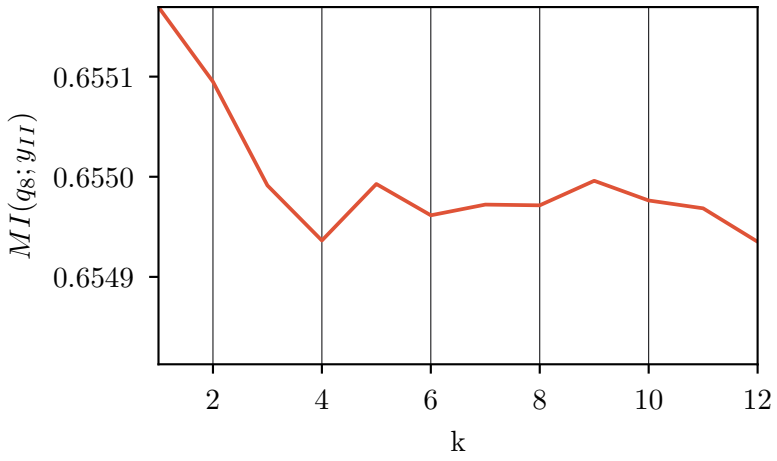


Figure 3.6.: Variation of mutual information $MI(q_8; y_{II})$ against the hyper-parameter k.

$\epsilon_{MI} = MI(q_8; y_{II}; N_{\max}) - MI(q_8; y_{II}; N_i)$. The errorbars indicate the 95 %-confidence interval based on 15 computations for each N_i.

The error and variance decrease steadily and converge towards $0.05\,\% \pm 0.075\,\%$. Therefore, 100 000 data points are sufficient to estimate the mutual information with negligible error, but decreased computational costs.

The dependence of $MI(q_8; y_{II})$ on the number of nearest neighbours k is presented in figure 3.6. The entropy estimate strongly depends on k as its bias increases for large k. Kraskov et al. recommends k to be a low integer within $\{2, 3, 4\}$ to minimise bias [44]. The behaviour of k for a combination of continuous q and discrete y is investigated in [47]. They use different k to estimate the mutual information for various q, y pairs and find little dependency on k. Following the recommendation of [44] and [47], the hyper-parameter is chosen as $k = 3$.

3.5. Sampling

The database contains in total $N \sim 5$ million samples from well-resolved large eddy simulation (LES) and direct numerical simulation (DNS). Combined with the number of features $N_q = 59$, the computational requirements for both algorithms become intractable. Furthermore, the number of samples from distinct simulations differs by up to three orders of magnitude, see chapter 4 and tables A.1 to A.5. This bias the models towards the set with more samples. Consequently, the database is subsampled to equalise the information and reduce computational costs.

Subsamples are drawn at random without replacement from the combined data sets. A combined data set includes all variations of one case presented in chapter 4. For example, the data for PH-Re includes 5 distinct cases with varying Reynolds number. These cases are combined into a single data set and samples are drawn.

Besides the difference in number of samples, the number of active and inactive error metrics on each data set is imbalanced. This imbalance is referred to as class imbalance in machine learning [48]. Class imbalancing as well as the assessment of an appropriate sample size is investigated below.

Class imbalancing

The true error metrics are evaluated on each dataset and show strong imbalance in the amount of active to inactive marker. On average, there are 29 % active marker for non-negativity and 13 % for anisotropy, see tables A.1 to A.5. The imbalance results, because RANS predictions are mostly correct and violations of the eddy-viscosity approximation are rare.

Class imbalancing is a common problem in machine learning. Weiss [48] recommends either cost-sensitive learning or sampling strategies to address imbalanced classes. Cost-sensitive learning manipulates the loss matrix which defines the costs of misprediction. Giving higher costs to a misclassification of the active metric, adjusts the learning process to improve predictions of active metrics. On the other hand, sampling strategies alter the amount of training data. These strategies include oversampling of the minority class or undersampling the majority class. Undersampling misses potentially useful information, but reduces training time. Whereas, oversampling uses exact copies of the minority class which causes overfitting.

Oversampling does not reduce the amount of training samples and, thus, is not considered. Undersampling and cost-sensitive learning have both been tested (results not shown). Manipulation of the loss matrix showed to be sensitive to the amount of data N. In comparison, undersampling achieved better generalisation, even with small sample sizes. Thus, undersampling is used to address class imbalancing.

Sample size calibration

A undersampled set contains equal amounts of active and inactive marker. The size of the set is calibrated by comparing the generalisation error for varying sample sizes s. The identifier model is trained with logistic regression, see section 3.3, and tested on a constant-size imbalanced test set. The generalisation error is assessed with the utility metric *F-measure*, which is introduced in section 3.6. This utility metric is equal to one for an ideal identifier. Since the samples are drawn at random, 15 repetitions are performed for each sample size and the F-measure is averaged.

Figure 3.7 shows the F-measure and 95 %-confidence intervals. For small samples sizes, the predictions have low utility and small variance, because the amount of samples is insufficient to predict well. The utility

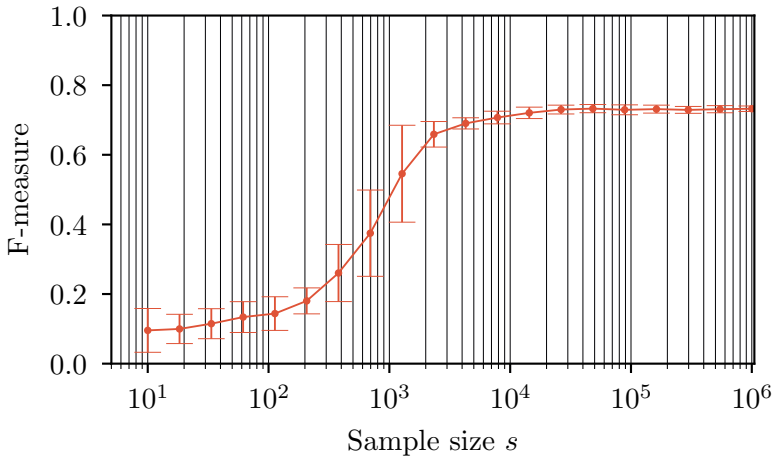


Figure 3.7.: The average generalisation error for identifier models when trained with varying sample size s.

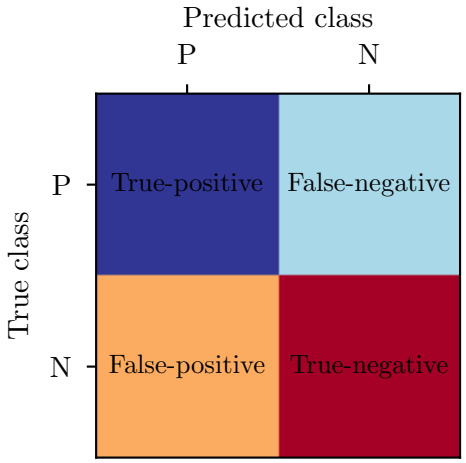


Figure 3.8.: The confusion matrix for binary classification. The positive and negative are abbreviated with P and N, respectively.

strongly increases with $s > 100$ samples. The model predictions depend on the randomly chosen samples which is indicated by high variance. For large s, the utility converges towards $72\,\% \pm 0.1\,\%$. The large bias of the model is a result of low model complexity. Nonetheless, the convergence indicates only marginal changes in the utility for sample sizes $s > 10\,000$. Therefore, the sample size is chosen to $s = 10\,000$.

3.6. Performance metrics

The training process of the algorithms requires a quantitative metric to distinguish the performance of models using a single quantity. The model evaluation uses performance metrics to guide the training process and identify good identifier. Furthermore, the performance metric determines the generalisation capability of a model.

In binary classification, performance metrics are based on the confusion matrix, see figure 3.8. The con-

fusion matrix compares model predictions $\tilde{y}$ against the true y. Possible evaluations of the matrix are true-positive (TP) and true-negative (TN), if the error metric y is correctly predicted. Otherwise, a false-positive (FP) indicates the prediction of an active y for a truly inactive y, and vice versa for false-negative (FN). For the present work, an active error metric belongs to the positive class and inactive metrics to the negative class.

For example, if a model for the anisotropy metric predicts an inactive marker close to the wall, while the true label is active. Then, the prediction of a negative class is incorrect, which is a FNin the confusion matrix.

The methodology of model evaluation aims at parsimonious, that is non-overfitting and interpretable, models [42]. Overfitting is avoided by considering the generalisation performance with a robust quantitative metric and interpretability observed with the model complexity. The best model must be chosen as a trade-off between generalisation performance and complexity of the model. This section introduces performance metrics for the assessment of generalisation and model complexity.

Accuracy

The simplest metric is *accuracy*. Accuracy measures the rate of correct predictions with

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.30}$$

The metric puts higher weight on the majority class. Consequently, a model achieves high accuracy, even if it only predicts the majority class [48]. Accuracy can be used for the learning process, because the undersampling provides balanced training data. Nonetheless, the evaluation on an unseen test case with the complete and, thus, imbalanced data requires a robust metric.

F-measure

[48] recommends the *F-measure* as a robust metric for imbalanced classes. The F-measure combines the *precision* and *true positive rate*. Precision considers the predictive performance of the model with respect to the positive class only. It measures the percentage of correct predictions of the positive class:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3.31}$$

TPR is the rate of correct positive predictions out of all possible positives in the data:

$$\text{True-positive rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{3.32}$$

The metrics are combined with the harmonic mean, cf. [49], to form the F-measure:

$$\text{F-measure} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}. \tag{3.33}$$

The F-measure is predicts within the range $[0, 1]$, see [30]. A F-measure equal to one indicates a perfect classifier. It decreases for increasing FPand FN, that is misclassifications.

Observing equations (3.31) to (3.33), the F-measure does not depend on correct predictions of inactive error metrics, that is TNs. TNs, however, are the majority class and easier to predict for a model compared to the positive class. In the context of RANS turbulence modelling, correct predictions of the positive class are more

expensive than the negative class, because the identifier does not invoke a change of the turbulence model for negatives. Therefore, TNs have less value for the classification and are not required for performance assessment.

In contrast to true predictions, the influence of false predictions impacts the RANS simulation. FPs invoke the CFD solver to change the turbulence model although the uncertainty in the LEVM is low. Similarly, FNs recommend no change in the turbulence model in spite of high RANS uncertainty and, thus, reducing the quality of a RANS solution. Both of these are separately considered in the precision and TPR, respectively. Therefore, the F-measure provides robust model evaluations with the positive class defined as minority class.

Model complexity

In addition to the above metrics, the model complexity is a criteria for model evaluation. Model complexity is not based on the confusion matrix, but counts the number of non-zero coefficients in a model. For a logistic regression model, the model complexity is the number of active features which is limited to 56. Instead, the SpaRTA algorithm linearly combines candidates to form model structures. Since the number of candidates exceeds the number of features by one or two orders of magnitude, the model complexity for SpaRTA models is possibly greater. Nonetheless, the model selection aims at interpretable models with minimal model complexity.

Receiver operating characteristic

The receiver operating characteristic (ROC) for classifier performance comparisons is discussed in [50]. It allows a visual comparison using the rate of true-positives to the rate of false-positive predictions. The true-positive rate is defined in equation (3.32) and the false-positive rate is defined as:

$$\text{False-positive rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}. \tag{3.34}$$

Clearly, the combination of the performance metrics accounts for changes in any element of the confusion matrix. Moreover, both performance metrics are not affected by class imbalancing. This is observable from the confusion matrix in figure 3.8. The true-positive rate depends on the first row, that is the positives. In comparison, the false-positive rate depends solely on the second row. Consequently, class imbalancing does not influence the metrics.

Figure 3.9 depicts the ROC. It visually indicates random classifiers with the black line. Any classifier on this line has the same probability of predicting a true-positive or false-positive resulting in a random outcome. Further, a strictly-positive classifier lies in the upper right corner (plus), and a strictly-negative classifier in the lower left corner (square). An ideal classifier in the characteristic lies in top left corner which is marked with a circle.

The overall performance of a classifier can be determined by its distance to the ideal state (circle). Comparing two arbitrary classifiers A and B, figure 3.9 shows that they have the same distance to an ideal classifier in spite of distinct true-positive and false-positive rates. The classifier A is conservative classifier, that is it classifies a positive only with strong evidence. On the other hand, B is a liberal classifier which classifies the positive class with weak evidence.

In the context of RANS turbulence modelling, switching or blending turbulence models increases the

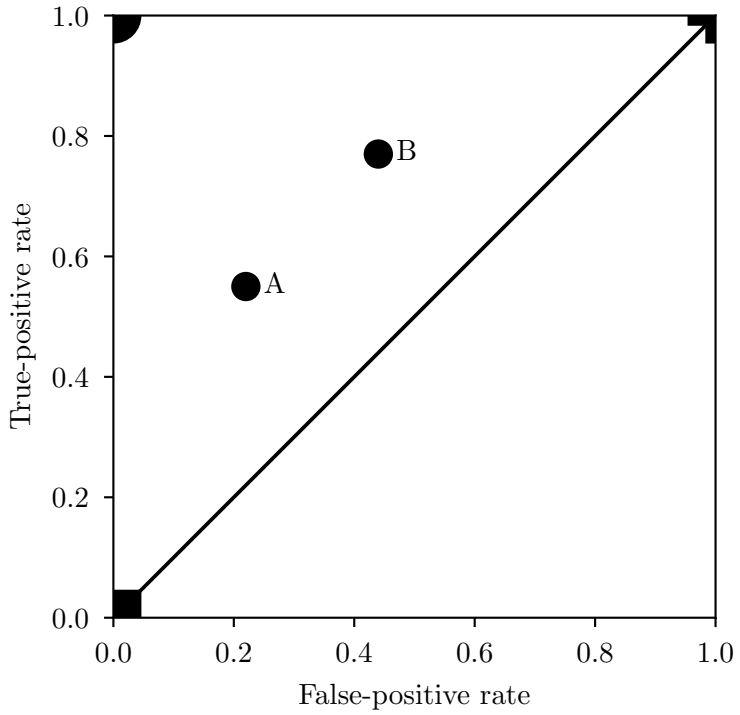


Figure 3.9.: The receiver operating characteristic for classifier performance analysis. The line represents a random, the circle an ideal, the square a strictly-negative and the plus a strictly-positive classifier.

computational costs of a simulation. For a universal and robust application of the identifiers, computational costs should only be increased when strong evidence for an improvement of RANS predictions is available. Therefore, the conservative classifier is favourable.

4. Database

Training data for the data-driven models is provided by a database of high-fidelity simulation data. Each data set is chosen according to three criteria: **(i)** the flow is difficult to predict for Reynolds-averaged Navier-Stokes (RANS); **(ii)** numerical methods of high-fidelity are employed; **(iii)** incompressible and non-reacting flow is simulated. The criteria are further discussed below.

(i) The data sets should cover weaknesses of RANS solutions in particular for Linear eddy viscosity models (LEVM). The reliability of RANS predictions depends on the capability of the turbulence model. The eddy-viscosity approximation for LEVM cannot predict anisotropy accurately in many flows, see section 2.2 for a detailed discussion. In order to identify where RANS solutions exhibit high uncertainty, the database includes cases with separation, curvature and pressure gradients.

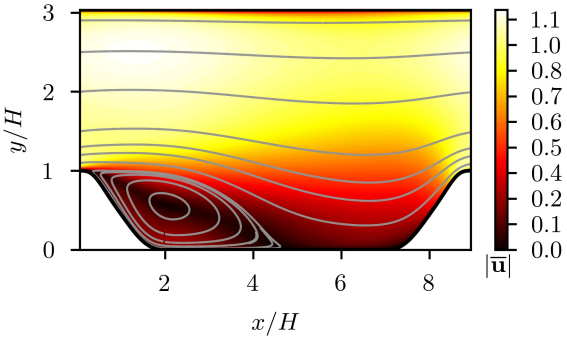
(ii) The data provides true information of the flow for training as well as testing models. To avoid ambiguity in the models, accurate solutions of the complex flow physics are required. High accuracy is provided by direct numerical simulation (DNS) and well-resolved large eddy simulation (LES) data. Nevertheless, the identifier models are designed to detect high uncertainty based on RANS input data. Therefore, only the mean flow and turbulence statistics from DNS and well-resolved LES data sets are included in the database.

(iii) The limitation to incompressible and non-reacting flow is set to reduce the complexity of the models. Compressibility, reacting and multi-phase flow increase the complexity of the physical phenomena. Data-driven models for compressible flow at high Mach numbers are trained in [51]. Besides other issues, the small number of training data reduced the generalisation of their models. Thus, more training data is required to provide sufficient information about complex flow physics. The rather simple choice of physical complexity promotes simpler models for the identification.

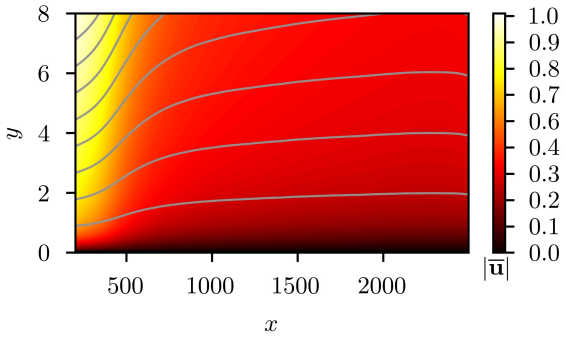
Table 4.1 gives an overview of the datasets employed in the present work. Each data set comprises multiple cases which vary in geometry, Reynolds number or boundary conditions. Flows with separation and curvature are included within the periodic hills (PH) cases PH-Re and PH-Geo as well as the curved-backwards-facing step (CBFS). The PH-Re data varies the Reynolds number on the same geometry. It provides information on the variation of the recirculation zone due to changing Reynolds number. On the other hand, the PH-Geo

Data set name	Description	Method	Source
PH-Re	Flow over periodic hills for $700 \leq Re_H \leq 10\,595$	DNS/LES	[52]
PH-Geo	Flow over scaled periodic hills for $Re_H = 5600$	DNS	[53]
CBFS	Curved-backwards-facing step for $Re_H = 13\,700$	LES	[54]
TBL-APG	Turbulent boundary layer on a flat plate with adverse pressure gradients and $910 \leq Re_\theta \leq 4320$	DNS	[55]
NACA	NACA4412 and NACA0012 aerofoils with 5° angle of attack and $100\,000 \leq Re_c \leq 1\,000\,000$	DNS	[56], [57]

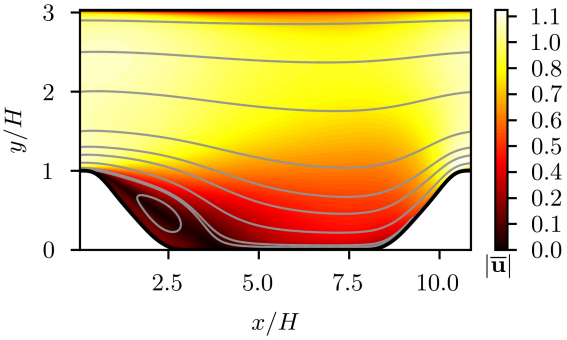
Table 4.1.: Overview of the database including the data set's name, a description, specification of the high-fidelity methodology and source.



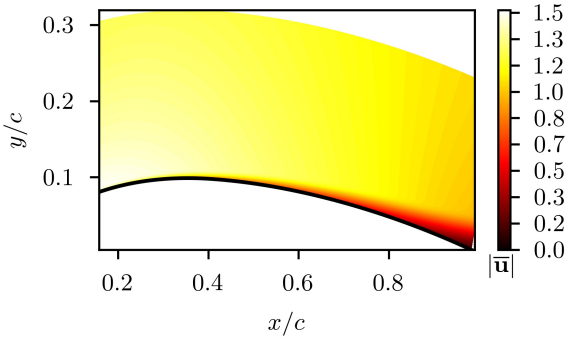
(a) Periodic hills at $Re_H = 5600$.



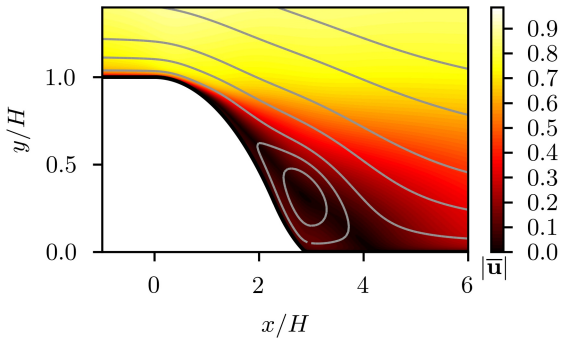
(b) Turbulent boundary layer subject to an APG.



(c) 1.5-scaled periodic hills at $Re_H = 5600$.



(d) NACA4412 at 5° AoA and $Re_c = 400\,000$.



(e) Curved-backwards-facing step at $Re_H = 13\,700$.

Figure 4.1.: Overview of data sets including flows with separation (left) and pressure gradients (right).

includes mild to strong separation by adjusting the curvature for the hill geometry. The difference between both periodic hill data sets becomes clear by comparing figure 4.1a and figure 4.1c. Furthermore, the CBFS data uses a gently curved step with a higher Reynolds number. Compared to the periodic hills, the CBFS does not constraint the flow. The geometry has a larger channel height and no second hill, see figure 4.1e.

The pressure gradient data includes National advisory committee for aeronautics (NACA) aerofoils and a turbulent boundary layer with adverse pressure gradient (TBL-APG). NACA aerofoils include favourable as well as adverse pressure gradients and a curved surface. The aerofoil data is limited to the turbulent boundary layer after it is tripped up to incipient separation close to the trailing edge. The TBL-APG data provides information purely on the effect of adverse pressure gradients of varying strength. Both flows are qualitatively shown in figure 4.1d and figure 4.1b.

The numerical method and validation of each data set is discussed below. Additionally, a brief overview of the anisotropy of turbulence on each data set is given. The visualisation of the anisotropy uses the red-green-blue (RGB) colormap for the barycentric triangle which is introduced in section 2.2.

4.1. Periodic hills

The periodic hills geometry is investigated with numerical and experimental methods in [52]. The investigation uses a range of Reynolds numbers within the range $Re_H = 700$ to $Re_H = 10\,595$, where Re_H is based on the hill height H. The numerical methods are well-resolved LES ($Re_H = 10\,595$) or DNS. Both are validated against particle image velocimetry (PIV) measurements. The flow over periodic hills is characterised by a separation bubble aft of the wind ward hill which causes challenging separation and reattachment dynamics for RANS turbulence models.

Anisotropy

The anisotropy in the periodic hills is visualised in figure 4.2, see section 2.2 for the definition of the map. The turbulence in the upper boundary layer transitions from two- to one-component turbulence with green and red colour, respectively. Two-component turbulence is a result of wall blocking. Furthermore, the expanding log region increases the peak in $\overline{u'^2}$ leading to thicker one-component turbulence. Closer to the windward hill, the flow is accelerated and, hence, the boundary layer becomes thinner. This boundary layer behaviour agrees with observations in channel flow [1].

The lower wall boundary layer between the hills in figure 4.2 is in the two-component state, because of stronger spanwise Reynolds stress inside and aft of the separation. Approaching the windward hill, the anisotropy tends towards the one-component state. Both states result from splatting. Splatting is described in [58] as the convection of large eddies from the point of reattachment towards the windward hill. The eddies strongly increase $\overline{w'^2}$ in the boundary layer and at the hill. Since wall blocking reduces $\overline{u'^2}$ and $\overline{v'^2}$, the turbulence approaches the one-component state at the downstream hill. The acceleration of the flow along the hill increases streamwise stresses and leads to two-component anisotropy at the hill crest.

In [52] two major effects are found as a consequence of varying Reynolds numbers. On the one hand, the upper wall boundary layer thickens causing thicker two- and one-component areas. On the other hand, the flow separation from the hill crest is delayed. Therefore, the effects of splatting along the lower wall are less pronounced for smaller Reynolds numbers.

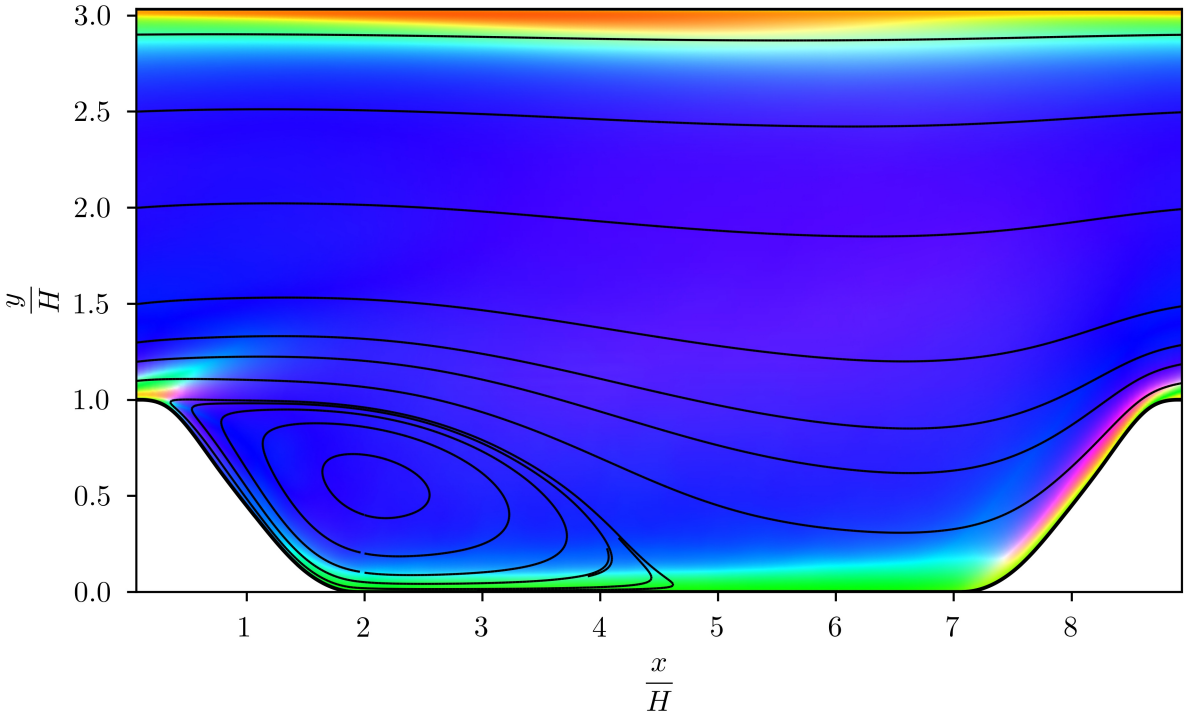


Figure 4.2.: States of turbulence on the periodic hills geometry from [52] at $Re_H = 5600$ visualised with RGB colormap.

4.2. Scaled periodic hills

In [53] a periodic hills geometry is studied aiming to create well-parameterised training data for data-driven turbulence models. The DNS are computed for constant Reynolds number based on hill height $Re_H = 5600$. In contrast to [52], the data set varies the hill geometry with a scaling factor α in the range $[0.5, 1.5]$. A scaling factor of $\alpha = 1$ corresponds to the geometry of the PH-Re data set which served as validation data in [53]. The scaling effectively varies the hill curvature and, thus, causes mild to strong separation.

Anisotropy

Figure 4.3 depicts the scaled periodic hill geometry at $Re = 5600$ with mild separation at $\alpha = 1.5$. The gentle curvature on the hills delays separation and decreases the recirculation zone size. Splatting still occurs, but has a decreased effect on the spanwise Reynolds stress at the windward hill compared to the non-scaled geometry. Anisotropy is mostly two-component at the hill as a direct result of wall blocking. One-component turbulence appears only in a thin layer before it transitions to an isotropic state.

The effects of acceleration above the hill crest and the upper wall boundary layer are similar to the PH-Re case. The reader is referred to the discussion in section 4.1.

4.3. Curved-backwards-facing step

The curved-backwards-facing step causes separation from the round surface which is investigated in [54]. Turbulence statistics, velocity and pressure fields are obtained with a DNS for $Re_H = 13\,700$. The Reynolds

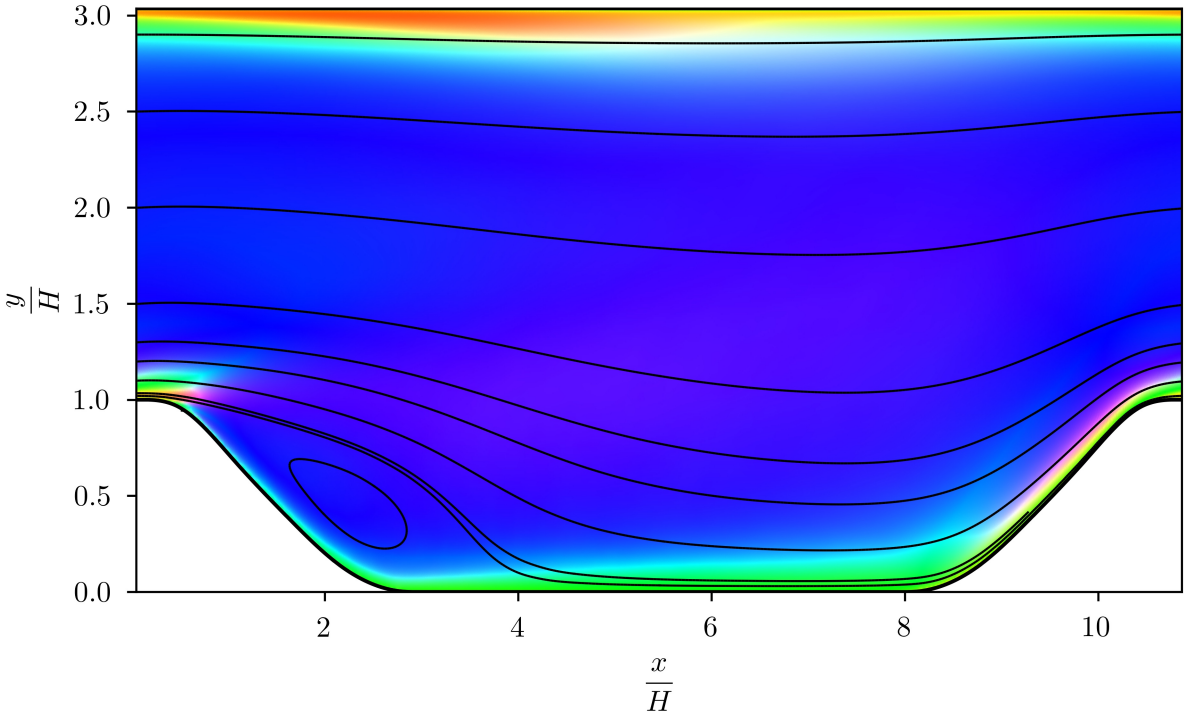


Figure 4.3.: Turbulence states on a mild-separation periodic hills geometry with $Re = 5600$.

number is defined with the step height H. Results are validated against experimental data in [59]. In [60], the experiments are carried out with PIV and laser doppler annemometry (LDA) measurements of the vortical structures and separated flow.

Anisotropy

The incoming boundary layer flow in figure 4.4 follows similar characteristics as described above for the data set PH-Re. The log region of the flow is dominated by streamwise fluctuations which result in one-component turbulence. The outer layer does not transition from the two-component state to isotropic characteristics, because the upper wall at $y/H > 8$ does not cause substantial blocking compared to the periodic hills data.

In [54], a strong increase in production of $\overline{u'^2}$ after the separation at about $y/H = 1$ is observed. The production decreases while redistribution increases inside the recirculation zone. Consequently, the transverse stress $\overline{v'^2}$ and $\overline{w'^2}$ increase leading to isotropic turbulence inside the separation bubble and two-component turbulence close to the wall.

Close to the reattachment point $y/H \approx 4.3$, [54] observed an increase in spanwise Reynolds stresses due to pressure-velocity interactions (splatting). Thus, turbulence close to the wall is close to the two-component limit with similar stream- and spanwise fluctuations.

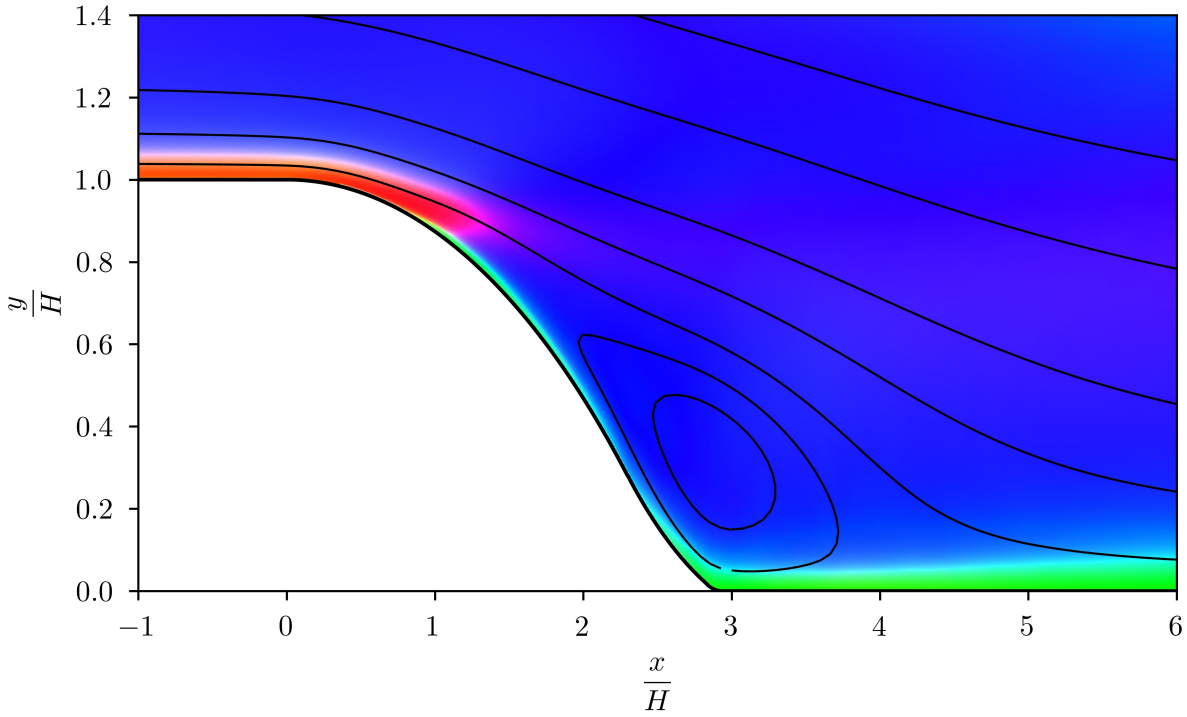


Figure 4.4.: Isotropic to 1C turbulence on the CBFS geometry at $Re_H = 13\,700$ visualised with RGB colormap.

4.4. Turbulent boundary layer with adverse pressure gradients

[55] studies the effect of APGs on a flat plate TBL with well-resolved LES. The TBL is in near-equilibrium condition which is achieved by prescribing the free stream velocity with a power law of the form $u_\infty = C(x - x_0)^m$ where C is a constant, x_0 a virtually defined origin and m the exponent of the law [61]. The pressure gradient is favourable for $m > 0$ and adverse for $m < 0$. This study compares non-constant APGs defined with $m = \{-0.13, -0.16, -0.18\}$ and $x_0 = 60$ against almost constant APGs with $m = \{-0.14, -0.18\}$ and $x_0 = 110$. The constant and non-constant pressure gradients are compared to investigate history effects on the TBL.

Anisotropy

The turbulent boundary layer for a non-constant adverse pressure gradient with $m = -0.13$ is shown in figure 4.5. The incoming profile is based on a DNS of a flat plate with pressure gradient equal to zero. This initial section is dominated by one-component turbulence as a result of wall blocking and the peak in production of streamwise fluctuations. The adverse pressure gradient is imposed at $x = 350$. Thereafter, the boundary layer is destabilised which leads to axisymmetric two-component turbulence in the near-wall region. The turbulence state transitions towards one-component and, later, isotropic turbulence in the outer layer. [55] found that the adverse pressure gradient exhibits a stronger wake with more energetic vortices in the outer layer as compare to a zero pressure gradient boundary layer. These vortices increase the wall-normal convection and lead to the change in the anisotropy.

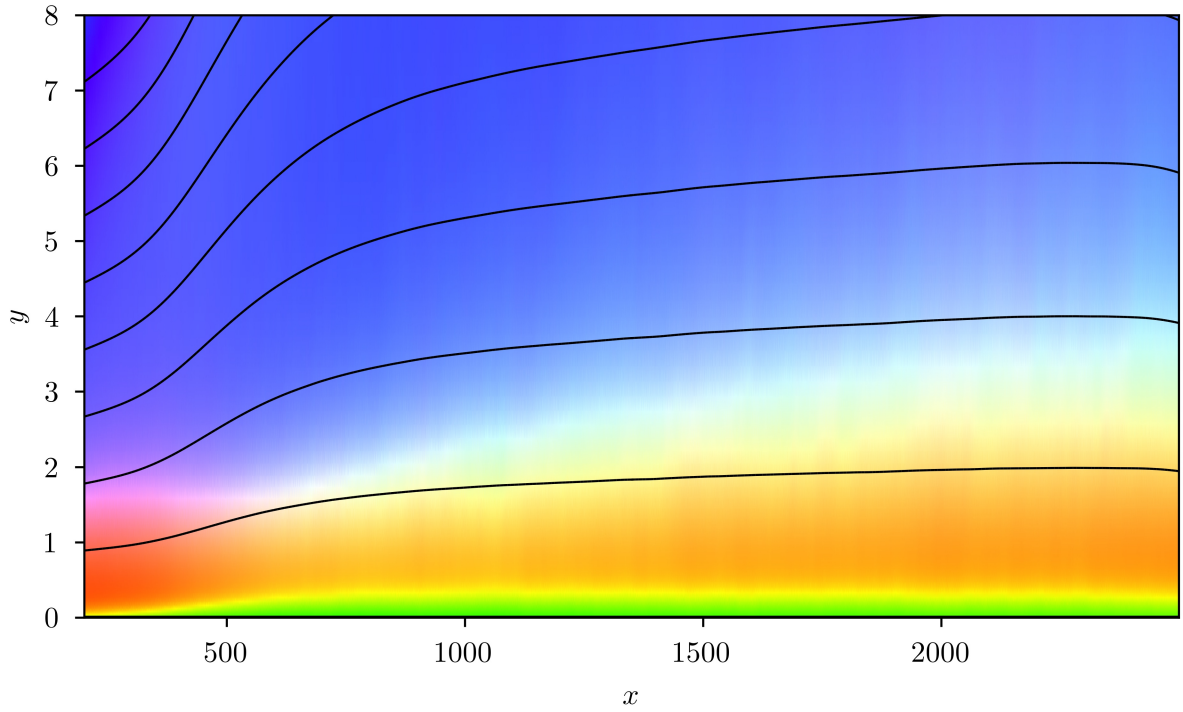


Figure 4.5.: 3C, 2C and 1C turbulence visualised with RGB colormap for a TBL under the influence of an APG.

4.5. NACA aerofoils

The data set includes NACA0012 and NACA4412 profiles which are studied with well-resolved LES in [57] and [56]. Simulations are performed for chord-length-based Reynolds number Re_c in the range 100 000 to 1 000 000 and an angle of attack of 5° Validation of the well-resolved LES method is performed in [62] against DNS data of a NACA4412 profile which originates from [63]. The NACA4412 data in [56] is created to study the Reynolds-number effect on a TBL subject to an APG on the suction side of aerofoils. Tanarro et al. [57] compare NACA4412 and NACA0012 profiles revealing strong influence of APGs on the outer region of TBLs. Both data sets provide training data for turbulent flows subject to strong pressure gradients.

Anisotropy

The turbulent boundary layer on the suction and pressure side of the wing in figure 4.6 is tripped at $x/c = 0.1$. Looking at the suction side, the anisotropy in the boundary layer is dominated by the production of streamwise fluctuations as a consequence of wall blocking. However, the favourable pressure gradient, and convex curvatures, stabilise the boundary layer by reducing the overall Reynolds stresses [1]. Aft of the crest, the pressure gradient decelerates the flow. In [56], this adverse pressure gradient leads to stronger wall normal convection of turbulence. The convection enhances redistribution of energy to the spanwise Reynolds stresses. Consequently, the turbulent boundary layer shifts towards the axisymmetric two-component state.

The pressure side boundary layer is subject to a weak adverse pressure gradient and concave curvature. Both act to destabilise the near-wall boundary layer [1]. The near-wall boundary layer is initially in the one-component state of turbulence. Since both the pressure gradient and curvature are subtle, the anisotropy

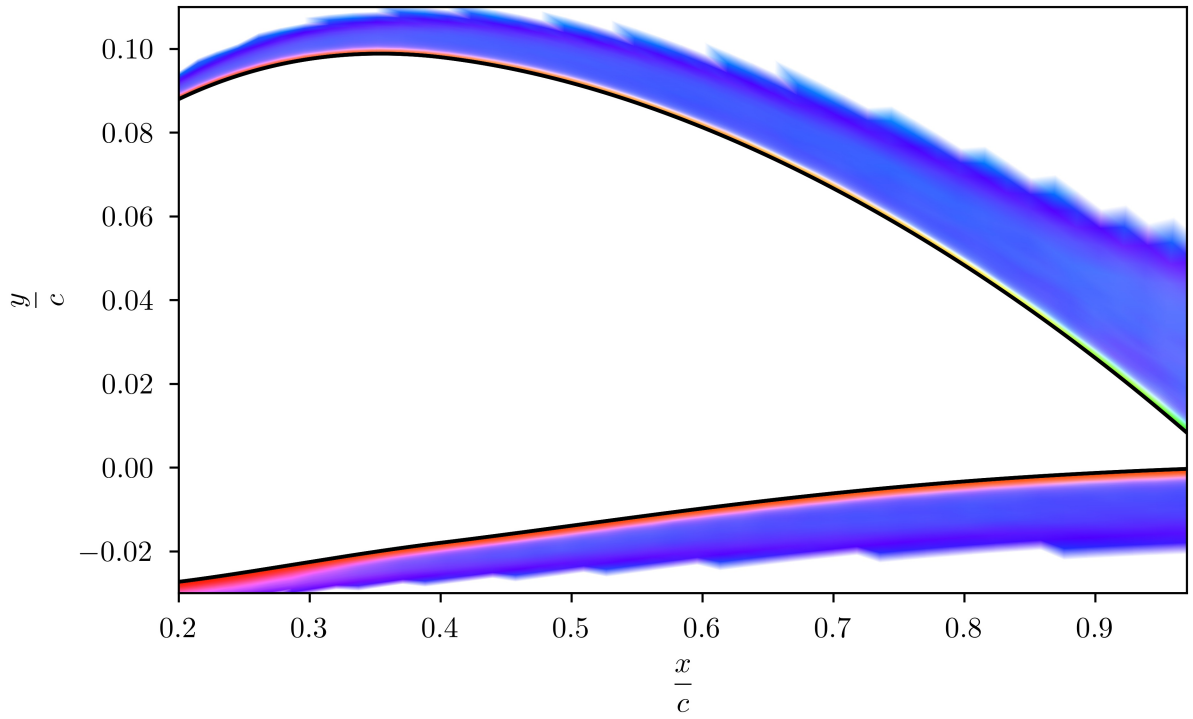


Figure 4.6.: Types of turbulence for the NACA4412 aerofoil at 5° AoA and $Re_c = 400\,000$.

state changes slowly along the chord length towards the axisymmetric two-component state.

5. Feature selection

The SpaRTA algorithm's computational complexity scales with the number of features as N_q^4. For computational tractability, N_q must be reduced. In order to use only the most-informative features, mutual information is employed to determine the dependency of each error metrics y on each feature q individually. This chapters presents the mutual information results for both error metrics and discusses the choice of features for SpaRTA.

5.1. Anisotropy features

The database comprises flows with either separation phenomena or pressure gradient effects, see figure 4.1. The analysis of the database showed that anisotropy occurs mainly through wall-blocking in flows characterised by pressure gradients, see figures 4.5 to 4.6. Further, in separated flows, turbulence becomes anisotropic in the recirculation region and aft of the reattachment point. Consequently, features that identify the boundary layer and the separation region are required.

Figure 5.1 presents the mutual information of the five most-informative features with respect to the anisotropy metric y_{II}. The features from highest to lowest information are: $k/\epsilon\|\mathbf{S}\|$, k, Re_d, deviation from parallel shear flow and the Q-criterion.

The ratio of turbulent time scale to mean straining $k/\epsilon\|\mathbf{S}\|$ is used to identify rapid distortions that is strong straining effects on the turbulence [1]. This is important in separation regions where the turbulent time scale indicates strong production of k. Further, it is informative of the outer layer and the near-wall region. In the outer layer, the ratio k/ϵ is large, because of high production and low dissipation of turbulence. In contrast, the dissipation becomes dominant near the wall, cf. [63, 55]. Similar to the turbulent time scale, the turbulent kinetic energy k conveys information about regions with strong turbulence including turbulent boundary layers, separation regions and the wake of separated flows, cf. [52]. Thus, both features provide information about regions of anisotropic turbulence.

The third feature is the wall-distance Reynolds number Re_d. Its definition in table 3.1 shows the dependency

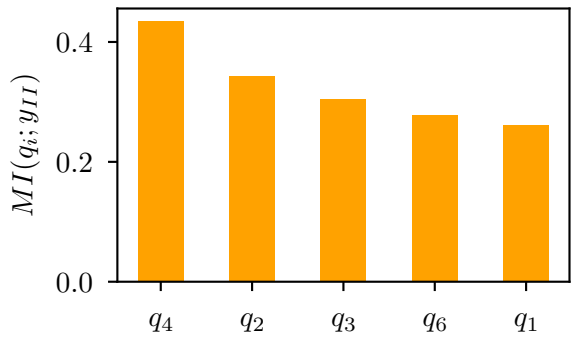


Figure 5.1.: Comparison of features based on their mutual information with respect to the anisotropy error metric y_{II}. The features q_i correspond to the physical and invariant features in table 3.1 and 3.3.

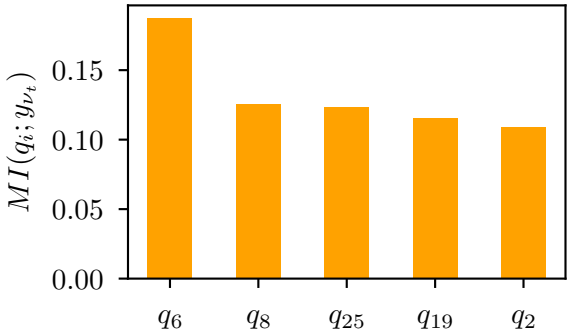


Figure 5.2.: The mutual information of the five most-informative features for the non-negativity error metric y_{ν_t}. The features q_i correspond to the physical and invariant features in table 3.1 and 3.3.

on k and the distance to the nearest wall d. The latter parameter clearly identifies boundary layers and is cannot be mistaken in the freestream, because of the limiter. Consequently, this Reynolds number is integrated in the blending function of the k-ω SST model to identify the near-wall region [10]. Also, in [15], Re_d reaches the highest importance for data-driven models that correct the anisotropy tensor indicating the information of the feature with respect to anisotropy. In summary, the wall-distance Reynolds number informs identifier models about the boundary layer where two- and one-component turbulence occurs.

The last features, deviations from parallel shear flow and the Q-criterion, both convey information about regions with rotating flow. While the former identifies curvature using streamlines [11], the latter includes the vorticity tensor $\mathbf{\Omega}$ which indicates strong vorticity [37]. Therefore, these features contribute information about the recirculation region.

Given the mutual information results, the most-informative features provide information about all regions of anisotropy observed in the database. Thus, the five features in figure 5.1 are chosen to train models with the SpaRTA algorithm.

5.2. Non-negativity features

A negative eddy-viscosity artificially reduces the viscous diffusion in a flow to account for turbulence. This non-physical effect occurs in separated flows, especially, in the recirculation region and downstream in the wake of the separation. On the other hand, negative eddy viscosities occur in boundary layers subject to adverse pressure gradients. The identifier models require information about both phenomena to effectively identify the non-negativity metric.

Figure 5.2 shows the mutual information for five features q with respect to the non-negativity metric y_{ν_t}. The five most-informative features from left to right are: the deviation from parallel shear flow, the pressure gradient along a streamline, the invariants $\overline{\mathbf{PK}}$ and $\overline{\mathbf{P}^2\mathbf{S}^2}$ and k.

The strongest dependency is found for the deviation from parallel shear flow. Deviations identify shear dominant regions where turbulence production is strong [11]. Thus, the feature indicates recirculation regions for the separation data sets where negative eddy viscosities are likely.

The second feature is the pressure gradient along a streamline. It informs models about the transition from a favourable to an adverse pressure gradient on the NACA aerofoil data. Further, the feature correlates with

the recirculation regions where the pressure gradient becomes stronger due to velocity changes. Considering the periodic hills cases, the acceleration and deceleration over the hills is largely identified by the pressure gradient. The fluid is compressed when accelerated towards the hill crest and expanded when decelerated on the leeward hill. Therefore, the pressure gradient along a streamline suggests regions of rapid distortions where LEVM are known to be invalid [1].

The invariant features both include the pressure gradient with the pseudo-tensor $\mathbf{P}$. Hence, they convey information according to the above explanation for the pressure gradient. Moreover, the invariant $\overline{\mathbf{PK}}$ includes the gradient of k. This gradient indicates the production and dissipation regions of turbulent kinetic energy within boundary layers as well as separation.

Following the argument for the anisotropy features, the turbulent kinetic energy identifies boundary layers and separation regions with strong turbulence. The addition to the features for identifier models presents information for typical regions of a violated Boussinesq hypothesis and, hence, negative eddy viscosities.

In summary, the choice of non-negativity features based on mutual information shows potential to identify regions of negative eddy viscosities. The five features are included to derive sparse identifier models with the SpaRTA algorithm for the non-negativity metric.

6. Identifier models

This chapter discusses the identification of high uncertainty in RANS solutions using data-driven models. High uncertainty is identified using two error metrics. The error metrics detect violations of Boussinesq's hypothesis due to anisotropic turbulence and negative eddy viscosities, see section 3.1.

The data-driven models are constructed with two distinct machine learning algorithms that allow interpretable model structures. Logistic regression constructs identifier models as linear combination of features. These models achieve low complexities and allow straightforward insight into the physics. Instead, the SpaRTA algorithm combines monomials of features into non-linear candidates. Linear combinations of these candidates form the model structure. In comparison, the SpaRTA models reach higher model complexity than Logistic Regression models which allows the models to describe complexer relationships. However, SpaRTA uses sparsity constraints to promote lower complexity in the model structures, see section 3.3.

The following sections evaluate identifier models for the two error metrics. The evaluation begins with the model selection where an interpretable model is chosen based on its performance and model complexity. Performances are only assessed as generalisation performance, that is model predictions on unseen data. After the selection of interpretable models, the generalisation of selected models is evaluated quantitatively and qualitatively. The quantitative evaluation compares the rate of true-positive and false-positive in the model predictions. Each rate is averaged over all cases that belong to an unseen data set. The qualitative performance analyses model predictions on a specific case to investigate the identifiers capabilities in the physical domain. The following section discusses the anisotropy metric and section 6.2 the non-negativity metric.

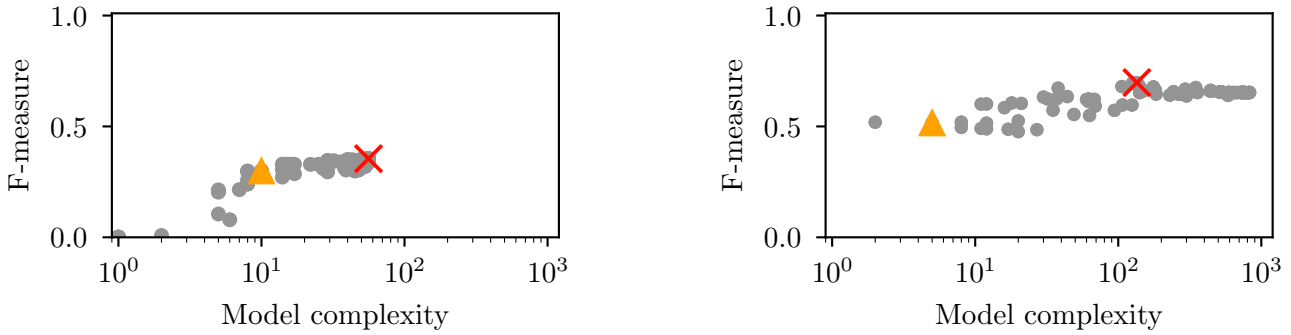
6.1. Anisotropy identification

Turbulence anisotropy causes inaccuracies in RANS solutions, because eddy-viscosity models rely on the invalid assumption that turbulence influences the velocity field analogous to molecular viscosity. Regions of anisotropic turbulence are identified with the error metric y_{II} which comprises the one- and, partially, two-component state of turbulence, see figure 3.1. Both algorithms constructed a number of identifier models for the metric. Within this section, an interpretable model for the anisotropy identification is selected for each algorithm and, subsequently, evaluated.

Model selection

An interpretable model requires low complexity and, ideally, does not decrease the performance. The balance between model complexity and performance is investigated in figure 6.1. Both graphs depict the number of non-zero model terms against F-measure. The F-measure is averaged over all cases of the PH-Re data set. A yellow triangle indicates the selected model which balances low complexity and high performance. Furthermore, the red cross shows the best model performance.

Beginning with the logistic regression models, the graph in figure 6.1a shows an exponential increase in model performance. Clearly, the model improves significantly for a small number of features and less



(a) Logistic regression models tested on the PH-Re data set.

(b) Models constructed with the SpaRTA algorithm and tested on the PH-Re data set.

Figure 6.1.: Model performance plotted against complexity for logistic regression and SpaRTA identifier models which detect anisotropy. A red cross indicates the best-performing model and a yellow triangle the selected algebraic model.

when the bulk of features is included suggesting that few features are strongly informative for the anisotropy metric. This is further supported when comparing the best to the interpretable model. The best model reaches an F-measure of 0.35 using 56 active features. However, the interpretable model decreases only slightly in performance to 0.29 while significantly reducing model complexity to 8 features. An increase in performance to 0.33 is possible, but would almost double the number of features to 14. Therefore, the model with 8 features is selected.

In spite of the model selection, the logistic regression models' performance is low. The best-performing model in figure 6.8a achieves a true-positive rate of 0.22, that is it predicts 22 % of all possible anisotropy identifiers. Comparing training and test performance, the same model reaches a F-measure of 0.91 on the training data (not shown), but only 0.35 on the test data. Thus, the models generalisation from training data to unseen data is limited, that is the logistic regression models underfit the anisotropy identification.

All logistic regression models are chosen with the same procedure leading to the following set of models for each test data:

$$M_{1,II} = -2.62q_3 + 0.69q_2 - 0.54q_{12} + 0.45q_4 - 0.36q_1 + 0.18q_6 - 0.13q_{25} - 0.05q_{39}, \tag{6.1}$$

$$M_{2,II} = -2.26q_3 - 0.71q_{12} + 0.45q_2 + 0.31q_4 + 0.24q_6 - 0.19q_9 + 0.17q_{13} + 0.15q_{23} + 0.08q_{19} - 0.03q_9, \tag{6.2}$$

$$M_{3,II} = -1.52q_3 - 0.51q_{12} + 0.31q_2 - 0.20q_1 + 0.11q_4 + 0.07q_{13} - 0.06q_7, \tag{6.3}$$

$$M_{4,II} = -1.54q_3 + 0.40q_2 + 0.23q_4 - 0.20q_{12} - 0.18q_1 - 0.06q_9 + 0.04q_6, \tag{6.4}$$

$$M_{5,II} = -1.54q_3 + 0.35q_2 - 0.34q_{12} + 0.26q_4 + 0.15q_6 - 0.12q_1 + 0.06q_{23} - 0.01q_{25}. \tag{6.5}$$

Model coefficients in equations (6.1) to (6.5) are rounded to the second decimal position. The models vary in complexity in the range 7 to 10 and use physical as well as invariant features. Nonetheless, the major feature in each model is the wall-distance based Reynolds number (q_3). This is expected, because Re_d identifies boundary layers which correlate well with turbulence anisotropy due to wall blocking, see anisotropy analysis on the test data in sections 4.1 to 4.5. Furthermore, the second strongest feature is either the tur-

bulence intensity (q_2) or the invariant $\hat{\mathbf{\Omega}}^2$ (q_{12}). Both features identify the outer layer of boundary layers where turbulent production becomes strong and, hence, are informative for anisotropy close to the wall. In conclusion, the strongest features indicate, that the logistic regression models identify mostly the boundary layers which possibly leads to low performances.

On the other hand, the SpaRTA models are shown in figure 6.1b. The model complexity is significantly higher for the SpaRTA algorithm, because the non-linear combination of features to candidates allows a large number of model terms. The trend of model performance over complexity is exponential, similar to the logistic regression models, but less pronounced. Nonetheless, the overall performance is on average 0.2 higher in F-measure. This is a direct result of the non-linear features which allow strong improvements in the prediction of the anisotropy metric.

The selected model achieves an F-measure of 0.60 using 11 model terms. Its performance is significantly lower compared to the best model with 135 model terms and F-measure of 0.69. An enhanced model for this test data uses 30 terms and achieves 0.67. However, models with 30 terms are hard to interpret, especially, when most terms include multiple features. As a result, the model with lower performance is selected.

Models for the other data sets follow the same procedure resulting in five independently chosen identifier models. The average reduction in performance compared to the best-performing model is 6 %. The structure of the chosen SpaRTA models reveal an interesting pattern. The algorithm repeatedly discovered the same model structure independent of the training data:

$$\begin{aligned} M_{6,II} &= -6.14 q_1^{2.0} q_3^{3.0} q_6^{2.0} + 4.73 q_2^{0.5} q_4^{2.0} \\ &\quad - 0.42 q_3^{2.5} - 0.45 q_3^{3.0} - 0.23 q_3^{3.5} + 0.06 q_3^{4.0} + 0.09 q_3^{4.5} \\ &\quad - 0.39 q_1 q_2^{0.5} q_4^{2.0} - 0.17 q_1 q_3^{3.0} q_4 - 0.04 q_1^{2.0} q_3^{3.0} q_4 - 0.04 q_2 q_4^{4.0}, \end{aligned} \tag{6.6}$$

$$M_{7,II} = 5.38 q_2^{0.5} q_4^{2.0} - 1.72 q_3^{3.0} - 0.14 q_3^{3.5} + 0.73 q_3^{4.0} - 0.11 q_3^{4.5}, \tag{6.7}$$

$$M_{8,II} = 5.28 q_2^{0.5} q_4^{2.0} - 1.60 q_3^{3.0} - 0.13 q_3^{3.5} + 0.68 q_3^{4.0} - 0.11 q_3^{4.5}, \tag{6.8}$$

$$M_{9,II} = 5.40 q_2^{0.5} q_4^{2.0} - 1.68 q_3^{3.0} - 0.12 q_3^{3.5} + 0.73 q_3^{4.0} - 0.13 q_3^{4.5}, \tag{6.9}$$

$$M_{10,II} = 5.35 q_2^{0.5} q_4^{2.0} - 1.54 q_3^{3.0} - 0.23 q_3^{3.5} + 0.56 q_3^{4.0} - 0.01 q_3^{4.5}. \tag{6.10}$$

For the PH-Re data, the algorithm also found the repeated model structure. However, it does only achieve a F-measure of 0.51 making the loss in performance unreasonable. The model coefficients are rounded to the second decimal position. These models predict a positive anisotropy metric when the model output is greater than zero and negative otherwise. Furthermore, the magnitude of the model coefficients demonstrates the relative importance of each term, because each feature is normalised onto the range $[-1, 1]$. With the exception of q_3, the wall-distance based Reynolds number lies within the interval $[0, 2]$ which effectively doubles the coefficients of q_3. This normalisation makes an analysis of the model's behaviour possible.

The models in equations (6.7) to (6.10) use a composite candidate of the turbulence intensity (q_2) and the ratio of the turbulent time scale to straining (q_4). Both features identify regions of turbulence production and, thus, the model potentially identifies separation regions as well as boundary layers. The first term is merged with a power series of the wall-distance based Reynolds number (q_3). The wall-distance based Reynolds number is only informative close to the wall, see section 5.1. Therefore, the model structure suggests a strong dependence on wall-proximity. Additionally, the model in equation (6.6) includes the q-criteria and the deviation from parallel shear flow. Both features are informative of the recirculation region potentially providing the model with more information about the separation region. Next, the selected models are

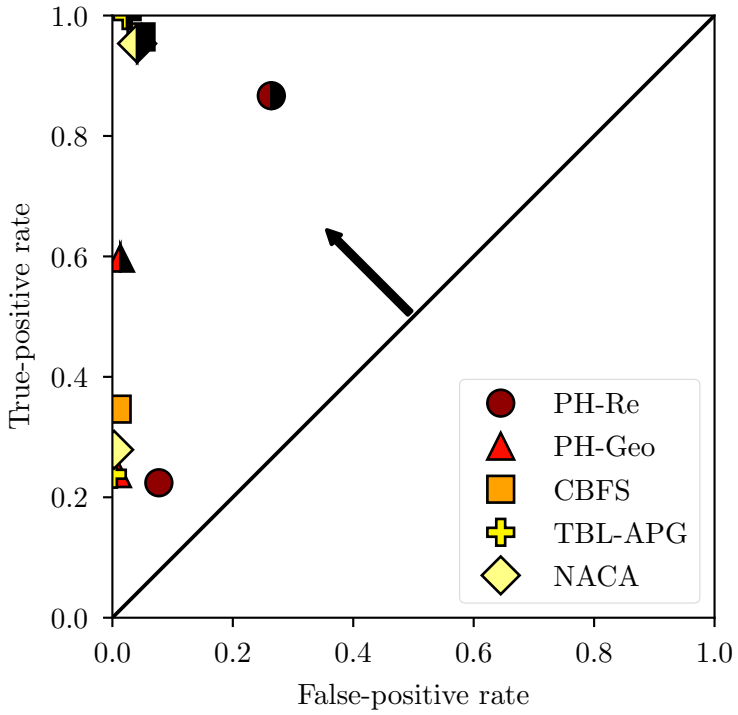


Figure 6.2.: Receiver operating characteristic for anisotropy identifiers. The shape of a symbol refers to the (unseen) test data set. Full symbols represent identifier models trained with logistic regression and half-filled symbols correspond to SpaRTA models. An optimal identifier lies in the upper left corner and models improve in performance along the arrow's direction.

evaluated quantitatively for their respective test data set.

Quantitative performance

The model selection suggested a limited generalisation of logistic regression models with identifications mostly close to the wall. Moreover, the interpretable SpaRTA models showed reduced performance compared to high-complexity models, but find a mutual form for the identification. In order to clarify the effect of these findings, the model performance is investigated with predictions on test data.

The receiver operating characteristic in figure 6.2 displays the performance for the selected model on each test data set. True-positive and false-positive rates are averaged over each case of the data sets. A random classifier is represented by the solid line and an ideal classifier reaches a true-positive rate equal to one and zero false-positive rate.

The logistic regression models achieve almost zero false-positive rates with the exception of the PH-Re data where about 10 % of the positive predictions are misclassified. Still, the true-positive rates are in the range 20 % to 40 % which means less than half of the points with active anisotropy metric are detected. Evidently, the previous suggestion of limited generalisation performance is confirmed.

The performance of the SpaRTA model on different test data shows strong differences between the test data sets. It achieves high performances on the curved step, aerofoil and flat plate geometry while significantly lower performance on both periodic hills data sets. In comparison, the former geometries and the periodic

hills geometry have decisively different characteristics. One the one hand, the upper wall on the periodic hills is at a distance of three step heights H while the distance is $8H$ on the curved step and, even larger, on the aerofoil and flat plate, cf. sections 4.1 to 4.5. The confinements leads to stronger acceleration and deceleration effects across the hills. On the other hand, the turbulence on the periodic hills includes not only the large separation region but also the interaction of the separated flow with the windward hill, that is splatting, cf. sections 4.1 to 4.2 and [58]. Consequently, the identification of anisotropy on the periodic hills is more challenging for the model compared to other data sets.

While the CBFS, TBL-APG and NACA data sets are predicted with similar rates, the predictions on periodic hills achieve distinct performances. Predictions for PH-Geo data achieves low false-positive rates of only 1 %, but predicts 60 % of all active anisotropy metrics. In contrast, the model for PH-Re data in equation (6.6) mispredicts 30 % of the positives, but achieves a true-positive rate of 86 %. According to [50], the PH-Geo model in equation (6.7) is considered a conservative model which makes decisions only with strong evidence, that is wall vicinity. On the other hand, the PH-Re model classifies liberally classifying almost all positive points correctly by relying on weak evidence. Weak evidence for this model is the number of distinct candidates included into the model which lessen the model's dependency on wall-distance. Thus, the lean model structure in the PH-Geo model leads to stronger dependency on the wall distance, and robuster identification of anisotropy.

Qualitative performance

The quantitative performance of distinct models shows large differences in model performance due to either the algorithm or the test data's geometry. So, the qualitative evaluations aims at providing insight into the origin of these differences. The evaluation inspects identifications with the selected models on distinct flows with separation or strong pressure gradients. Each identification is investigated point-wise using the confusion matrix which shows the true and false predictions visually. Each grid point is coloured with respect to a positive or negative anisotropy metric, that is active or inactive. The true predictions use blue and red for true-positive and true-negative points, respectively. Furthermore, false predictions use adapted colours with orange for false positives and light blue for false negatives. As a reference, the confusion matrix is displayed in figure 3.8.

The interpretable logistic regression models showed strong dependence on wall-distance, see equations (6.1) to (6.5) and overall low generalisation performance. The limitations of the logistic regression models are tested on the curved step geometry with the best interpretable model in equation (6.3). Figure 6.3 depicts the predictions evaluated with the confusion matrix. Dark blue points are true-positives and dark red are true-negatives showing that the free-stream is correctly predicted as isotropic and the near-wall region as anisotropic. However, the model does only predict the anisotropy close to the wall and does not capture the anisotropy further removed. This leads to low true-positive rates and, also, confirms the strong wall-dependence of the models. Since this is the best performance of an interpretable model, the predictions on other data sets have similar limitations which further proves the narrow performance of low complexity models. Therefore, the logistic regression models are not suitable identifier models for anisotropy identification.

Next, the SpaRTA models are analysed. The analysis focusses on the prediction of a strongly separated flow and a flow with strong pressure gradient to investigate the differences in the quantitative performance. Each prediction utilises the respective interpretable model in equations (6.6) to (6.10).

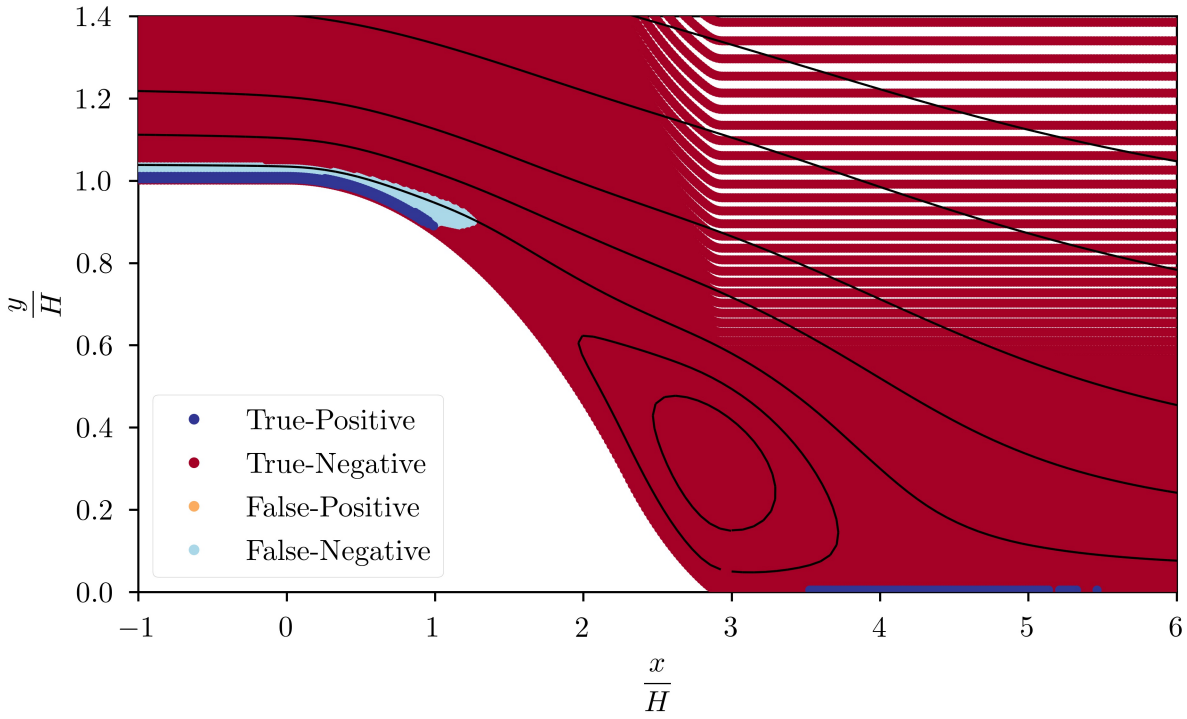


Figure 6.3.: Qualitative evaluation of interpretable logistic regression model on the curved step geometry with $Re_H = 13\,700$. Streamlines visualise the recirculation region aft of the step. White lines on the right of the plot result from variations in mesh resolution.

Beginning with separated flow, the model in equation (6.6) predicts the anisotropy metric on an unscaled periodic hills geometry with Reynolds number $Re_H = 5600$, that is PH-Re data. Figure 6.4a shows model predictions using the confusion matrix. The mean flow is identified by the model as isotropic turbulence with true negative (red) points. Whereas, anisotropy is detected along the upper and lower walls.

A detailed section of the upper wall is shown in figure 6.4b. Here, the identifier model predicts the region of one-component anisotropy with almost constant thickness along the wall. However, the boundary layer grows thicker aft of the upstream hill causing a growing region of anisotropy, cf. figure 4.2. The thicker region of one-component turbulence is not predicted by the model as indicated by the false negative predictions.

Looking at the lower wall, both hills are displayed in more detail in figure 6.4c and 6.4d, respectively. The upstream hill includes a small region of one-component turbulence close to the hill crest. This region is correctly classified for points close to the wall. However, points remote from the wall are mispredicted as inactive. Moreover, the model mispredicts the curved surface along the leeward side of the hill. Although two-component anisotropy occurs close to the surface, cf. figure 4.2, the anisotropy metric defined in [12] does not include the axisymmetric contraction of turbulence. Thus, the true error metric is inactive. Nonetheless, the model erroneously predicts active points along the complete leeward side of the hill which leads to the false-positive points.

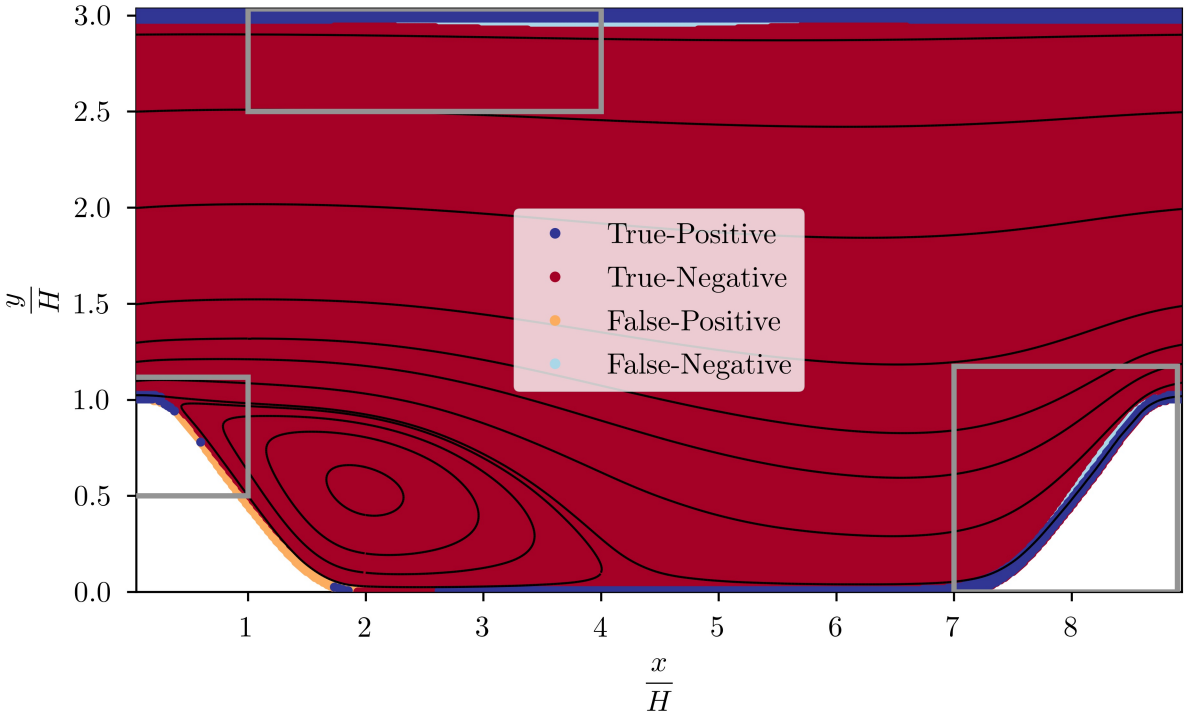
A similar scheme is found on the windward side of the hill in figure 6.4d. This hill side includes a larger region of one-component turbulence remote from the wall which results from the convection of turbulent structures. The anisotropy in the vicinity of the wall is well-predicted by the model. Whereas, the remote region of one-component turbulence is mistakenly predicted as negatives as indicated by the false-negatives.

The analysis of the periodic hills case shows, that the model predicts anisotropy as a consequence of wall-blocking. But, it can only partially capture the effects of complex flow phenomena, for example, boundary layer growth and splatting. Therefore, the anisotropy identification on separated flows is limited to wall-blocking induced anisotropy.

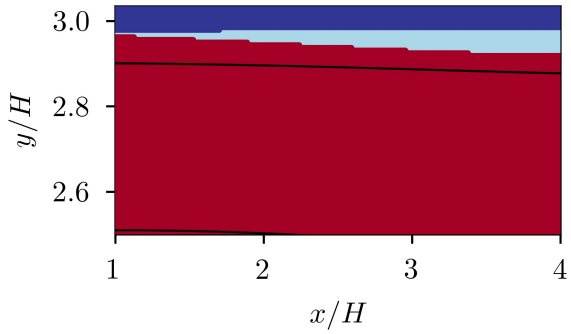
This limitation is clearly visible in the smooth model prediction in figure 6.5. The smooth prediction transforms the model with the logistic function $\sigma(M(q))$, cf. figure 3.2, and depicts the probability of an active metric with the contour. Identifications close to the wall have high probability which corresponds to an active metric. The probability changes rapidly in a small interval towards the mean flow which implies low uncertainty for the active anisotropy predictions. While the mean flow reaches a probability of up to 0.2, the increased uncertainty does not lead to misclassifications by the identifier. Consequently, active and inactive metrics are clearly distinguished and have small uncertainty, but do not capture the more challenging effects of turbulence in separated flow.

Apart from separated flows, anisotropy identifications on a NACA4412 profile with $Re_c = 100\,000$ are investigated. The predictions are presented in figure 6.6. In figure 6.6a, a view of the complete geometry shows an overall good prediction of inactive points in the freestream and active points close to the profile. Still, the crest of the profile around $x/c = 0.35$ and the aft region in the range $x/c = [0.8, 0.98]$ are not perfectly identified.

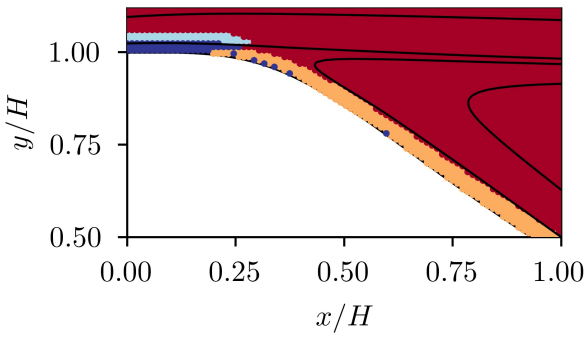
The crest of the profile is presented in figure 6.6b with mispredictions in front and aft of the profile's crest. Across the crest, the freestream pressure gradient transitions from a favourable to an adverse pressure gradient. In front of the crest, the favourable pressure gradient does not influence the one-component turbulence. Since the boundary layer is growing despite the favourable pressure gradient, the region of active error metrics increases. This growing region however is overpredicted by the model which leads to an increase



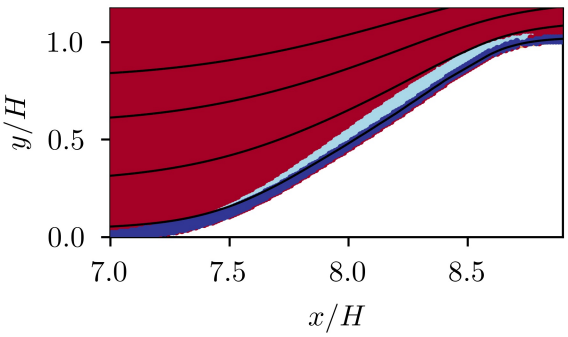
(a) Overview of the identification.



(b) The upper wall.



(c) The leeward side of the hill.



(d) The windward side of the hill

Figure 6.4.: Identification of the anisotropy metric on the perioidic hills geometry with $Re_H = 5600$. Predictions are performed with the algebraic SpaRTA model defined in equation (6.8). An overview and three enlarged sections of both hill sides and the upper wall are shown. Grey boxes indicate the enlarged sections.

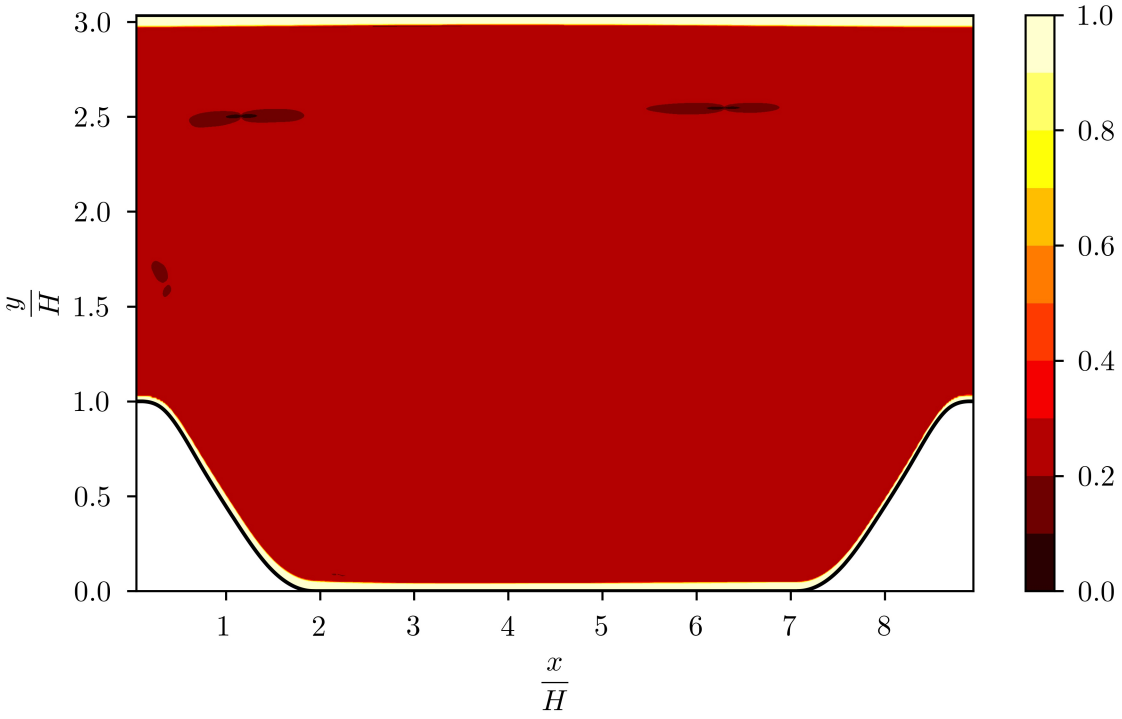


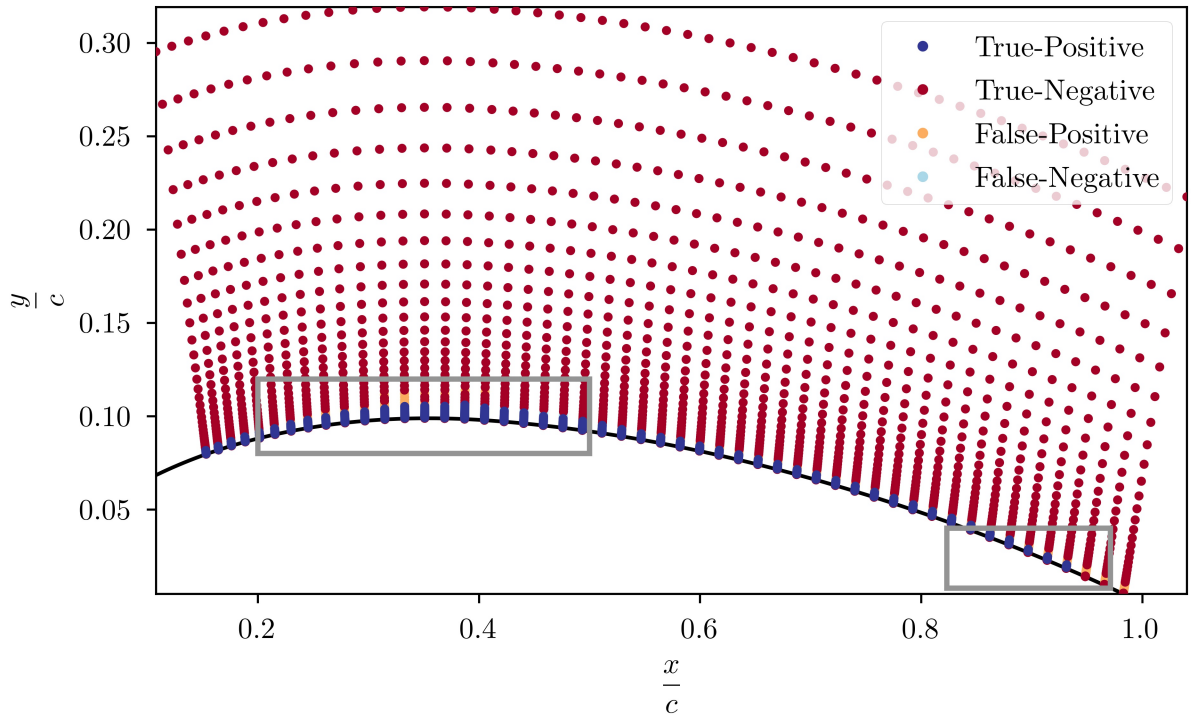
Figure 6.5.: Smooth identification of the anisotropy metric on periodic hills with $Re_H = 5600$ using the algebraic SpaRTA model defined in equation (6.8).

in false-positives on the upper side of the region. In contrast, the adverse pressure gradient, aft of the crest, convects turbulent structures from the wall towards the outer layer which shifts the anisotropy towards the two-component state, see [55, 56, 63]. The identifier does not capture this effect and underpredicts anisotropy downstream of the crest resulting in false-negatives.

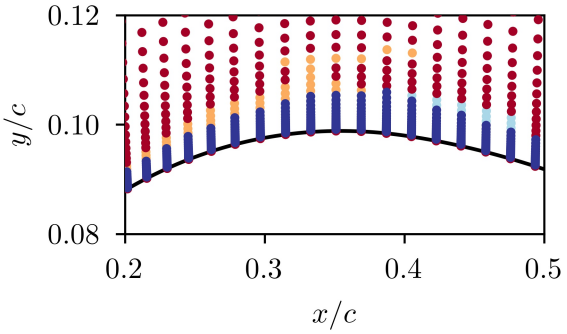
The aft section of the aerofoil in figure 6.6c shows the effect of the adverse pressure gradient. The anisotropy continues to transition towards the two-component state of turbulence. Accordingly, the region of active anisotropy metric shrinks, and disappears completely for $x/c > 0.94$. Considering the model predictions, the aft section of the aerofoil is increasingly mispredicted showing the models limitation to react to the adverse pressure gradient.

These effects are similarly observed from the smooth model prediction for the aerofoil in figure 6.7. The model predicts a growing layer of anisotropy with high probability of an active marker. While this region is predicted with low uncertainty, it confirms that the model does not react to the pressure gradient's influence on the anisotropy. Further, it depicts an arbitrary region of higher probability which leads to mispredictions of active points, see figure 6.6b.

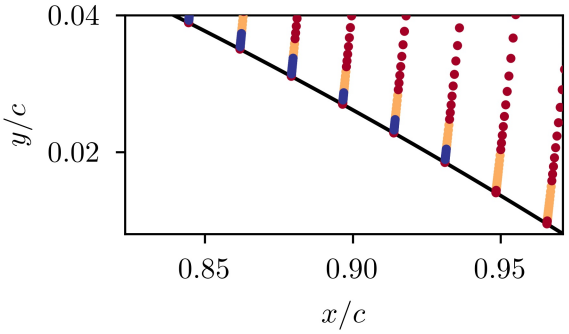
In summary, the algebraic model achieves good anisotropy predictions for the aerofoil, but it does not capture the challenging effect of the pressure gradient on turbulence. This shortcoming is potentially circumvented by adding additional features that convey information of the pressure gradient to the identifier model. Nonetheless, the model shows a limitation to wall-induced anisotropy which correlates with the prediction on separated flows.



(a) Overview of the identification.



(b) The area around the profile's crest.



(c) The windward side of the hill

Figure 6.6.: Identification of the anisotropy metric on the NACA4412 profile with $Re_c = 100\,000$. Predictions are performed with the algebraic SpaRTA model defined in equation (6.8). An overview of the aerofoil and two detailed views of the crest and the tail are presented. Grey boxes in the overview indicate detailed views.

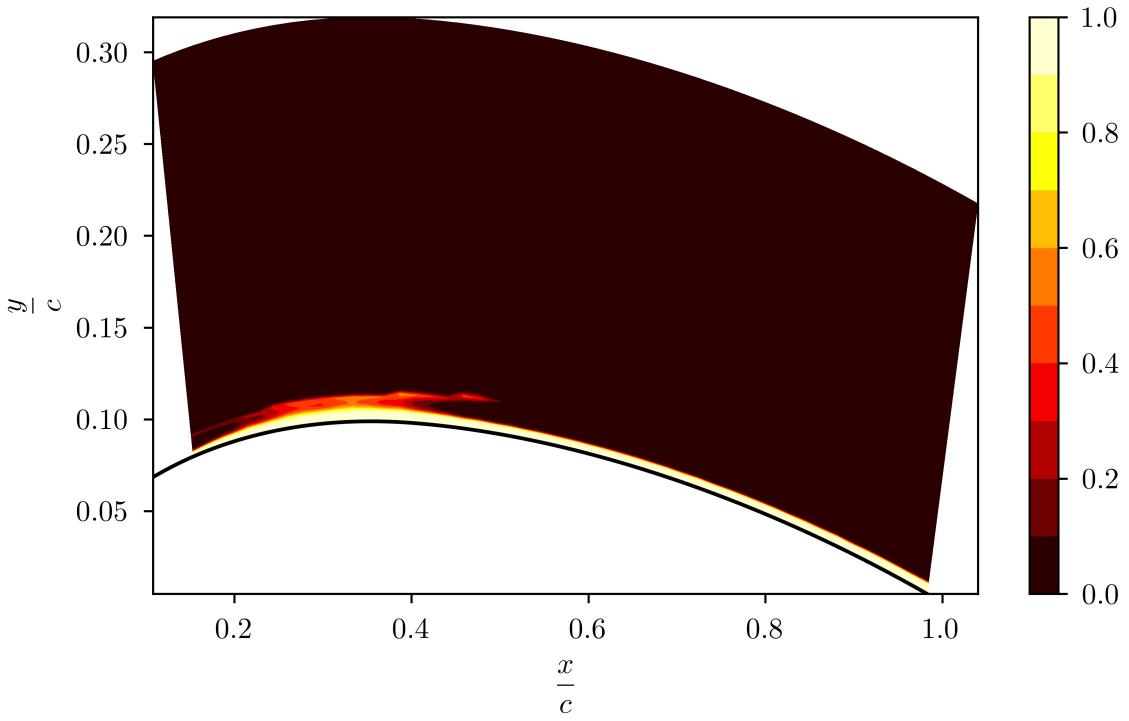
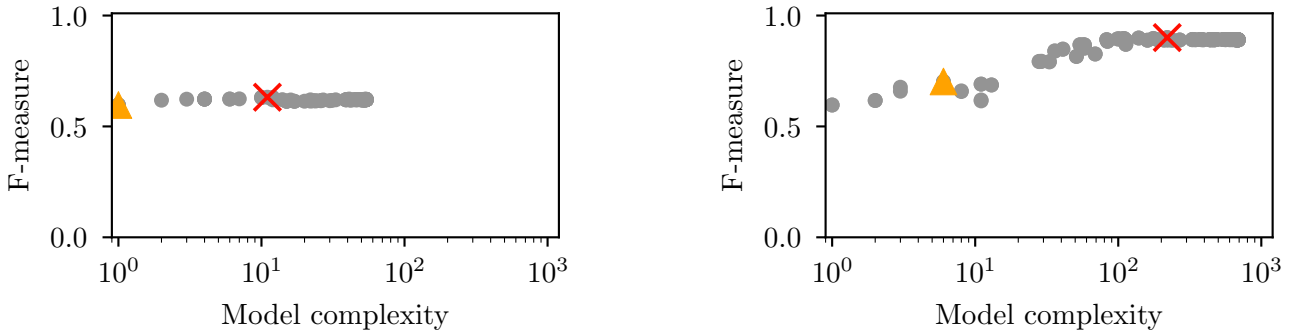


Figure 6.7.: Probability of active anisotropy metric on the NACA4412 aerofoil with $Re_c = 100\,000$ using the algebraic SpaRTA model defined in equation (6.10).



(a) Logistic Regression models.

(b) Models trained with the SpaRTA algorithm.

Figure 6.8.: Model performance plotted against complexity for logistic regression and SpaRTA models that identify the non-negativity metric. Both graphs show the generalisation performance to NACA data. A red cross indicates the best-performing model and a yellow triangle an algebraic SpaRTA model.

6.2. Non-negativity identification

Linear eddy viscosity models approximate the influence of turbulence on the mean flow with the eddy-viscosity approximation. The approximation defines the eddy viscosity in analogy to molecular viscosity, see section 2.2. A negative eddy viscosity appears in flows with strong mean velocity gradients and causes the eddy-viscosity assuption to become invalid, see [1, 6]. The identifier models learn to predict when the eddy viscosity becomes negative from local mean flow data. Both algorithms construct a variety of models. Thus, a model is selected that balances low model complexity, to allow physical insight, and high performance. Subsequently, the selected models are evaluated based on their quantitative and qualitative performance.

Model selection

The diagrams in figure 6.8 analyse the change in performance with varying model complexity. It provides insight into the capability of the algorithm to find suitable identifiers for the non-negativity metric. Furthermore, the plot provides a basis to select an interpretable identifier model which balances high performance and low complexity.

Figure 6.8a shows all logistic regression models and their performance. The complexity varies from a single term model to 56 terms, that is a linear combination of all physical and invariant features. The trend is approximately constant which implies that a single term model is comparable to a full model. Consequently, the model with complexity equal to one balances both parameters the best. However, inspection of the model structure reveals, that the simple model has the form $|w_2|k$ where w_2 is the model coefficient. Since k is positive semi-definite, the model predicts only active metrics. In addition, the high complexity models do not achieve significantly higher performances. This suggests that the linear combination of features cannot physically describe the non-negativity metric. The next section evaluates the performance for each test data set to investigate the finding.

In contrast, the SpaRTA models in figure 6.8b reach higher complexity with up to 690 active candidates, that is non-linear combinations of selected features. Their performance increases exponentially and converges towards the best-performing model (red cross) with an F-measure of 0.90. Thus, the complex identifier mod-

els are capable of identifying the non-negativity with a small bias. In comparison, an interpretable model is selected with 6 active candidates and a reduction in performance of 0.20 compared to the complex identifier. An enhanced model is available with 28 model terms which reaches a reduction of 0.10 compared to the best-performing model. The large amount of 28 terms makes physical insight into the model difficult. Instead, the algebraic model with 6 model terms is selected accepting a larger bias in the prediction of negative eddy viscosities.

For other test data sets, the performance reduction for interpretable SpaRTA models is similar with an average decrease of 0.19 or 31 %. Interestingly, almost all selected SpaRTA models share a mutual model structure:

$$M_{1,\nu_t} = -59q_2^{3.0}q_6^{5.0} - 27q_{10}q_{27}q_6^{2.0} + 24q_2^{3.0}q_{27}^{2.0}q_6 + 19q_2^{2.5} - 18q_2^{2.0} - 3.6q_{27}^{2.0} - 0.78q_{10}, \tag{6.11}$$

$$M_{2,\nu_t} = -61q_2^{3.0}q_6^{5.0} - 26q_{10}q_{27}q_6^{2.0} + 25q_2^{3.0}q_{27}^{2.0}q_6 + 20q_2^{2.5} - 19q_2^{2.0} - 3.6q_{27}^{2.0} - 1.02q_{10} + 0.22q_{10}q_2^{0.5}, \tag{6.12}$$

$$M_{3,\nu_t} = -59q_2^{3.0}q_6^{5.0} - 28q_{10}q_{27}q_6^{2.0} + 25q_2^{3.0}q_{27}^{2.0}q_6 + 18q_2^{2.5} - 18q_2^{2.0} - 3.7q_{27}^{2.0} - 0.84q_{10}, \tag{6.13}$$

$$M_{4,\nu_t} = -60q_2^{3.0}q_6^{5.0} - 26q_{10}q_{27}q_6^{2.0} + 26q_2^{3.0}q_{27}^{2.0}q_6 + 19q_2^{2.5} - 18q_2^{2.0} - 3.7q_{27}^{2.0} - 0.80q_{10}, \tag{6.14}$$

$$M_{5,\nu_t} = -47q_2^{2.0} + 38q_2^{2.5} + 30q_2^{3.0} + 3.7q_2^{3.5} - 13q_2^{4.0} - 12q_2^{4.5}. \tag{6.15}$$

The model coefficients are presented with two significant figures. Since all features are normalised onto the range $[-1, 1]$, the magnitude of coefficients indicates the relative importance of distinct terms.

The models in equations (6.11) to (6.14) use the turbulence intensity feature (q_2) in multiple candidates with power of 2.0, 2.5 and 3.0, that is they strongly depend on the turbulence intensity. Moreover, the invariant $\overline{\mathbf{PK}}$ (q_{27}) and the deviation from parallel shear flow (q_6) appear in most of the terms with high magnitudes and with exponents greater than one. Surprisingly, the pressure gradient along a streamline (q_{10}) has smaller coefficients than the invariant q_{27}. Although q_{10} provides direct information about the pressure gradient, the invariant feature, including the gradient of pressure and k, appears more versatile for the identifier. In stark contrast, the model in equation (6.15) utilises a simple power series of the turbulence intensity feature q_2 for identifications. This implies a strong dependence on either a separation region or boundary layers.

Quantitative performance

The receiver operating characteristic in figure 6.9 compares the performance of the selected models for each test data set. The diagram presents the performance of a random identifier with a black line, and an ideal identifier would lie in the upper left corner. Full symbols depict logistic regression models while half-filled symbols present the performance of SpaRTA models.

The logistic regression model performances lie all in the upper right corner in figure 6.9. The model predicts all positive points and achieves a true-positive rate equal to one, but reaches an almost equally high false-positive rate for all data sets. Thus, the logistic regression models achieve a random performance by predicting virtually only an active non-negativity metric. This supports the above suggestion: logistic regression models, with linear combinations of features, cannot identify the non-negativity metric.

The SpaRTA model achieve a better overall performance than the logistic regression models. With a reduction in both rates, selected SpaRTA models reduce false-positive rates strongly to an average of 33 % and mean true-positive rates to 77 %. This reduction shifts the performance on each test data closer to the ideal performance. Consequently, the SpaRTA algorithm discovers models with higher performance compared to the logistic regression algorithm.

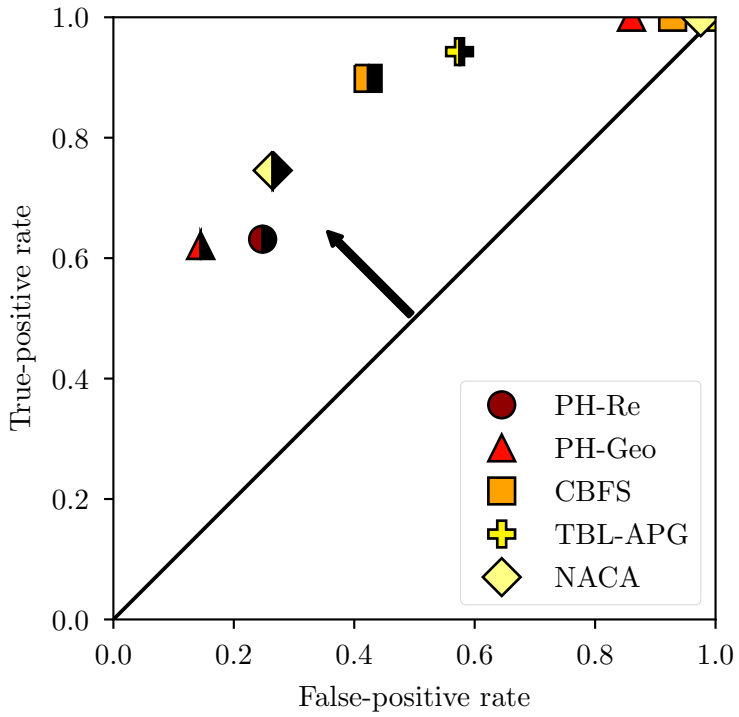


Figure 6.9.: Receiver operating characteristic for non-negativity identifiers. The shape of a symbol refers to the (unseen) test data set. Full symbols represent identifier models trained with logistic regression and half-filled symbols correspond to SpaRTA models. An optimal identifier lies in the upper left corner and models improve in performance along the arrow's direction.

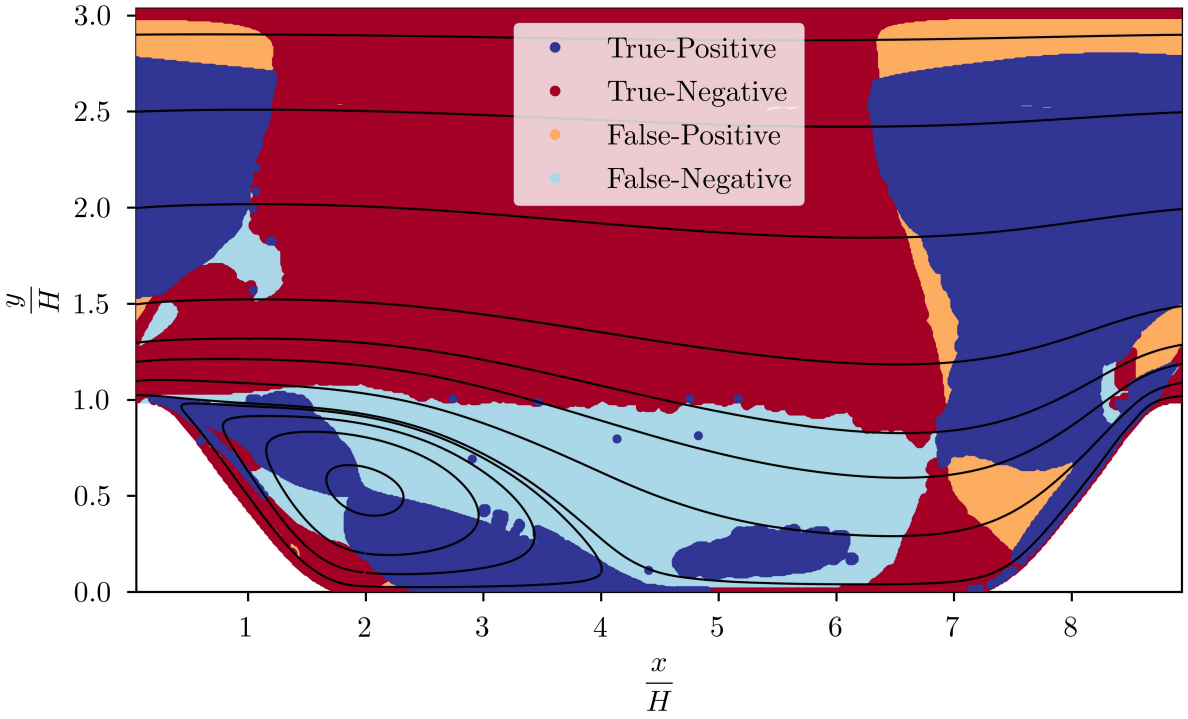


Figure 6.10.: Non-negativity identification for the flow over periodic hills at $Re_H = 5600$ using the algebraic SpaRTA model defined in equation (6.12).

Despite the improvement in the SpaRTA models, the predictions on each test set reach moderate performances. For example, the PH-Geo model predicts 62 % of all active metrics. While the mispredictions are low with 14 %, the model does not predict a considerable amount of the points with negative eddy-viscosities. Predictions on the NACA data result in a higher true-positive rate of 74 %, but also an increased false-positive rate to 26 %. This trend continues to worse performances on the curved step and flat plate geometry, that is the distance to the ideal classifier increases.

In summary, all identifications show a considerable margin towards an ideal classifier. The margin is a result of the aforementioned larger bias of interpretable models compared to complex models.

Qualitative performance

The quantitative evaluation of model performances shows that interpretable logistic regression models predict non-negativity almost at random. Hence, these models are not considered in the qualitative evaluation. In contrast, the interpretable SpaRTA models reach moderate performances as a consequence of their bias. In order to assess the capability of the SpaRTA models, predictions on a flow with separation (PH-Geo) and strong pressure gradients (NACA) are investigated. The model predictions provide insight into the capability to identify negative eddy viscosities.

Figure 6.10 presents predictions on the unscaled periodic hills geometry with $Re_H = 5600$. The predictions are marked by color with respect to the confusion matrix: dark blue and red points correspond to true-positives and true-negatives whereas orange and cyan points are false-positives and false-negatives, respectively. The coloured confusion matrix is depicted in figure 3.8.

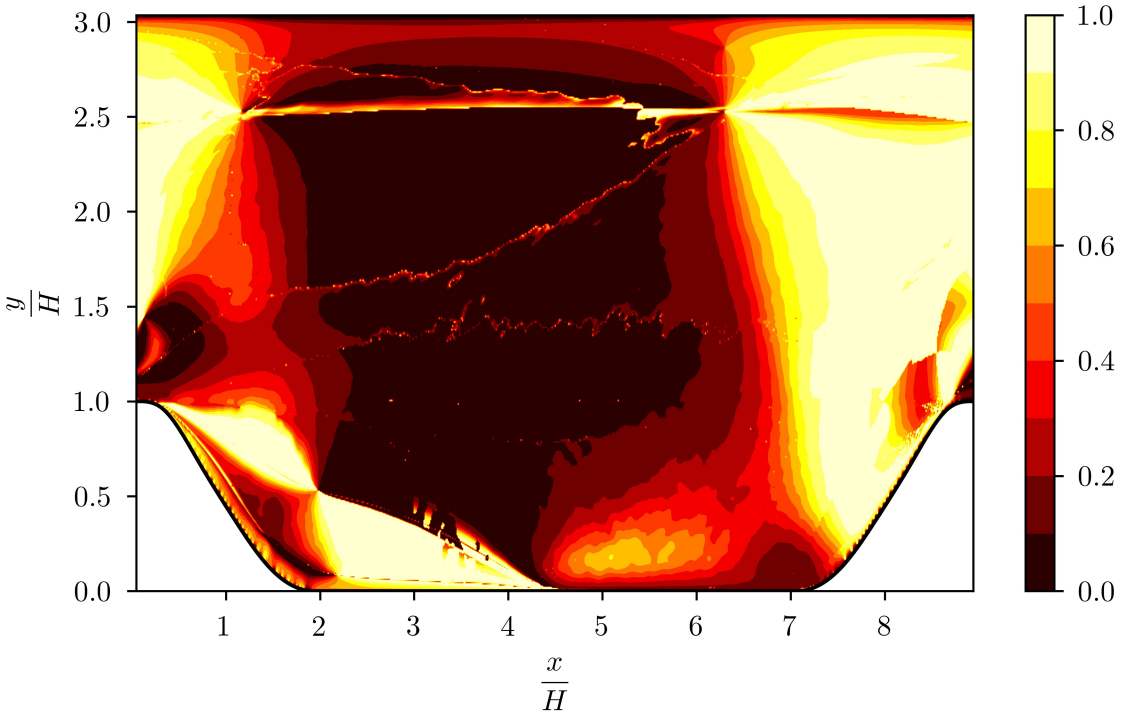


Figure 6.11.: Smooth identification of non-negativity for the flow over periodic hills at $Re_H = 5600$ using the algebraic SpaRTA model defined in equation (6.11).

The true-positives and -negatives indicate that negative eddy viscosities appear in two regions: the separation region and its wake as well as in the accelerated flow across the hill. Model predictions for the recirculation region are to a large extent true. The model correctly captures non-negativity along the locus of the recirculation and near-wall region. Both regions are well described by the non-orthogonality feature (q_6) which is dominant in the model in equation (6.11), cf. chapter 5. Thus, the strong contribution of this feature to the model in is identified by the SpaRTA algorithm. In contrast, the wake of the separation, where turbulent structures are transported (splatting), is largely misclassified. This suggests that the model is missing information about the convection of turbulence. Interestingly, the invariant $\overline{\mathbf{PK}}$ (q_{27}) provides information about the gradient of the turbulent kinetic energy, but, apparently, it is not specifically informative of the wake.

Apart from the recirculation, figure 6.10 shows predictions of the accelerated flow across the hill crest. The bulk of this region is correctly identified. Nonetheless, regions of false predictions occur adjacent to the bulk regions of negative eddy viscosity. False-positives and -negatives result on the upstream and downstream side of the hill which suggests insufficient information about the effect of strong straining on the turbulence.

In addition to the confusion plots, applying the logistic function to the model in equation (6.11) leads to a smooth identification of RANS uncertainty. Figure 6.11 presents a contour of the probability of the positive class $\Pr(y_{\nu_t})$, that is an active non-negativity, where a point with $\Pr(y_{\nu_t}) > 0.5$ is classified as active. The smooth prediction allows to analyse the uncertainty of identifications.

The identification around the locus of the recirculation shows little uncertainty with a probability $\Pr(y_{\nu_t}) > 0.9$. In contrast, the wake of the separation shows almost no uncertainty in the region $2 \leq x/c \leq 4.5$, even

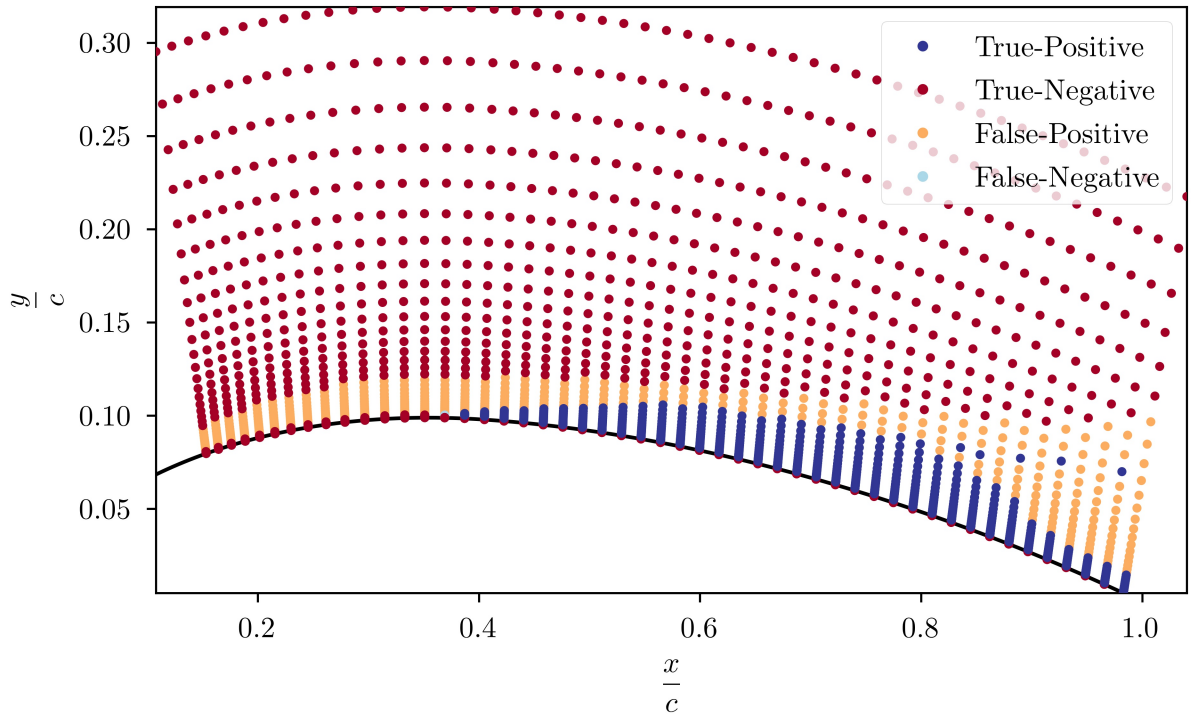


Figure 6.12.: Identification of negative eddy viscosity on the NACA4412 profile with $Re_c = 100\,000$. The confusion matrix is evaluated with predictions performed by the algebraic SpaRTA model defined in equation (6.15).

though, it is misclassified. The probability is lower than 0.2 for this region. This supports the above finding, that additional information about the convection of turbulence can improve non-negativity predictions. Similarly, the accelerated flow regions do not show strong uncertainty even in misclassified regions, for example, close to the upper wall.

In summary, the mispredictions in the separation and accelerated flow suggest that the models cannot adequatly identify rapid distortions which lead to violations of Boussinesq's hypothesis, see [1].

Figure 6.12 shows non-negativity identifications for the NACA4412 aerofoil. Identifications are performed with the interpretable model in equation (6.15). Looking at the overall predictions, the flow shows a large region of negative eddy viscosity that stretches from $x/c = 0.4$ to the end of the profile. The growth and shrinkage of the region is a consequence of the strong adverse pressure gradient. The pressure gradient steeply increases towards its maximum at $x/c = 0.4$ and decreases aft of $x/c = 0.8$. It effects the turbulence by strengthening turbulence production near the wall and convecting turbulent structures towards the outer layer, see [63, 55]. This influence of the turbulence requires a negative eddy viscosity and, thus, a violation of Boussinesq's hypothesis.

The identifier model overpredicts the region of negative eddy viscosity. It identifies a growing region of negative eddy viscosity along the complete profile, see the false-positives in figure 6.12. Since the model depends solely on the turbulent kinetic energy k, the prediction is similar to a boundary layer. Consequently, it wrongly activates the error metric in front of the profile's crest and does not account for the region's shrinking aft of $x/c = 0.8$.

Figure 6.13 depicts the smooth model prediction for the same flow. Clearly, the freestream is predicted with

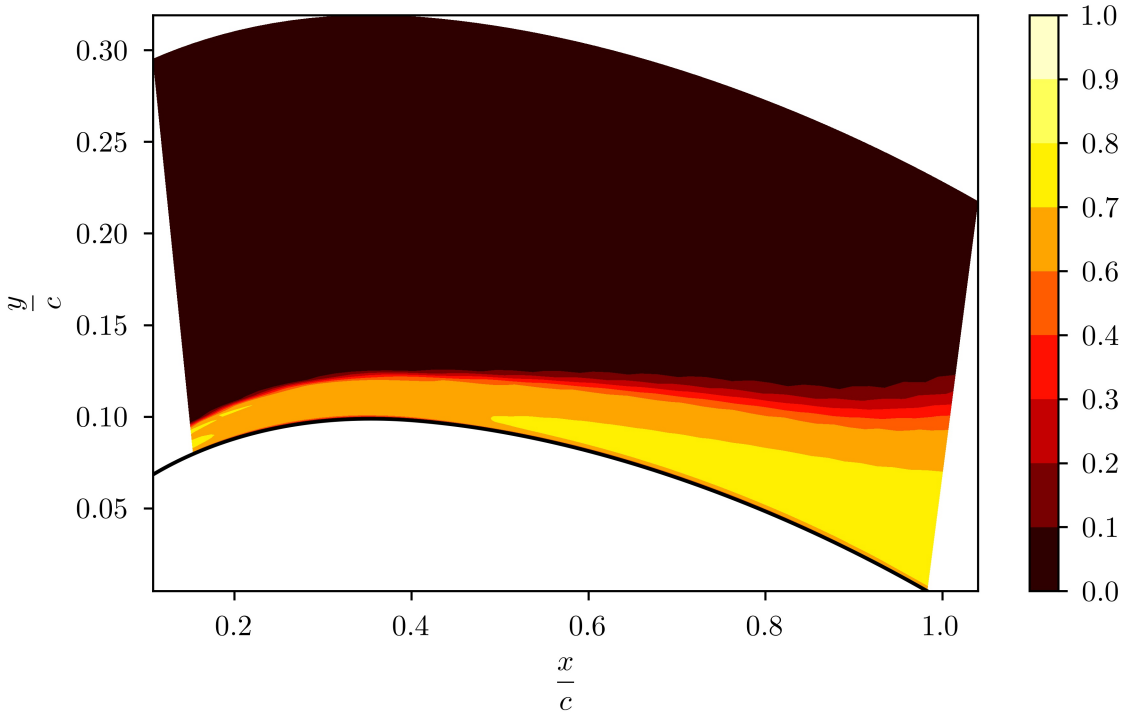


Figure 6.13.: Smooth identification of non-negativity on the NACA4412 profile with $Re_c = 100\,000$. The identifier model is defined in equation (6.11).

low uncertainty, that is the probability of an active metric is less than 0.1. Moreover, the erroneous prediction of a layer of negative eddy viscosity instead of a region is visible. Thus, the k-polynomial structure predicts a growing layer of non-negativity similar to the boundary layer while truly the negative eddy viscosity occurs in a bubble-like region.

All in all, the non-negativity identification is challenging. The low complexity of logistic regression models does not allow an adequate identification of the phenomena. On the other hand, SpaRTA models are capable of identifying non-negativity. However, the quality of interpretable models is limited. The selected interpretable models partially capture the effect of strong velocity gradients in accelerated flow and negative eddy viscosity in the separation bubble. Nonetheless, the wake region of the separation and strong adverse pressure gradients are not well predicted by the interpretable models. Thus, the identification of rapid distortions is only partially possible.

6.3. Algorithmic performance

The previous sections studied the performance of anisotropy and non-negativity identifier models focussed on one test data set. This section focusses on the performance of the models when averaged over all test data sets yielding the overall performance of each algorithm for one error metric. Additionally, limits of the interpretable models are investigated comparing them to non-interpretable models with more than 100 model terms. The comparisons utilise a receiver operating characteristic, cf. figure 3.9.

Figure 6.14 summarises the performance of the logistic regression and SpaRTA identifier for both error

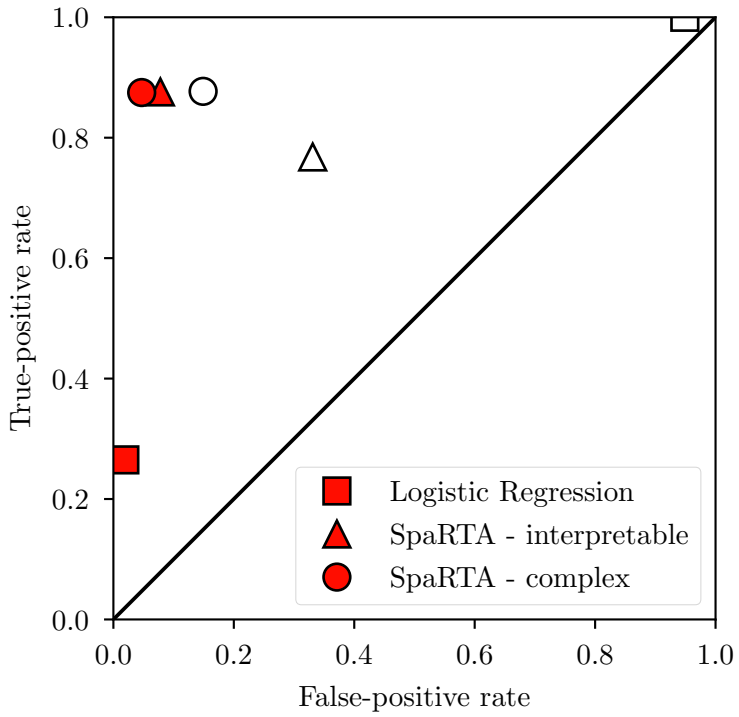


Figure 6.14.: Receiver operating characteristic with average performance per algorithm and error metric. Anisotropy identifier are marked with filled symbols and non-negativity identifier with hollow symbols. The mean performance of non-interpretable SpaRTA models are based on the identifiers with maximum F-measure for each test data set. An ideal identifier would lie in the upper left corner.

metrics. Looking at the anisotropy metric, interpretable SpaRTA models clearly outperform the Logistic regression models with a true-positive rate of 0.88, that is an increase of 330 %. Simultaneously, the false-positive rate increases marginally with 6 % additional false-positives. On the other hand, non-negativity identifier constructed with logistic regression cannot discriminate positive from negative eddy viscosities. This suggests, that the non-negativity metric requires higher complexity compared to anisotropy predictions. Using the SpaRTA models, non-negativity is discriminated decreasing mispredictions by 62 p.p. to 33 % and correct predictions by only 23 p.p.. Therefore, predictions of both error metrics boost with the non-linear combinations of features in the SpaRTA algorithm are largely boost identifications.

Beyond interpretable models, figure 6.14 shows the average performance of non-interpretable identifiers trained for both error metrics using the SpaRTA algorithm. The models are selected based on the greatest *F-measure* and reach complexities well above 100 terms, see figures 6.1b to 6.8b. Nonetheless, they represent the highest possible performance for the five selected features.

Evidently, the anisotropy identifier halve their misprediction rate, but achieve virtually no improvement in the true-positive rate. The identifiers appear more capable of distinguishing anisotropy near the wall leading to lower false-positive rates without substantially improving predictions of complexer anisotropy. Hence, the identification of anisotropic turbulence is not limited by the model complexity, but requires additional features for better predictions.

Figure 6.14 clearly shows, that non-negativity identifications enhance with high model complexity. The complex models predict 88 % of the points with negative eddy viscosity and mispredict 15 %, thus, shifting the performance closer to the ideal performance (upper left corner). Consequently, the non-negativity metric is likely to improve further by enriching the candidate library $\mathcal{B}$ with additional features and non-linear function transformations of the features.

7. Summary

The present work constructs data-driven identifier models for high uncertainty in Reynolds-averaged Navier-Stokes (RANS) solutions. The identification is based on violations of Boussinesq's hypothesis. Therefore, error metrics are employed which identify anisotropic turbulence and negative eddy viscosity. In order to improve predictions, the models smoothly predict the probability of a violation based on local mean flow quantities. This allows point-wise blending and switching of turbulence models in regions of high uncertainty.

The incorporated machine learning algorithms investigate the capabilities of interpretable identifier models. An interpretable model uses a sparse model structure to enable physical insight into the identification. Suitable models are constructed with two distinct algorithms: *Logistic Regression* uses a linear combination of features to construct models with minimum complexity; *SpaRTA* allows non-linear combinations of the features and uses sparsity constraints which allows models to predict complexer physics while achieving interpretable model structures. For computational tractability, the features for the SpaRTA algorithm are selected based on mutual information of features and both error metrics. All identifiers are assessed on a database of flows with strong pressure gradients and separation. The database is restricted to time-averaged high-fidelity data to test the capability of interpretable models.

7.1. Conclusions

The mutual information discloses the most-informative features with respect to each error metric. Accordingly, the anisotropy metric correlates with features which identify boundary layers and separation regions. These features include the turbulent time scale, turbulence intensity and wall-distance based Reynolds number. The negative eddy viscosity metric finds strong mutual information with features, that inform the identifier model about recirculation zones and pressure gradients.

The logistic regression models exhibit limited applicability for the identification of anisotropic turbulence and negative eddy viscosity. Identifier for negative eddy viscosity do not not reach favourable performances. Independent of the model's complexity, models do not achieve a significant ability to predict negative eddy viscosity. For anisotropic turbulence, models identify anisotropy dominantly close to the wall using 7 to 10 active model terms. However, the identification does not generalise well from the training to unseen flows. The performance evaluation shows limitations of identifier models to predict anisotropy on the flows included in the database. They capture at most 35 % of the grid points with anisotropic turbulence on a curved-backwards-facing step geometry. Thus, the analysis does not show favourable performance for the minimum-complexity identifier models.

The SpaRTA algorithm finds identifier models with more than 100 terms. Interpretable models require a reduction of the terms by two orders of magnitude and, as a consequence, reduce in performance. Anisotropy identifiers share a mutual model structure with 5 active terms. Active terms solely use the turbulent time scale, turbulence intensity and wall-distance based Reynolds number for the identification. The average performance reduction compared to a high-complexity model is 6 %. Similarly, the non-negativity identifiers

share a mutual model structure with 7 to 8 active candidates. The model structure dominantly uses features for turbulence intensity, pressure gradient and the deviation from parallel shear flow for identifications. An interpretable model structure decreased performance on average by 31 %.

Quantitative and qualitative model evaluations investigate identifier performances on flows characterised by either separation or strong pressure gradients. Considering anisotropic turbulence, identifier models are capable of identifying wall-blocking induced anisotropy. However, they do not capture the effect of separation phenomena, varying boundary layer thickness and strong adverse pressure gradients on anisotropy. On the other hand, non-negativity identifications inside the recirculation zone and in accelerated flow are possible. Still, their accuracy is limited. The identifiers do not physically predict negative eddy viscosity as a result of convected turbulence and strong adverse pressure gradients.

The overall performance of interpretable and complex models are compared to evaluate limitations of the interpretable models. Complex anisotropy identifiers half the false-positive rate to 5 %, but do not significantly increase in the true-positive rate. Hence, model complexity on its own does not increase anisotropy predictions and, instead, additional features are required. For predictions of negative eddy viscosity, the increase in model complexity leads to a boost in true- and false-positive rate of 11 p.p. and 18 p.p., respectively. Clearly, the non-negativity requires higher model complexity for accurate identifications.

7.2. Recommendations

The analysis of identifier models constructed with both algorithms reveal possible improvements and further studies for both, features and algorithms.

The logistic regression algorithm constructs models as linear combinations of the physical and invariant features. Within the investigations, these models do not achieve favourable performance. Therefore, future approaches for identifier models require either highly-informative features for the error metrics or increased model complexity, similar to the SpaRTA algorithm.

The SpaRTA algorithm is capable of finding interpretable models for the anisotropy identification. However, the models do not predict the complex physical phenomena of turbulence anisotropy or negative eddy viscosity accurately. Improvements are possible by including a larger number of features to the SpaRTA algorithm which then provide further information about the distinct flow physics. Moreover, constructing models for a narrower range of flows potentially identifies very informative model terms. Subsequently, this additional knowledge enriches the feature selection for the SpaRTA algorithm.

The algorithm in this study are both deterministic symbolic regression algorithms. Deterministic algorithms derive models from discrete features or candidates for the model terms. An alternative are stochastic symbolic regression algorithms. Stochastic algorithms are known to explore the model space efficiently and, potentially, find improved identifier models. One example is Gene-expression programming. This algorithm is applied in [20, 21] to derive correction models for the anisotropy tensor.

Bibliography

[1] Pope, S. B., *Turbulent flows*, Cambridge University Press, ISBN 978-0511840531, 2001

[2] Slotnick, J., Khodadoust, A., Alonso, J., Darmofal, D., Gropp, W., Lurie, E., and Mavriplis, D., *CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences*, Technical report, NASA, 2014, URL: `https://ntrs.nasa.gov/citations/20140003093`

[3] Launder, B. E. and Sharma, B., *Application of the energy-dissipation model of turbulence to the calculation of flow near a spinning disc*, Letters in heat and mass transfer, Vol. 1, No. 2, pp. 131–137, 1974

[4] Wilcox, D. C., *Formulation of the k-w turbulence model revisited*, AIAA Journal, Vol. 46, No. 11, pp. 2823–2838, 2008

[5] Spalart, P. and Allmaras, S., *A one-equation turbulence model for aerodynamic flows*, *30th Aerospace Sciences Meeting and Exhibit*, Jan. 6–9, 1992, Reno, USA, 1992

[6] Leschziner, M., *Statistical turbulence modelling for fluid dynamics - demystified: An introductory text for graduate engineering students*. Imperial College Press, ISBN 978-1783266609, 2015

[7] Wilcox, D. C., *Turbulence modeling for CFD*, DCW industries, ISBN 978-1-928729-08-2, 2006

[8] Duraisamy, K., Iaccarino, G., and Xiao, H., *Turbulence modeling in the age of data*, Annual Review of Fluid Mechanics, Vol. 51, No. 1, pp. 357–377, 2019

[9] Wu, J.-L., Xiao, H., and Paterson, E., *Physics-informed machine learning approach for augmenting turbulence models: a comprehensive framework*, Physical Review Fluids, Vol. 3, No. 074602, 2018

[10] Menter, F. R., Kuntz, M., and Langtry, R., *Ten years of industrial experience with the SST turbulence model*, Turbulence, heat and mass transfer, Vol. 4, No. 1, pp. 625–632, 2003

[11] Gorlé, C., Larsson, J., Emory, M., and Iaccarino, G., *The deviation from parallel shear flow as an indicator of linear eddy-viscosity model inaccuracy*, Physics of Fluids, Vol. 26, No. 051702, 2014

[12] Ling, J. and Templeton, J., *Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty*, Physics of Fluids, Vol. 27, No. 085103, 2015

[13] Ling, J., Jones, R., and Templeton, J., *Machine learning strategies for systems with invariance properties*, Journal of Computational Physics, Vol. 318, pp. 22–35, 2016

[14] Ling, J., Kurzawski, A., and Templeton, J., *Reynolds averaged turbulence modelling using deep neural networks with embedded invariance*, Journal of Fluid Mechanics, Vol. 807, pp. 155–166, 2016

[15] Wang, J.-X., Wu, J.-L., and Xiao, H., *Physics-informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on DNS data*, Physical Review Fluids, Vol. 2, No. 034603, 2017

[16] Koza, J. R., *Genetic programming as a means for programming computers by natural selection*, Statistics and computing, Vol. 4, No. 2, pp. 87–112, 1994

[17] Bishop, C. M., *Pattern recognition and machine learning*, Springer, ISBN 978-0387310732, 2006

[18] Hastie, T., Tibshirani, R., and Friedman, J., *The elements of statistical learning: Data mining, inference, and prediction*, Springer Science & Business Media, ISBN 978-0-387-84858-7, 2009

[19] Schmelzer, M., Dwight, R. P., and Cinnella, P., *Discovery of algebraic reynolds-stress models using sparse symbolic regression*, Flow, Turbulence and Combustion, Vol. 104, pp. 579–603, 2020

[20] Weatheritt, J. and Sandberg, R., *A novel evolutionary algorithm applied to algebraic modifications of the RANS stress-strain relationship*, Journal of Computational Physics, Vol. 325, pp. 22–37, 2016

[21] Weatheritt, J. and Sandberg, R., *The development of algrabice stress models using a novel evolutionary algorithm*, International Journal of Heat and Fluid Flow, Vol. 68, pp. 298–318, 2017

[22] Blazek, J., *Computational fluid dynamics: principles and applications*, Butterworth-Heinemann, ISBN 978-0-08-099995-1, 2015

[23] Fröhlich, J., *Large Eddy Simulation turbulenter Strömungen*, Springer, ISBN 978-3-8351-0104-3, 2006

[24] Schumann, U., *Realizability of Reynolds-stress turbulence models*, The Physics of Fluids, Vol. 20, No. 721, pp. 721–725, 1977

[25] Banerjee, S., Krahl, R., Durst, F., and Zenger, C., *Presentation of anisotropy properties of turbulence, invariants versus eigenvalue approaches*, Journal of Turbulence, No. 8, 2007

[26] Lumley, J. L. and Newman, G. R., *The return to isotropy of homogeneous turbulence*, Journal of Fluid Mechanics, Vol. 82, pp. 161–178, 1977

[27] Lumley, J. L., *Computational modeling of turbulent flows*, in: Yih, C.-S. (Eds.) *Advances in Applied Mechanics*, Elsevier, 1979

[28] Emory, M. and Iaccarino, G., *Visualizing turbulence anisotropy in the spatial domain with componentality contours*, Center for Turbulence Research Annual Research Briefs, pp. 123–138, 2014

[29] Mohri, M., Rostamizadeh, A., and Talwalkar, A., *Foundations of machine learning*, MIT press, ISBN 978-0262039406, 2018

[30] Russell, S. J. and Norvig, P., *Artificial intelligence a modern approach*, Pearson Education, ISBN 978-0-13-604259-4, 2010

[31] Bergstra, J. and Bengio, Y., *Random search for hyper-parameter optimization*, The Journal of Machine Learning Research, Vol. 13, No. 1, pp. 281–305, 2012

[32] Tang, J., Alelyani, S., and Liu, H., *Feature selection for classification: a review*, in: Aggarwal, C. C. (Eds.) *Data classification algorithms and applications*, CRC press, 2014

[33] Dash, M. and Liu, H., *Feature selection for classification*, Intelligent data analysis, Vol. 1, No. 3, pp. 131–156, 1997

[34] McConaghy, T., *FFX: fast, scalable, deterministic symbolic regression technology*, in: Riolo, R., Vladislavleva, E., and Moore, J. H. (Eds.) *Genetic Programming Theory and Practice IX*, Springer, 2011

[35] Craft, T. J., Launder, B. E., and Suga, K., *Development and application of a cubic eddy-viscosity model of turbulence*, International Journal of Heat and Fluid Flow, Vol. 17, pp. 108–115, 1996

[36] Wiesler, S. and Ney, H., *A convergence analysis of log-linear training*, in: Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q. (Eds.) *Advances in Neural Information Processing Systems 24*, Curran Associates, 2011

[37] Chakraborty, P., Balachandar, S., and Adrian, R. J., *On the relationships between local vortex identification schemes*, Journal of Fluid Mechanics, Vol. 535, pp. 189–214, 2005

[38] Johnson, R. W., *Handbook of fluid dynamics*, CRC Press, ISBN 978-1-4398-4957-6, 2016

[39] Defazio, A., Bach, F., and Lacoste-Julien, S., *Saga: a fast incremental gradient method with support for non-strongly convex composite objectives*, in: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (Eds.) *Advances in Neural Information Processing Systems 27*, Curran Associates, 2014

[40] Zou, H. and Hastie, T., *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), Vol. 67, No. 2, pp. 301–320, 2005

[41] Friedman, J., Hastie, T., and Tibshirani, R., *Regularization paths for generalized linear models via coordinate descent*. Journal of statistical software, Vol. 33, pp. 1–22, 2010

[42] Brunton, S. L. and Kutz, J. N., *Data-driven science and engineering: machine learning, dynamical systems, and control*, Cambridge University Press, ISBN 978-1108380690, 2019

[43] Smith, R. C., *Uncertainty quantification: theory, implementation, and applications*, SIAM, ISBN 978-1-611973-21-1, 2013

[44] Kraskov, A., Stögbauer, H., and Grassberger, P., *Estimating mutual information*, Phys. Rev. E, Vol. 69, p. 066138, 2004

[45] Papana, A. and Kugiumtzis, D., *Evaluation of mutual information estimators on nonlinear dynamic systems*, Nonlinear Phenomena in Complex Systems, Vol 11, No 2, pp 225-232, 2008, 2008

[46] Walters-Williams, J. and Li, Y., *Estimation of mutual information: a survey*, *4th International Conference on Rough Sets and Knowledge Technology*, July 14–16, 2009, Gold Coast, Australia, 2009

[47] Ross, B. C., *Mutual information between discrete and continuous data sets*, PLOS ONE, Vol. 9, No. 2, pp. 1–5, 2014

[48] Weiss, G. M., *Mining with rarity: a unifying framework*, SIGKDD Explor. Newsl. Vol. 6, No. 1, pp. 7–19, 2004

[49] Bronstein, I. N., Hromkovic, J., Luderer, B., Schwarz, H.-R., Blath, J., Schied, A., Dempe, S., Wanka, G., and Gottwald, S., *Taschenbuch der Mathematik*, Springer-Verlag, 2012

[50] Fawcett, T., *An introduction to roc analysis*, Pattern Recognition Letters, Vol. 27, No. 8, pp. 861–874, 2006

[51] Wang, J.-X., Huang, J., Duan, L., and Xiao, H., *Prediction of reynolds stresses in high-Mach-number turbulent boundary layers using physics-informed machine learning*, Theoretical and Computational Fluid Dynamics, Vol. 33, No. 1, pp. 1–19, 2019

[52] Breuer, M., Peller, N., Rapp, C., and Manhart, M., *Flow over periodic hills–numerical and experimental study in a wide range of reynolds numbers*, Computers & Fluids, Vol. 38, No. 2, pp. 433–457, 2009

[53] Xiao, H., Wu, J.-L., Laizet, S., and Duan, L., *Flows over periodic hills of parameterized geometries: a dataset for data-driven turbulence modeling from direct simulations*, Computers & Fluids, p. 104431, 2020

[54] Bentaleb, Y., Lardeau, S., and Leschziner, M. A., *Large-Eddy Simulation of turbulent boundary layer separation from a rounded step*, Journal of Turbulence, No. 13, 2012

[55] Bobke, A., Vinuesa, R., Örlü, R., and Schlatter, P., *History effects and near equilibrium in adverse-pressure-gradient turbulent boundary layers*, Journal of Fluid Mechanics, Vol. 820, pp. 667–692, 2017

[56] Vinuesa, R., Negi, P. S., Atzori, M., Hanifi, A., Henningson, D. S., and Schlatter, P., *Turbulent boundary layers around wing sections up to Re_c=1,000,000*, International Journal of Heat and Fluid Flow, Vol. 72, pp. 86–99, 2018

[57] Tanarro, Á., Vinuesa, R., and Schlatter, P., *Effect of adverse pressure gradients on turbulent wing boundary layers*, Journal of Fluid Mechanics, Vol. 883, No. A8, 2020

[58] Fröhlich, J., Mellen, C. P., Rodi, W., Temmerman, L., and Leschziner, M. A., *Highly resolved large-eddy simulation of separated flow in a channel with streamwise periodic constrictions*, Journal of Fluid Mechanics, Vol. 526, p. 19, 2005

[59] Lardeau, S. and Leschziner, M., *The interaction of round synthetic jets with a turbulent boundary layer separating from a rounded ramp*, Journal of fluid mechanics, Vol. 683, pp. 172–211, 2011

[60] Zhang, S. and Zhong, S., *An experimental investigation of turbulent flow separation control by an array of synthetic jets*, AIAA 2010-4582, *5th Flow Control Conference*, June 28–July 1, 2010, Chicago, USA, 2010

[61] Townsend, A. A., *The structure of turbulent shear flow*, Journal of Fluid Mechanics, Vol. 1, No. 5, pp. 554–560, 1956

[62] Negi, P. S., Vinuesa, R., Schlatter, P., Hanifi, A., and Henningson, D. S., *Unsteady aerodynamic effects in pitching airfoils studied through large-eddy simulations*, *Tenth International Symposium on Turbulence and Shear Flow Phenomena*, July 7–Sept. 7, 2017, Chicago, USA, 2017

[63] Hosseini, S., Vinuesa, R., Schlatter, P., Hanifi, A., and Henningson, D., *Direct numerical simulation of the flow around a wing section at moderate reynolds number*, International Journal of Heat and Fluid Flow, Vol. 61, pp. 117–128, 2016

List of Figures

List of Tables

A. Database statistics

The database includes a variety of data sets. Each data set includes a number of cases with varying geometry, Reynolds number or inflow conditions. This chapter provides information on the configuration of each case along with the amount of data points, the relative amount of active error metrics for a case and for the total data set.

Case names	Re_H	L_c	N	Active y_{ν_t}	Active y_{II}
PH-Re-700	700	$H = 1$	60 500	33 %	32 %
PH-Re-1400	1400	$H = 1$	60 500	35 %	27 %
PH-Re-2800	2800	$H = 1$	60 500	38 %	21 %
PH-Re-5600	5600	$H = 1$	60 500	39 %	16 %
PH-Re-10595	10595	$H = 1$	60 500	40 %	13 %
Total	-	-	302 500	37 %	22 %

Table A.1.: Summarised data for all PH-Re cases. Re_H is the Reynolds number with respect to hill height H. L_c is the characteristic length. N is the number of samples. The percentage of active labels for the non-negativity and anisotropy error metrics are shown with y_{ν_t} and y_{II}, respectively.

Case names	Re_H	L_c	N	Active y_{ν_t}	Active y_{II}
PH-Geo-05	5600	$H = 1$	283 360	44 %	4 %
PH-Geo-08	5600	$H = 1$	271 040	42 %	4 %
PH-Geo-10	5600	$H = 1$	295 680	41 %	4 %
PH-Geo-12	5600	$H = 1$	320 320	39 %	5 %
PH-Geo-15	5600	$H = 1$	359 590	39 %	5 %
Total	-	-	1 529 990	41 %	4 %

Table A.2.: Summarised data for all PH-Geo cases. Re_H is the Reynolds number with respect to hill height H. L_c is the characteristic length. N is the number of samples. The percentage of active labels for the non-negativity and anisotropy error metrics are shown with y_{ν_t} and y_{II}, respectively.

Case names	Re_H	L_c	N	Active y_{ν_t}	Active y_{II}
CBFS	13700	$H = 1$	122 880	9 %	5 %

Table A.3.: Summarised data for the CBFS case. Re_H is the Reynolds number with respect to the step height H. L_c is the characteristic length. L_c is the characteristic length and H the hill height. N is the number of samples. The percentage of active labels for the non-negativity and anisotropy error metrics are shown with y_{ν_t} and y_{II}, respectively.

Case names	Re_θ	L_c	N	Active y_{ν_t}	Active y_{II}
TBL-APG-Bobke-b1	$910 \leq Re_\theta \leq 3360$	$\delta_{99}(x)$	472 570	2 %	5 %
TBL-APG-Bobke-m13	$990 \leq Re_\theta \leq 3515$	$\delta_{99}(x)$	472 570	2 %	5 %
TBL-APG-Bobke-m16	$1010 \leq Re_\theta \leq 4000$	$\delta_{99}(x)$	566 770	5 %	4 %
TBL-APG-Bobke-m18	$990 \leq Re_\theta \leq 4320$	$\delta_{99}(x)$	629 570	6 %	3 %
Total	-	-	2 708 250	6 %	6 %

Table A.4.: Summarised data for all TBL-APG cases. Re_θ is the momentum thickness Reynolds number. L_c is the characteristic length and δ_{99} the 99 % boundary layer thickness. N is the number of samples. The percentage of active labels for the non-negativity and anisotropy error metrics are shown with y_{ν_t} and y_{II}, respectively.

Case names	Re_c	L_c	N	Active y_{ν_t}	Active y_{II}
NACA4412-top-1	100 000	$\delta_{99}(x)$	2600	36 %	29 %
NACA4412-top-2	200 000	$\delta_{99}(x)$	2600	38 %	26 %
NACA4412-top-4	400 000	$\delta_{99}(x)$	2600	39 %	20 %
NACA4412-top-10	1 000 000	$\delta_{99}(x)$	3050	41 %	22 %
NACA4412-bottom-1	100 000	$\delta_{99}(x)$	2600	39 %	36 %
NACA4412-bottom-2	200 000	$\delta_{99}(x)$	2600	46 %	33 %
NACA4412-bottom-4	400 000	$\delta_{99}(x)$	2600	48 %	27 %
NACA4412-bottom-10	1 000 000	$\delta_{99}(x)$	3050	49 %	25 %
NACA0012-top-4	400 000	$\delta_{99}(x)$	2600	40 %	24 %
Total	-	-	24 300	42 %	27 %

Table A.5.: Summarised data for all NACA cases. Re_c is the Reynolds number with respect to chord length. L_c is the characteristic length and δ_{99} the 99 % boundary layer thickness. N is the number of samples. The percentage of active labels for the non-negativity and anisotropy error metrics are shown with y_{ν_t} and y_{II}, respectively.