TU Delft

Delft University of Technology

CFD-driven symbolic identification of algebraic Reynolds-stress models

Ben Hassan Saïdi, Ismaïl; Schmelzer, Martin; Cinnella, Paola; Grasso, Francesco

DOI
10.1016/j.jcp.2022.111037

Publication date
2022

Document Version
Final published version

Published in
Journal of Computational Physics

Citation (APA)
Ben Hassan Saïdi, I., Schmelzer, M., Cinnella, P., & Grasso, F. (2022). CFD-driven symbolic identification of algebraic Reynolds-stress models. *Journal of Computational Physics*, *457*, Article 111037. https://doi.org/10.1016/j.jcp.2022.111037

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Green Open Access added to TU Delft Institutional Repository

'You share, we take care!' - Taverne project

https://www.openaccess.nl/en/you-share-we-take-care

Otherwise as indicated in the copyright section: the publisher is the copyright holder of this work and the author uses the Dutch legislation to make this work public.

Journal of Computational Physics 457 (2022) 111037

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

CFD-driven symbolic identification of algebraic Reynolds-stress models

Ismaïl Ben Hassan Saïdi [a,*], Martin Schmelzer [b], Paola Cinnella [c], Francesco Grasso [d]

[a] Institut Aérotechnique, CNAM, 15 Rue Marat, 78210 Saint-Cyr-l'École, France
[b] Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 2, Delft, the Netherlands
[c] Institut Jean Le Rond D'Alembert, Sorbonne Université, 4 Place Jussieu, 75005 Paris, France
[d] Laboratoire DynFluid, CNAM, 151 Boulevard de l'Hopital, 75013 Paris, France

ARTICLE INFO

Article history:
Received 18 April 2021
Received in revised form 11 January 2022
Accepted 29 January 2022
Available online 14 February 2022

Keywords:
Symbolic identification
Turbulence modelling
Separated flows

ABSTRACT

A CFD-driven deterministic symbolic identification algorithm for learning explicit algebraic Reynolds-stress models (EARSM) from high-fidelity data is developed building on the frozen-training SpaRTA algorithm of [1]. Corrections for the Reynolds stress tensor and the production of transported turbulent quantities of a baseline linear eddy viscosity model (LEVM) are expressed as functions of tensor polynomials selected from a library of candidate functions. The CFD-driven training consists in solving a blackbox optimization problem in which the fitness of candidate EARSM models is evaluated by running RANS simulations. The procedure enables training models against any target quantity of interest, computable as an output of the CFD model. Unlike the frozen-training approach, the proposed methodology is not restricted to data sets for which full fields of high-fidelity data, including second flow order statistics, are available. However, the solution of a high-dimensional expensive blackbox function optimization problem is required. Several steps are then undertaken to reduce the associated computational burden. First, a sensitivity analysis is used to identify the most influential terms and to reduce the dimensionality of the search space. Afterwards, the Constrained Optimization using Response Surface (CORS) algorithm, which approximates the black-box cost function using a response surface constructed from a limited number of CFD solves, is used to find the optimal model parameters. Model discovery and cross-validation is performed for three configurations of 2D turbulent separated flows in channels of variable section using different sets of training data to show the flexibility of the method. The discovered models are then applied to the prediction of an unseen 2D separated flow with higher Reynolds number and different geometry. The predictions of the discovered models for the new case are shown to be not only more accurate than the baseline LEVM, but also of a multi-purpose EARSM model derived from purely physical arguments. The proposed deterministic symbolic identification approach constitutes a promising candidate for building accurate and robust RANS models customized for a given class of flows at moderate computational cost.

© 2022 Elsevier Inc. All rights reserved.

* Corresponding author. Present affiliation: Laboratoire DynFluid, Arts et Métiers Institute of Technology, 151 Boulevard de l'Hopital, 75013 Paris, France.
E-mail address: ismail.benhassansaidi@ensam.eu (I. Ben Hassan Saïdi).

https://doi.org/10.1016/j.jcp.2022.111037
0021-9991/© 2022 Elsevier Inc. All rights reserved.

1. Introduction

Computational Fluid Dynamics (CFD) simulations of turbulent flows for industrial applications largely rely on the Reynolds-Averaged Navier-Stokes (RANS) equations supplemented with linear eddy viscosity models (LEVM) for the Reynolds stresses. Despite their generally inaccurate predictions of non-equilibrium turbulent flow with strong pressure gradients, streamline curvature, anisotropy and separation [2,3] LEVM are usually preferred to more sophisticated models –namely, Reynolds stress models (RSM)– due to their superior robustness and lower computational cost. RSM involve additional transport equations for the Reynolds stresses [4–6], which in principle can be derived exactly from the Navier–Stokes equations. Unfortunately, this process introduces a number of additional unclosed terms (pressure-strain correlations, diffusion and dissipation rates) that need suitable mathematical modeling, as well as the calibration of the associated closure coefficients. The predictive performance and the numerical robustness of RSM is sensitive to the modeling of these terms. Furthermore, RSM involve the solution of seven additional transport equations (for 3D flows), thus increasing computational cost compared to LEVM.

A trade-off between LEVM and RSM consists in the development of so-called generalized eddy viscosity models [7]. Within this class of models, Explicit Algebraic Reynolds Stress Models (EARSM) generalize the linear eddy viscosity concept by assuming that the anisotropic part of the Reynolds stress not only depends on the mean strain rate tensor S_{ij} but also on the rotation rate tensor Ω_{ij} through a non linear relationship. Specifically, Reynolds stress anisotropy is written as a tensor polynomial of S_{ij} and Ω_{ij} [7], the coefficients of which can be determined from the RSM equations simplified under local equilibrium [8] as functions of tensor invariants.

In this sense, classical EARSM partly inherit from modeling hypothesis used for the development of RSM. Furthermore, they also require the calibration of a number of closure coefficients, which is usually done for a set of simple canonical flows. Due to the uncertainties associated with the choice of both the mathematical structure and the closure coefficients the EARSM, as well as other RANS models, are not universally valid and may perform poorly for flow configurations very dissimilar to the calibration ones. The reader may refer to [9] for a discussion of RANS modeling uncertainties.

Several recent studies have illustrated the potential of data-driven approaches for the development of turbulence models, customized for specific classes of flows (see, e.g., [10,9]). Such approaches use machine learning techniques to infer the functional form of the Reynolds stress tensor from high-fidelity experimental or, most often, numerical data. While many of them use black-box machine learning such as Gaussian processes, Deep Neural Networks or Random Forests [11–13] to reconstruct or correct the Reynolds stress tensor, some others have considered open-box machine learning techniques, relying on the use of a set of candidate solutions or function dictionaries to build physically interpretable turbulence models similar to EARSM. In [14], this goal is achieved by using non deterministic symbolic regression based on an evolutionary algorithm (Gene Expression Programming, GEP). The GEP is efficient in constructing models with improved accuracy for separated flows. Moreover, learned models involving only a few nonlinear terms are found to exhibit lower training and prediction error and higher numerical robustness [15–17]. However, the random nature of the search algorithm discovers a model with a different mathematical form at each run, using the same training data. A deterministic symbolic identification algorithm called SpaRTA (Sparse Regression of Turbulent Stress Anisotropy) has been proposed in [1]. SpaRTA combines functions from a predefined library of candidates without any random recombination. The algorithm constrains the search towards algebraic models of minimal complexity using sparsity-promoting regression techniques [18,19]. This method has also shown its efficiency in improving RANS models for the simulation of separated flows behind steps or in convergent-divergent channels. Recently, SpaRTA has been applied to flows around 3D bluff body geometries [20], to wind turbine wake modeling [21] and it has been coupled with LES in a multi-fidelity optimization framework, allowing significant improvements over designs based on a simple LEVM [22]. Furthermore, the SpaRTA methodology has also been applied and extended to discover constitutive laws for elastic solids [23]. A similar deterministic identification approach has been also proposed in [24], whereby unclosed tensors in the RANS or Reynolds-stress transport equations are written as linear combinations of a tensor basis, whose coefficients are learned from synthetic or DNS data of simple flows using l_1-constrained mean square regression. The results are compared to a constitutive law using a similar decomposition, but with coefficients built on Neural Networks (NN) [12]. Because the deterministic identification methodology seeks to uncover underlying physics, far less data is required to achieve reasonable learned models. As a consequence, it represents potentially an excellent modeling construct for sparse datasets, such as those available from experiments.

A limitation of the above-mentioned symbolic identification algorithms relies in their training procedure. The latter uses a frozen-field (or *a priori* [25]) mode, whereby frozen mean-field and Reynolds stress high-fidelity data are used to train the constitutive equation for the Reynolds stresses. Once the training is completed, the resulting model is plugged in a CFD solver and used to predict a new flow. Such a strategy has the advantage of simplicity and extremely low cost (no CFD solves are required for the training procedure). The drawbacks are: 1) the need for full-field, high quality data for first and second-order flow statistics (namely, the Reynolds stresses); the latter are generally available only for simple geometries and low to moderate Reynolds numbers, for which DNS or well-resolved LES are possible; 2) due to the offline training, the learned model may cause numerical stiffness once coupled with a CFD code, requiring the introduction of relaxation coefficients [1] or implicit treatment to stabilize the CFD calculations and improve their conditioning [24]; 3) offline training using frozen high-fidelity fields does not guarantee consistency with the system of RANS equations (see [25] for a more detailed discussion).

The above-mentioned limitations have motivated recent work toward the development of so-called model-consistent or CFD-driven training algorithms [26–28]. In such an approach, the CFD model is embedded within the training process. This is expected to enable flexible use of incomplete or noisy training data for any flow quantity at hand. Moreover as each candidate model is evaluated through a RANS simulation, models preventing the convergence of the solution can be eliminated during the training process. Hence, learned models are expected to be robust. Finally, the CFD-driven approach ensures coupling of the mean field and turbulent transport equations at any stage of the training process, thus guaranteeing model consistency. However, the price to pay is a much higher computational cost of the training procedure, due to the multiple CFD solves required to evaluate model fitness.

In the present paper, we propose a CFD-driven extension of the SpaRTA algorithm [1] that evaluates the fitness of candidate models by running RANS simulations. The new method generates data-driven EARSM of the same general form of the frozen-training SpaRTA algorithm. Unlike the latter, however, it is applicable to configurations for which full fields of first and second-order turbulent statistics are not available (for example, experimental databases), as it can be trained with any kind of observable data. Examples include incomplete velocity or pressure fields, wall distributions of pressure or skin friction coefficients, or integral quantities such as aerodynamic coefficients. Such data can be made available through experiments in addition to simulations, opening the way to the development of data-driven models for complex, high-Reynolds flows. In order to reduce the computational burden associated with model training in a highly dimensional parameter space, a preliminary local sensitivity analysis is used to identify the most influential parameters. Subsequently, a pre-screening of candidate models is carried out using the Reynolds stress anisotropy barycentric map representation [29], so that parameter ranges leading to violation of the realizability conditions are discarded. An important advance over the state of the art is represented by the introduction of a surrogate-based CFD-driven training strategy. More precisely, we use a response surface methodology [30] to build an inexpensive surrogate of the costly black-box function from a moderate number of CFD evaluations. An approximate solution to the optimization problem is then computed by running the optimizer on the response surface. As in the frozen-training SpaRTA approach, the turbulence models obtained at the end of the procedure are sparse data-driven EARSM models involving a small subset of the initial library of candidate functions. Numerical experiments for selected separated flows are used to illustrate the accuracy and the generalization capabilities of the models.

The paper is organized as follows. In Section 2, we briefly recall the formulation of LEVM and EARSM models of turbulence. Datasets used for model training and validation are described in Section 3. In Section 4, we recall the main features of the frozen-training SpaRTA algorithm. Section 5 is devoted to numerical results and discussion. Conclusions and perspectives for future work are provided in Section 6.

2. Reynolds-averaged Navier–Stokes equations and turbulence models

The incompressible RANS equations read:

$$\frac{\partial \overline{u}_i}{\partial x_i} = 0 \tag{1}$$

$$\frac{\partial \overline{u}_i}{\partial t} + \frac{\partial (\overline{u}_i \overline{u}_j)}{\partial x_j} = -\frac{1}{\rho}\frac{\partial \overline{p}}{\partial x_i} + \frac{\partial}{\partial x_j}\left(\nu \frac{\partial \overline{u}_i}{\partial x_j} - \overline{u_i' u_j'}\right), \tag{2}$$

where ρ is the constant density, ν the fluid kinematic viscosity, $\overline{u_i}$ and $\overline{p}$ the averaged velocity components and pressure and u_i' the velocity fluctuations.

The unclosed term $\tau_{ij} = \overline{u_i' u_j'}$ is the kinematic Reynolds stress tensor. It can be decomposed into an isotropic part $\frac{2}{3}k\delta_{ij}$ and an anisotropic part $a_{ij} = 2kb_{ij}$ (with b_{ij} a non-dimensional anisotropy tensor), such that:

$$\tau_{ij} = \overline{u_i' u_j'} = 2k\left(b_{ij} + \frac{1}{3}\delta_{ij}\right), \tag{3}$$

where $k = tr(\tau_{ij})/2$ is the turbulent kinetic energy. Therefore, the modeling of the entire range of turbulent scales in the RANS approach reduces to the modeling of k and of the anisotropy tensor. This is generally done by introducing a constitutive relation for b_{ij}, i.e. a functional form relating its components to the mean field.

2.1. Boussinesq analogy and LEVM models

Most turbulence models currently used in industrial codes are LEVM. Such models rely on the so-called Boussinesq analogy, whereby the anisotropic part of τ_{ij} is written as a linear function of the mean rate-of-strain $S_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right)$:

$$b_{ij} = -\frac{\nu_t}{k} S_{ij} \tag{4}$$

with ν_t the eddy viscosity coefficient.

The LEVM constitutive relation (4) is then closed by expressing ν_t as a function of the mean flow. Most often, ν_t is computed using a transport-equation model such as the $k-\omega$ SST [31], used here as the baseline model. For completeness,

the SST $k-\omega$ equations are recalled in Appendix A. Although this model is believed to be one of the most accurate LEVM models on average, it still shows limitations in providing reliable results for flows with separation, streamline curvature or strong pressure gradients [2,3].

2.2. Augmented Boussinesq analogy and EARSM models

To overcome the limitations of LEVM models, more sophisticated approaches have been developed, such as the generalized eddy viscosity models [7]. In particular, EARSM generalize the linear eddy viscosity concept by assuming that the anisotropy of the Reynolds-stress depends not only on the strain rate tensor S_{ij} but also on the rotation rate tensor $\Omega_{ij} = \frac{1}{2}\left(\frac{\partial \overline{u}_i}{\partial x_j} - \frac{\partial \overline{u}_j}{\partial x_i}\right)$ through a non linear relationship.

Such models can be constructed from a baseline LEVM using the so-called augmented Boussinesq analogy:

$$b_{ij} = b_{ij}^{bl} + b_{ij}^{\Delta}, \tag{5}$$

where $b_{ij}^{bl} = -\frac{\nu_t}{k} S_{ij}$ and b_{ij}^{Δ} is the extra anisotropy tensor, used to correct the baseline Boussinesq model b_{ij}^{bl}.
Following this approach, an EARSM can be constructed by specifying a non linear relationship between the extra anisotropy tensor b_{ij}^{Δ} and the strain rate tensor and the rotation rate tensor. Following [7], the most general form of the anisotropic Reynolds-stress correction is derived via the Cayley-Hamilton theorem:

$$b_{ij}^{\Delta}(\hat{S}_{ij}, \hat{\Omega}_{ij}) = \sum_{\lambda=1}^{10} G_\lambda(I_1, ..., I_5) T_{ij}^{(\lambda)}, \tag{6}$$

using a minimal integrity basis of ten tensors $T_{ij}^{(\lambda)}$ and five invariants I_m, written as functions of the non-dimensional mean strain rate $\hat{S}_{ij} = \frac{1}{2\omega}\left(\frac{\partial \overline{u}_i}{\partial x_j} + \frac{\partial \overline{u}_j}{\partial x_i}\right)$ and rotation rate $\hat{\Omega}_{ij} = \frac{1}{2\omega}\left(\frac{\partial \overline{u}_i}{\partial x_j} - \frac{\partial \overline{u}_j}{\partial x_i}\right)$. In the following, we focus on two-dimensional flow cases, for which the first three base tensors form a linearly independent basis and only the first two invariants are nonzero [7], and the set of base tensors and invariants reads

$$\begin{aligned} T_{ij}^{(1)} &= \hat{S}_{ij},\ T_{ij}^{(2)} = \hat{S}_{ik}\hat{\Omega}_{kj} - \hat{\Omega}_{ik}\hat{S}_{kj}, \\ T_{ij}^{(3)} &= \hat{S}_{ik}\hat{S}_{kj} - \tfrac{1}{3}\delta_{ij}\hat{S}_{mn}\hat{S}_{nm} \\ I_1 &= \hat{S}_{mn}\hat{S}_{nm},\ I_2 = \hat{\Omega}_{mn}\hat{\Omega}_{nm}. \end{aligned} \tag{7}$$

Deriving EARSM models then consists in finding a set of scalar functions G_λ, which are classically obtained from physical considerations [7,8]. In the following, we present a deterministic symbolic identification method to build EARSM models from high-fidelity data sets, cast in the form of equations (3), (5) and (6).

In the numerical experiments of section 5, we also consider a general purpose EARSM, recently proposed in [32] and referred-to as BSL-EARSM, for comparison with customized data-driven models. Its formulation is briefly recalled in Appendix B.

2.3. CFD solver

The CFD-SpaRTA methodology requires the implementation of the general formulation of the augmented EARSM in a CFD solver. In the present work, the open-source finite-volume code OpenFOAM [33] has been used. The RANS equation system (equations (1) and (2)) are solved using the SIMPLE algorithm [34]; the convective terms are discretized using linear upwinding and viscous terms with 2nd order central differencing. The solution is advanced to the steady state using a Gauss-Seidel smoother.

3. Flow configurations and data sets

The aim of the present work being the development of data-driven corrections to turbulence models, all numerical experiments are carried out for a family of flow configurations that challenge standard RANS models, and specifically LEVM. In particular, we have considered the same cases reported in [1], representing flows in 2D channels with variable cross sections. High fidelity data from DNS or well-resolved LES are available for model training and validation. To ensure that discretization errors remain low with respect to turbulence modeling errors, the RANS simulation use sufficiently fine meshes.

More details on the flow configurations and data sets used in the present numerical experiments are given below.

3.1. Periodic hill flow (PH_{10595})

The so-called periodic hill flow consists of a flow through a channel constrained by periodic restrictions (hills) of height h. For a channel segment comprised between two adjacent hills, the flow separates on the lee-side of the first hill and reattaches between the hills. The test case has been widely investigated in the literature, both experimentally and numerically. The high-fidelity LES data used in the present work are from [35] for $Re = 10595$ (PH_{10595}), where Re is a Reynolds number based on the bulk velocity in the restricted section and the hill height. Our RANS simulations use a computational grid consisting of 120 x 130 cells. Cyclic boundary conditions are used at the inlet and outlet and a forcing term is applied to maintain a constant flow rate through the channel.

3.2. Converging-diverging channel at Re=12600 (CD_{12600})

This configuration corresponds to a 2D channel of half-height H with an asymmetric bump of height $h \approx 2/3H$ located on the bottom wall. The Reynolds number (based on the channel half-height and inlet conditions) is $Re = 12600$. A small separation occurs on the lee-side of the bump. For this test case, high-fidelity DNS data from [36] are available. The RANS simulations are based on a mesh of 140 x 100 cells. A velocity profile obtained from a companion channel-flow simulation is imposed at the inlet of the computational domain.

3.3. Curved backward-facing step ($CBFS_{13700}$)

The case consists in a 2D flow over a gently-curved backward-facing step of height h, producing a separation bubble. The upstream channel height is $8.52h$ and the Reynolds number, based on the inlet velocity and step height, is 13700. High-fidelity LES data from [37] are used for training. For the RANS simulations, the mesh consists of 140 x 150 cells. Slip conditions are used at the upper boundary, and a velocity profile obtained from a fully-developed boundary layer simulation is set at the domain inlet.

4. Symbolic identification methodology

In this section, we introduce the general form of candidate models, constructed from a library of functions. Then we recall the frozen-training SpaRTA algorithm of [1]. Finally, we describe in detail the steps of the new CFD-driven symbolic identification algorithm.

4.1. Symbolic identification of EARSM

Following the approach proposed in [1], the scalar functions G_λ are approximated as linear combinations of candidate functions of I_1 and I_2, consisting of monomials up to order 6:

$$\begin{aligned}\mathbf{B} = [&c, I_1, I_2, I_1^2, I_1 I_2, I_2^2, I_1^3, I_1^2 I_2, I_1 I_2^2, I_2^3, I_1^4, I_1^3 I_2, I_1^2 I_2^2, I_1 I_2^3,\\ &I_2^4, I_1^5, I_1^4 I_2, I_1^3 I_2^2, I_1^2 I_2^3, I_1 I_2^4, I_2^5, I_1^6, I_1^5 I_2, I_1^4 I_2^2, I_1^3 I_2^3, I_1^2 I_2^4, I_1 I_2^5, I_2^6]^T.\end{aligned} \tag{8}$$

Given this library, the full expression of eq. (6) becomes:

$$\begin{aligned} b_{ij}^{\Delta}(S_{ij}, \Omega_{ij}) \quad &= \theta_0^{b_{ij}^{\Delta}} T_{ij}^{(1)} + \theta_1^{b_{ij}^{\Delta}} I_1 T_{ij}^{(1)} + \theta_2^{b_{ij}^{\Delta}} I_2 T_{ij}^{(1)} + \dots \\ &+ \theta_{26}^{b_{ij}^{\Delta}} I_1 I_2^5 T_{ij}^{(1)} + \theta_{27}^{b_{ij}^{\Delta}} I_2^6 T_{ij}^{(1)} \\ &+ \theta_{28}^{b_{ij}^{\Delta}} T_{ij}^{(2)} + \theta_{29}^{b_{ij}^{\Delta}} I_1 T_{ij}^{(2)} + \theta_{30}^{b_{ij}^{\Delta}} I_2 T_{ij}^{(2)} + \dots \\ &+ \theta_{54}^{b_{ij}^{\Delta}} I_1 I_2^5 T_{ij}^{(2)} + \theta_{55}^{b_{ij}^{\Delta}} I_2^6 T_{ij}^{(2)} \\ &+ \theta_{56}^{b_{ij}^{\Delta}} T_{ij}^{(3)} + \theta_{57}^{b_{ij}^{\Delta}} I_1 T_{ij}^{(3)} + \theta_{58}^{b_{ij}^{\Delta}} I_2 T_{ij}^{(3)} + \dots \\ &+ \theta_{82}^{b_{ij}^{\Delta}} I_1 I_2^5 T_{ij}^{(3)} + \theta_{83}^{b_{ij}^{\Delta}} I_2^6 T_{ij}^{(3)}, \end{aligned} \tag{9}$$

where the vector of coefficients:

$$\Theta^{b_{ij}^{\Delta}} = \left[\theta_0^{b_{ij}^{\Delta}}, \theta_1^{b_{ij}^{\Delta}}, \theta_2^{b_{ij}^{\Delta}}, \dots, \theta_{83}^{b_{ij}^{\Delta}}\right]^T \quad \in \mathbb{R}^{84} \tag{10}$$

has to be determined.

A major improvement proposed in [1] consists in the introduction of a corrective term in the turbulent transport equations, in addition to the use of the generalized eddy viscosity formulation. Specifically, a residual term R is added to the model transport equations and the augmented k-ω-SST model becomes:

$$\frac{\partial k}{\partial t}+\overline{u}_j\frac{\partial k}{\partial x_j}=P_k+R-\beta^* k\omega+\frac{\partial}{\partial x_j}[(\nu+\sigma_k\nu_t)\frac{\partial k}{\partial x_j}] \tag{11}$$

$$\frac{\partial \omega}{\partial t}+\overline{u}_j\frac{\partial \omega}{\partial x_j}=\frac{\alpha}{\nu_t}(P_k+R)-\beta\omega^2+\frac{\partial}{\partial x_j}[(\nu+\sigma_\omega\nu_t)\frac{\partial \omega}{\partial x_j}]+CD_{k\omega}, \tag{12}$$

with $P_k=\min(-2k(b_{ij}^{bl}+b_{ij}^{\Delta})\partial_j U_i, 10\beta^*\omega k)$.

The residual R can be positive or negative and it represents a correction of the model production term. Introducing a suitable non-dimensional correction tensor b_{ij}^R, the model for R writes:

$$R=2kb_{ij}^R\frac{\partial \overline{u}_j}{\partial x_i}, \tag{13}$$

where b_{ij}^R is also decomposed according to the Cayley-Hamilton theorem:

$$\begin{aligned} b_{ij}^R(S_{ij},\Omega_{ij}) \quad &= \theta_0^R T_{ij}^{(1)}+\theta_1^R I_1 T_{ij}^{(1)}+\theta_2^R I_2 T_{ij}^{(1)}+\dots \\ &+\theta_{26}^R I_1 I_2^5 T_{ij}^{(1)}+\theta_{27}^R I_2^6 T_{ij}^{(1)} \\ &+\theta_{28}^R T_{ij}^{(2)}+\theta_{29}^R I_1 T_{ij}^{(2)}+\theta_{30}^R I_2 T_{ij}^{(2)}+\dots \\ &+\theta_{54}^R I_1 I_2^5 T_{ij}^{(2)}+\theta_{55}^R I_2^6 T_{ij}^{(2)} \\ &+\theta_{56}^R T_{ij}^{(3)}+\theta_{57}^R I_1 T_{ij}^{(3)}+\theta_{58}^R I_2 T_{ij}^{(3)}+\dots \\ &+\theta_{82}^R I_1 I_2^5 T_{ij}^{(3)}+\theta_{83}^R I_2^6 T_{ij}^{(3)}, \end{aligned} \tag{14}$$

The vector of coefficients:

$$\Theta^R=\left[\theta_0^R,\theta_1^R,\theta_2^R,\dots,\theta_{83}^R\right]^T \quad \in\mathbb{R}^{84} \tag{15}$$

needs to be determined alongside with $\Theta^{b_{ij}^{\Delta}}$.

Once the general library of candidate models has been chosen, the difference between the frozen-training and the CFD-driven approaches relies in the method used to determine $\Theta^{b_{ij}^{\Delta}}$ and Θ^R.

4.2. Frozen-training SpaRTA algorithm

In this approach, high fidelity data for the velocity field, $\overline{u}_i^*$, turbulent kinetic energy k^* and Reynolds-stresses τ_{ij}^* are used to train the model. Such data are pre-processed prior to use for model learning.

The high-fidelity anisotropy tensor b_{ij}^* is obtained from equation (3). The next step of the method is the determination of the high fidelity values for b_{ij}^{Δ} and R. Knowing b_{ij}^*, $b_{ij}^{\Delta,*}=b_{ij}^*+\frac{\nu_t^*}{k^*}S_{ij}^*$ is directly derived if a high-fidelity ansatz of the turbulent viscosity ν_t^* is provided. The latter can be computed using the relation (A.3), setting $k=k^*$. However, a high fidelity value of ω is also needed. For this purpose, a so-called k-corrective-frozen-RANS algorithm is used, whereby a high-fidelity estimate of the residual R^* is computed alongside ω^*. This process consists in injecting $\overline{u_i}^*$, k^* and b_{ij}^* in equations (11) and (12) and iteratively computing ω^* by solving equation (12). At each iteration, R^* is computed as the residual of equation (11) and fed back into equation (12). On completion of the preceding manipulations, high fidelity fields for $b_{ij}^{\Delta,*}$ and R^* are available.

High-fidelity counterparts of the terms in the library $\mathbf{B}^*$ and of the basis tensors $T_{ij}^{(\lambda,*)}$ are computed from the high-fidelity velocity field by using relations (7), (8); $\mathbf{B}^*$ and $T_{ij}^{(\lambda,*)}$ are then used to compute the b_{ij}^{Δ} and R fields corresponding to candidate models defined by specifying the sets of coefficients $\Theta^{b_{ij}^{\Delta}}$ and Θ^R. For each field Δ (Δ indicating either b_{ij}^{Δ} or R) the following optimization problem can be constructed:

$$\Theta=\underset{\hat{\Theta}}{\operatorname{argmin}}\|\Delta^*-\Delta(\hat{\Theta})\|_2^2+\text{regularization terms}, \tag{16}$$

where $\|\bullet\|_2$ stands for the l_2 norm of the error, $\hat{\Theta}$ is the vector $\Theta^{b_{ij}^{\Delta}}$ (respectively Θ^R) if Δ represents b_{ij}^{Δ} (respectively R). Δ^* is the high fidelity value of the field and $\Delta(\hat{\Theta})$ is the field reconstructed using the model corresponding to $\hat{\Theta}$.

The choice of the regularization terms is crucial to efficiently select a small subset of relevant terms among those constituting the highly dimensional symbolic library. The goal of the whole methodology being the construction of augmented RANS models optimized for a specific class of flows, the learnt model should be general enough to be effective for the simulation of flows belonging to the same class (e.g. flows with separations). Without the regularization terms, the problem formulated by equation (16) reduces to a standard mean-squares optimization problem whose resolution leads to dense coefficient vectors $\Theta^{b_{ij}^{\Delta}}$ and Θ^R, possibly overfitting the data.

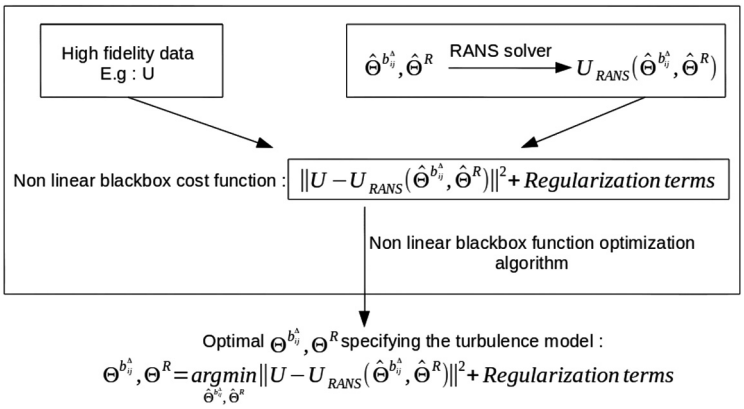


Fig. 1. Technical flow diagram of the CFD-driven SpaRTA method. U represents any quantity of interest considered as the target for model training.

In [1], this effect is mitigated by using an elastic net regularization, known to be sparsity promoting [38,39]:

$$\text{regularization terms} = \lambda\rho\|\hat{\Theta}\|_1 + 0.5\lambda(1-\rho)\|\hat{\Theta}\|_2^2, \tag{17}$$

where $\rho \in [0,1]$ is the blending parameter between the l_1- and l_2- norm regularization and λ is the regularization weight. The blending parameter ρ relaxes the l_1- regularization, that tends to select only one function in sets of correlated functions and therefore degrade the predictive performance of the inferred models. This relaxation allows to find sparse models with good predictive performance. The regularization weight λ also controls the sparsity of the inferred models (independently of ρ). As the optimal value of (ρ,λ) is not known a priori, the optimization problem given in equation (16) must be solved for a grid of pairs $(\rho_{1\leq i\leq N}, \lambda_{1\leq j\leq M})$. In Schmelzer et al. [1], a set of optimizations is carried out a grid of 9×100, leading to the discovery of 900 concurrent models, some of which have the same abstract form (i.e. involve the same functional terms). Once the abstract model forms have been identified, an inference step based on ridge regression is used to calibrate the model coefficients. The computational cost of model discovery and calibration is of the order of minutes, since no additional CFD solve is required at this stage. Finally, best-performing models are selected from the wide set of learned models through cross validation [40]. In order to not overcharge the role of the training data from k-corrective-frozen-RANS, the validation task is performed *a posteriori* by implementing candidate models in a CFD solver. This allows validation against quantities other than those used in the training process, namely velocity data. Specifically, data for cases PH_{10595}, CD_{12600} and CBFS_{13700} are used, leading to approximately 200 CFD calculations.

The CFD solver used in [1] is OpenFOAM [33], in which the general formulation of the augmented EARSM has been implemented as a new class of models. The coefficients are passed by the optimizer to the OpenFoam input file RASproperties using a python interface.

An implementation of a general EARSM model in OpenFOAM is available here: https://github.com/shmlzr/general_earsm.git.

4.3. CFD-driven symbolic identification

The technical flow diagram of the CFD-driven algorithm that we have developed in the present work to determine $\Theta^{b_{ij}^{\Delta}}$ and Θ^R is sketched in Fig. 1.

The first step of the method consists in gathering high fidelity data for a flow or a class of flows for which we want to improve the accuracy of the baseline model. It should be emphasized that, while the frozen-training approach requires data for second-order quantities like k and τ_{ij}, observations available for any quantity of interest (velocity or pressure measurements, global aerodynamic coefficients, etc.) can be used as a target for training. In the flowchart of Fig. 1, we assumed that high fidelity data are available for a quantity of interest U: this could be, for instance, the streamwise velocity component $\overline{u}_1$ but the same procedure applies for any other data set. Typically, U is a vector of high-fidelity samples extracted at selected locations in the flow field and concatenated into a single vector.

In the second step, the general form of the augmented turbulence model is implemented in the CFD-solver, as described in Section 4.2. The latter is used to evaluate the quantity U using a candidate augmented turbulence model specified by any instance of the vectors $\hat{\Theta}^{b_{ij}^{\Delta}}$ and $\hat{\Theta}^R$. The output RANS data for U are indicated as $U_{RANS}(\hat{\Theta}^{b_{ij}^{\Delta}}, \hat{\Theta}^R)$. Then, the fitness of candidate models is evaluated in terms of the mean squared error $\|U^* - U_{RANS}(\hat{\Theta}^{b_{ij}^{\Delta}}, \hat{\Theta}^R)\|_2^2$ between the high fidelity data U^* and the RANS model prediction $U_{RANS}(\hat{\Theta}^{b_{ij}^{\Delta}}, \hat{\Theta}^R)$. The model can then been constructed by solving the following optimization problem:

$$(\Theta^{b_{ij}^{\Delta}}, \Theta^R) = \underset{(\hat{\Theta}^{b_{ij}^{\Delta}}, \hat{\Theta}^R)}{\text{argmin}} \|U - U_{RANS}(\hat{\Theta}^{b_{ij}^{\Delta}}, \hat{\Theta}^R)\|_2^2 + \text{regularization terms}, \tag{18}$$

where regularization terms are added to the least squares formulation to prevent overfitting of the high fidelity data. The choice of the regularization is discussed in the following.

The preceding cost function is a non-linear black-box function involving a RANS simulation for any choice of $\hat{\Theta}^{b^\Delta_{ij}}, \hat{\Theta}^R$, and is therefore costly in terms of CPU time. In the following, problem (18) is optimized using the Constrained Optimization using Response Surface (CORS) algorithm [30], which uses the response surface methodology to approximate the expensive function using a limited number of function evaluations. The optimization is subsequently performed on this response surface which is further enriched at each iteration with a new function evaluation at the provisional optimum location (see Appendix C for a short description of the algorithm). Here we use a version of the CORS algorithm adapted from the package available in the *blackbox* python module (see Ref. [41]).

A first modification is introduced to account for diverging or poorly converging CFD simulations due to non realizable/non robust candidate models. If convergence is not achieved according to the prescribed criteria, the candidate model is eliminated and the coefficients are resampled by slightly decreasing their magnitude. The procedure is repeated until convergence is obtained. Then, the initial response surface is built using n function evaluations, while additional m evaluations are used to converge the solution. The parameters m and n are set independently and typically $n < m$ and $n + m > 2D$ to ensure a good exploration of the parameter space and a satisfactory convergence to the optimum, with D the dimensionality of the search space. Note that in the original CORS algorithm $n = m = N/2$, with $N > 2D$.

Given the difficulty of constructing accurate response surfaces in highly dimensional parameter spaces (here, $D = 168$), a preliminary local sensitivity analysis is conducted to identify the most influential parameters and reduce the search space. Afterwards, search ranges for model parameters are assigned based on Reynolds-stress realizability considerations. Finally, the regularization terms in the fitness function definition (18) are specified. These three steps are described into more details in the following.

4.3.1. Local sensitivity analysis

In the CORS algorithm here implemented, the total number of evaluations $N = n + m$ must satisfy the constraint $N > 2D$. As the general formulation of the augmented model involves 168 coefficients, we have $N > 336$. In addition to the large numerical cost of the training procedure, for highly dimensional spaces, the response surface may not provide an accurate enough approximation of the true cost function.

To reduce the size of the parameter space, we have carried out a local sensitivity analysis around the baseline model formulation (corresponding to $\theta^{b^\Delta} = 0$ and $\theta^R = 0$) to determine the most influential coefficients. For that purpose, each coefficient is varied by a small perturbation ($O(10^{-3})$), and the corresponding CFD solution is determined. The derivatives of the cost function are then evaluated by one-sided finite differences. More precisely, we investigate the sensitivity of the mean squared error on prescribed reference data to parameters $\Theta^{b^\Delta_{ij}}$ and Θ^R of the general EARSM.

The results of the sensitivity study are reported in the following for the three flow configurations PH_{10595}, CD_{12600} and $CBFS_{13700}$ of Section 3. For the three flows, the cost function is based on high fidelity velocity and Reynolds stress profiles. The profiles are selected at the following streamwise locations: for PH_{10595} at $x/h = (0.05, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0)$; for CD_{12600} at $x/H = (5.5, 6.5, 7.5, 8.5, 9.5, 10.5, 11.5, 12)$; and for $CBFS_{13700}$ $x/h = (0.1, 1.5, 3.0, 4.5, 6.0, 7.5)$.

The sensitivity derivatives of the cost function with respect to the parameters of the b^Δ model are reported in Fig. 2, where the abscissas correspond to parameter subscripts. Parameters with subscripts $0 \div 27$, $28 \div 55$ and $56 \div 83$ appear in the function coefficients of $T^{(1)}_{ij}$, $T^{(2)}_{ij}$ and $T^{(3)}_{ij}$, respectively. Several comments are in order. First, the parameters associated with $T^{(1)}_{ij}$ (corresponding to a correction of the linear eddy viscosity coefficient) are overall more influential than the parameters involved in the quadratic tensorial terms $T^{(2)}_{ij}$ and $T^{(3)}_{ij}$, which govern higher-order corrections of the constitutive law. Second, the parameter sensitivity decays rapidly with the degree of the monomial functions, and terms up to the second degree are at least one order of magnitude larger than the coefficients of the higher order terms. This has led us to retain only monomials up to the second order in the library of candidate functions (equation (8)). Similar results are found for the parameters of b^R. On the basis of this analysis, the dimensionality of the search space is reduced to 36 instead of the initial value 168.

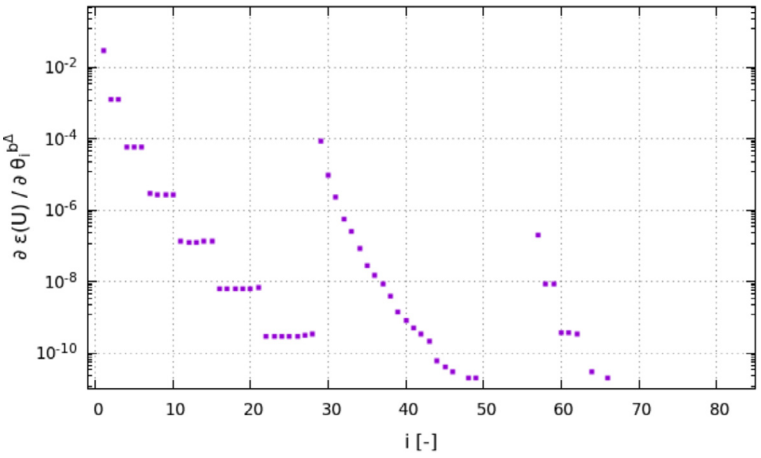
4.3.2. Selection of parameter ranges

The initial range of each model coefficient must be specified prior to the CORS optimization. For each $\Theta_{1 \le i \le 36}$ of $\Theta^{b^\Delta_{ij}}$ and Θ^R, the search space is set as:

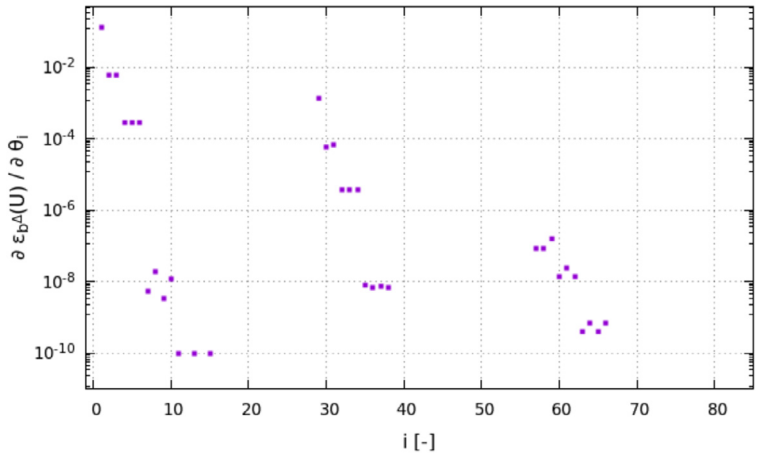
$$a_i \le \Theta_i \le b_i \tag{19}$$

where $a_i \in \mathbb{R}$ and $b_i \in \mathbb{R}$ are respectively the lower and upper bound of the range of variation of candidate model coefficients. In our setup of the CFD-driven method, the coefficients ranges are set with the aim to ensure the realizability constraint [42].

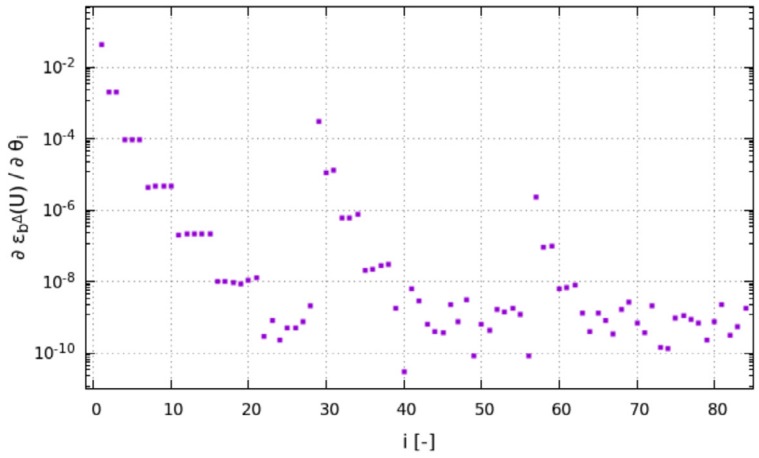
As stressed in Section 4.3 and in Appendix C candidate models are evaluated by running RANS simulations at each iteration of the optimization loop. When a candidate model is found to be non robust, preventing convergence, it is eliminated and resampled until a robust model is obtained. This procedure introduces an additional computational cost that depends on the number of non robust models to be evaluated. This number increases with the dimension of the search space D.



(a) PH_{10595}



(b) CD_{12600}



(c) CBFS_{13700}

Fig. 2. Sensitivity derivatives of the cost function (based on the velocity components) to the parameters θ^{b^Δ} for the three test cases.

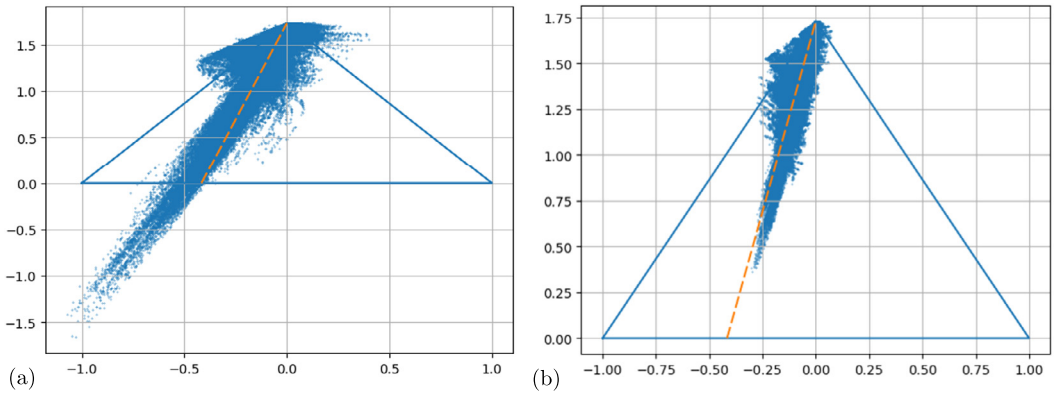


Fig. 3. Modeled anisotropy tensor at each mesh cell placed in the barycentric map for 100 random vectors $\Theta^{b_{ij}^{\Delta}}$ sampled using uniform Latin Hypercube sampling. (a) sampling in range $[-7,7]^{18}$. (b) Sampling in range $[-2,2]^{18}$.

Hence, the *a priori* specification of parameter ranges results from the following trade-off. On one hand, the parameter range must be large enough to obtain substantial modification of the baseline model. On the other, it must be as small as possible to reduce the number of non robust candidate models and avoid the overcost associated with testing / elimination and resampling.

For that purpose, a realizability study is performed using the Reynolds stress anisotropy barycentric map representation [29,43,44].

A priori realizability study. The modeled anisotropy tensor is initially computed from high fidelity data with each parameter $\Theta^{b_{ij}^{\Delta}}$ varying in the range $[-7,7]$. The tensor eigenvalues are noted λ_1, λ_2 and λ_3 (with $\lambda_1 \geq \lambda_2 \geq \lambda_3$). Following [29], we construct an equilateral triangle defined by the three corners $\mathbf{x}_{1C}$, $\mathbf{x}_{2C}$ and $\mathbf{x}_{3C}$, where $\mathbf{x} = (x, y)$ is a position vector. The three corners the triangle represent the three limiting physical states of turbulence. We then construct the following linear mapping between the anisotropy eigenvalues and the coordinates x:

$$\mathbf{x} = \mathbf{x}_{1C}(\lambda_1 - \lambda_2) + \mathbf{x}_{2C}(2\lambda_2 - 2\lambda_3) + \mathbf{x}_{3C}(3\lambda_3 + 1). \tag{20}$$

As any realizable state of turbulence is a convex combination of the three limiting states of turbulence, $\mathbf{x}$ must lie within the triangle if λ_1, λ_2 and λ_3 correspond to a realizable anisotropy tensor. For a given set of coefficients $\Theta^{b_{ij}^{\Delta}}$, the eigenvalues of b_{ij} can be computed for each mesh point and the corresponding $\mathbf{x}$ can be placed in the barycentric map in order to verify if the resulting anisotropy tensor field corresponds to a physically realizable flow.

A total of 50000 vectors $\Theta^{b_{ij}^{\Delta}}$ are generated using uniform Latin Hypercube sampling. For each sample, the corresponding anisotropy tensor eigenvalues are plotted in the barycentric map coordinate system for each mesh point. Sets of parameters leading to a violation of the realizability condition are then discarded. For the remaining sets of parameters, we compute the average value, the standard deviation and the minimum and maximum values taken by each component of $\Theta_l^{b_{ij}^{\Delta}}$. Realizable values of the coefficients are equal to 0 in the mean, with a variance of approximately 4. Therefore, ranges $[-2,2]$ are finally adopted for the parameters $\Theta^{b_{ij}^{\Delta}}$ and the same search range is assumed to be appropriate for the Θ_{ij}^{R}.

In Fig. 3a, results for the PH_10595 case are reported. For the sake of clarity, the modeled anisotropy tensor at each mesh cell has been placed in the barycentric map using only 100 random vectors $\Theta^{b_{ij}^{\Delta}}$ sampled using uniform Latin Hypercube sampling in range $[-7,7]^{18}$. The figure shows that the anisotropy tensor is not realizable for a large fraction of mesh points. In Fig. 3b, the sampling has been performed in range $[-2,2]^{18}$. The fraction of mesh points possibly corresponding to non realizable anisotropy tensors is drastically reduced. Therefore, an initial parameter range $[-2,2]$ ensures that most candidate models lead to realizable anisotropy tensor fields in the entire computational domain. Generally, this choice yields more robust models and limits the risk non-converging CFD simulations. The remaining non-robust models are excluded from the CORS optimization as mentioned in Section 4.3 (see also Appendix C).

CFD-based realizability study. As an alternative to using high-fidelity full field data, a CFD-based realizability study can be performed prior to the optimization step. Here, a set of 300 random vectors $\Theta^{b_{ij}^{\Delta}}$ are sampled using the uniform Latin Hypercube, with $\Theta^{b_{ij}^{\Delta}} \in [-7,7]^{18}$. Each $\Theta^{b_{ij}^{\Delta}}$ vector corresponds to a candidate model, for which an OpenFOAM simulation of the PH_{10595} case is carried out. With the chosen range, approximately 94% of the sampled models prevents solution convergence. For the remaining simulations, we extract the anisotropy tensor b_{ij} at each mesh point and we examine the location of its eigenvalues with respect to the barycentric map. We observe that 84% of the robust models yield realizable states at all mesh points. For these models, we compute the mean, the standard deviation and the minimum and maximum values for each component of $\Theta_l^{b_{ij}^{\Delta}}$. These quantities are displayed in Fig. 4, which shows that the coefficients yielding

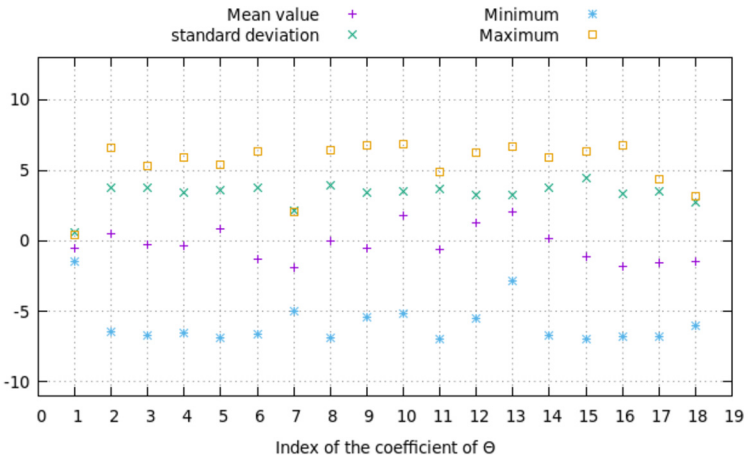


Fig. 4. Mean value, variance, minimum and maximum value of each coefficient $\Theta^{b_{ij}^{\Delta}}_{1\leq l\leq 18}$ among vector samples corresponding to realizable models.

realizable states are approximately equal to 0 in the mean and have a variance of approximately 4. For the study conducted here, only 300 random samples have been considered. Even though this number is not sufficient for the statistical quantities to achieve convergence, the results are in good accordance with the previously discussed *a priori* realizability study, since each sample generates a Reynolds stress field and realizability can be checked at each mesh point.

We then performed the same study with 300 random vectors $\Theta^{b_{ij}^{\Delta}}$ sampled in the range $[-2, 2]^{18}$. More than 50 % of the samples led to convergence of the simulation and approximately 70% of these models yielded to fully realizable Reynolds stress fields. We then optimize the model in the narrower parameter range $[-2, 2]$, and check *a posteriori* that the algorithm selects a realizable optimum. The computations can be run in parallel and some of them can be re-used to initialize the CORS algorithm at the subsequent optimization step.

We point out that the CFD-based realizability study can be skipped, since unrealizable models tend to be numerically unstable and in any case less accurate, and tend to be naturally discarded by the optimizer. In this case, arbitrary search ranges must be assigned by the user to the optimizer, and the realizability of the learned models is only assessed *a posteriori*. Based on the present *a priori* and *a posteriori* studies, we recommend using ranges of $[-2, 2]$ or less.

4.3.3. Regularization terms

The regularization terms of the optimization problem (18) have to be specified. Unlike the frozen-training SpaRTA, the CFD-driven training involves a nonlinear dependency of the cost function on the model parameters, due to the nonlinearity of the RANS equations. Sparsity-promoting regularization of nonlinear parameters would require an approximate computation of the cost function derivatives with respect to the parameters or function linearization [45,46]. This can be efficiently done for CFD using an adjoint solver, which is not always promptly available. For this reason, we choose to simply add a LASSO regularization, and we examine its effect on the complexity and generalizability of the resulting models. The optimization problem of equation (18) is reformulated as:

$$\Theta = \underset{\hat{\Theta}}{\operatorname{argmin}} \|U - U_{RANS}(\hat{\Theta})\|_2^2 + \lambda \|\hat{\Theta}\|_1, \tag{21}$$

where λ is the regularization weight Since its optimal value is not known a priori, the optimization problem given in equation (21) must be solved for N values λ_i with $1 \leq i \leq N$. In particular in this work we have selected $\lambda = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$. This results in six alternative models for each test case of section 3. Cross-validation [40] is then used to select the best-performing model across all validation data sets. The performance of each candidate is arbitrarily evaluated as the mean squared error with respect to the high fidelity velocity data. Other choices are possible.

5. Numerical results

5.1. Model discovery

In the following, the CFD-driven SpaRTA is applied to the flow configurations of Section 3. Two series of numerical experiments are carried out, using different training data.

In a first series of experiments, high-fidelity data for the Reynolds stresses are used to train the model, as in the frozen-training SpaRTA algorithm. However, in this case no additional data for the velocity field or the transported turbulent quantities is needed and the k-corrective frozen-RANS procedure is not applied. More precisely, the training data are vertical profiles of τ_{11}, τ_{22} and τ_{12}. Profiles at $x/h = (0.05, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0)$ are used for

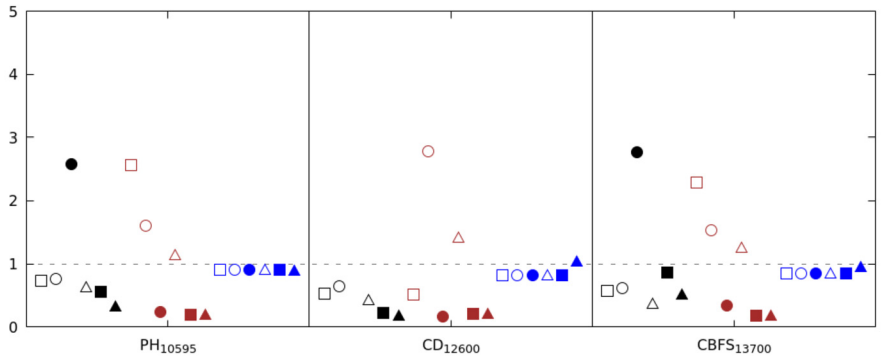


Fig. 5. Cross-validation errors based on velocity profiles ($\overline{u_1}$, $\overline{u_2}$) for models trained on Reynolds-stress data. The errors are normalized with the mean-squared error of the baseline k-ω SST model. Black: models trained on PH_{10595}. Red: models trained on CD_{12600}. Blue: models trained on $CBFS_{13700}$. □: $\lambda = 10^{-1}$. ○: $\lambda = 10^{-2}$. ●: $\lambda = 10^{-3}$. △: $\lambda = 10^{-4}$. ■: $\lambda = 10^{-5}$. ▲: $\lambda = 10^{-6}$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

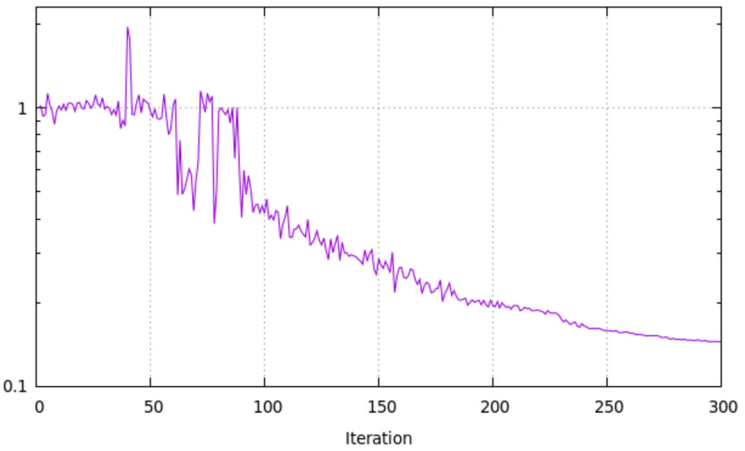


Fig. 6. Convergence of the optimization for Model 1: value of the cost function with respect to the iteration.

the PH_{10595} case; for the CD_{12600} we use profiles at $x/H = (5.5, 6.5, 7.5, 8.5, 9.5, 10.5, 11.5, 12)$; finally, the profiles at $x/h = (0.1, 1.5, 3.0, 4.5, 6.0, 7.5)$ are considered for case $CBFS_{13700}$. The amount of data used for the training is very sparse compared to the number of degrees of freedom used to solve the CFD problem: they represent 8.3%, 5.7% and 4.28% of the total number of mesh points, used respectively for PH_{10595}, CD_{12600} and $CBFS_{13700}$. Six values of the regularization weight λ are considered, as discussed previously. For each of the six values of λ, the CORS algorithm has been run using $N = 300$ function evaluations, and setting $n = 39$ and $m = 261$. The model training step leads to the discovery of 3×6 alternative models that are submitted to cross-validation: models trained for a flow configuration are applied to the two other cases. A total of 5400 RANS simulations (using 12 cores each on a standard workstation) have therefore been performed for the model discovery and the cross-validation. The overall cost of the learning procedure corresponds to approximately 16500 core-hours. This is obviously much more expensive than the frozen-training procedure but it remains affordable without use of intensive computational facilities. Mean squared errors on velocities, normalized by the mean-squared error of the baseline $k-\omega$ SST model, are reported in Fig. 5. Most models show an improvement over the baseline. Models trained on $CBFS_{13700}$ produce errors of the same order than the baseline model. Models trained on CD_{12600} with low values of λ produce the smallest errors averaged on the validation set.

The best-performing model, identified as the one having the lowest average error across the validation set, is trained on CD_{12600} with $\lambda = 10^{-5}$, and is referred-to as Model 1 in the following. The convergence history of the cost function residual for this model is shown in Fig. 6. We see that a satisfactory convergence of the optimization is reached. The model exhibits only a few terms that are bigger than 10^{-1} in magnitude. The other terms, are mostly $O(10^{-2})$ or lower, and are neglected in the selected formulation of the model.

The resulting mathematical expressions for the two correction tensors involve 4 out of the 36 candidate functions left after parameter space reduction:

$$b_{ij}^{\Delta} = -1.47 \times 10^{-1} I_1^2 T_{ij}^{(1)} - 2.6791 \times 10^{-1} T_{ij}^{(2)} \tag{22}$$

$$b_{ij}^{R} = 4.6018 \times 10^{-1} T_{ij}^{(1)} - 1.6779 \times 10^{-1} T_{ij}^{(3)} \tag{23}$$

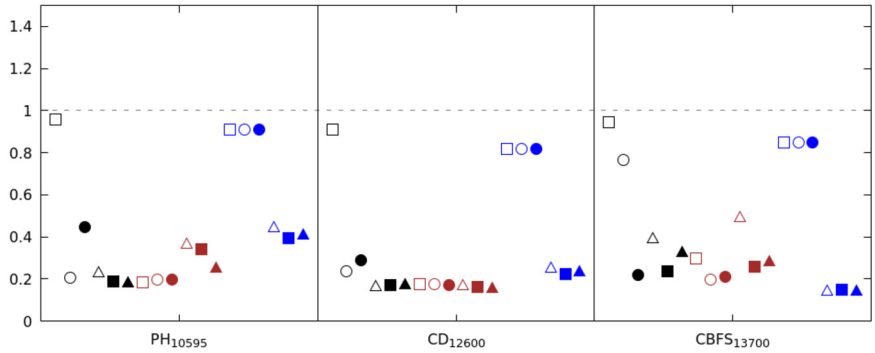


Fig. 7. Cross-validation errors based on velocity profiles ($\overline{u_1}$, $\overline{u_2}$) for models trained on velocity and skin friction data. The errors are normalized with the mean-squared error of the baseline k-ω SST model. Black: models trained on PH_{10595}. Red: models trained on CD_{12600}. Blue: models trained on $CBFS_{13700}$. □: $\lambda = 10^{-1}$. ○: $\lambda = 10^{-2}$. ●: $\lambda = 10^{-3}$. △: $\lambda = 10^{-4}$. ■: $\lambda = 10^{-5}$. ▲: $\lambda = 10^{-6}$.

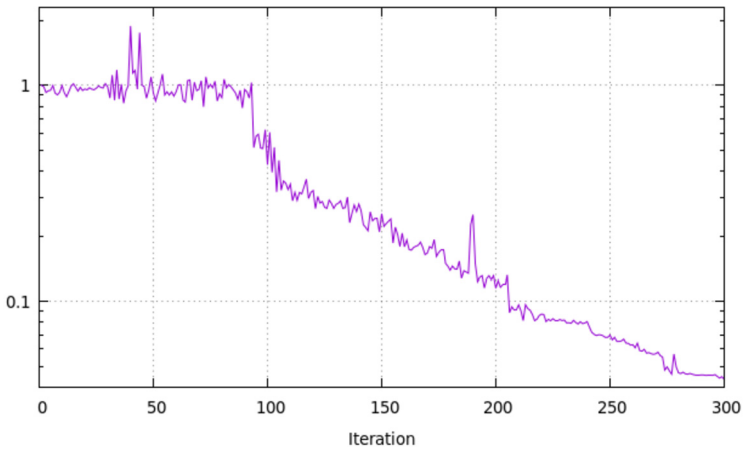


Fig. 8. Convergence of the optimization for Model 2: value of the cost function with respect to the iteration.

In the second series of experiments, the method is trained against high fidelity velocity profiles and skin friction data to test its ability to discover models optimized for any target quantity of interest and to assess the sensitivity of the symbolic identification procedure to the kind of training data. Velocity profiles ($\overline{u_1}$ and $\overline{u_2}$) are extracted at the same locations of the Reynolds stress profiles, whereas the skin friction data are extracted along the bottom walls in all cases. In this setting the data represent 9.1%, 6.7% and 4.95% of the total number of mesh points for PH_{10595}, CD_{12600} and $CBFS_{13700}$, respectively. Cross validation errors for the 18 discovered models are reported in Fig. 7. The target variable (the longitudinal velocity) being the same as the training one, the validation errors are lower.

The best-performing model is the one learned on CD_{12600} with $\lambda = 10^{-2}$ and it is hereafter named Model 2. The convergence history of the cost function residual for training of Model 2 is shown in Fig. 8. Again, a satisfactory convergence of the optimization is reached.

As for Model 1, Model 2 is sparse and the retained mathematical expression with coefficients $O(10^{-1})$ or higher is:

$$b_{ij}^{\Delta} = -2.8356 \times 10^{-1} T_{ij}^{(1)} - 1.4738 \times 10^{-1} I_2^2 T_{ij}^{(2)} \tag{24}$$

$$b_{ij}^{R} = \quad (1.0375 \times 10^{-1} I_2 - 2.8833 \times 10^{-1} I_1 I_2) T_{ij}^{(3)} \tag{25}$$

For both models we checked a posteriori that the realizability conditions are satisfied throughout the flow domain for all cases.

In the following, a detailed assessment of the two learned models against the high-fidelity data is carried out for a number of flow quantities, and the results are compared to those of the $k-\omega$ SST model. For comparison, we also include the results of the BSL-EARSM model of Menter et al. [32] and those of a frozen-training SpaRTA model reported in Ref. [1], given in the following:

$$b_{ij}^{\Delta} = 0 \tag{26}$$

$$b_{ij}^{R} = 3.9 \times 10^{-1} T_{ij}^{(1)} \tag{27}$$

Table 1
Mean-squared errors of predicted velocity, turbulent kinetic energy and Reynolds shear-stress profiles and skin friction distribution with respect to high-fidelity data for various models. The errors are normalized by the mean-squared error of the baseline k-ω SST model.

Velocity $((\overline{u}_1, \overline{u}_2)$			
	PH_{10595}	CD_{12600}	$CBFS_{13700}$
Model 1	0.1901	0.1993	0.1687
Model 2	0.1932	0.1767	0.2012
frozen-training	0.2067	0.2067	0.1711
BSL-EARSM	0.3207	0.5999	0.7762
Turbulent kinetic energy k			
	PH_{10595}	CD_{12600}	$CBFS_{13700}$
Model 1	0.3677	0.7279	0.6136
Model 2	0.4740	0.9510	0.9889
Frozen-training	0.3478	0.7472	0.6135
BSL-EARSM	0.6881	0.9821	1.0065
Reynolds shear stress τ_{12}			
	PH_{10595}	CD_{12600}	$CBFS_{13700}$
Model 1	0.5485	0.9204	0.721
Model 2	0.519	0.9545	0.834
Frozen-training	0.5535	0.9394	0.7162
BSL-EARSM	1.2229	0.9524	0.8061
Skin friction coefficient C_f			
	PH_{10595}	CD_{12600}	$CBFS_{13700}$
Model 1	0.6828	0.6004	0.3194
Model 2	0.5888	0.5548	0.5836
Frozen-training	0.6427	0.5958	0.3125
BSL-EARSM	0.8443	0.6383	0.2203

and that corresponds to a simple correction of the production terms in the $k-\omega$ transport equations.

Finally, the learned models are applied to an unseen flow case not included in the training and testing datasets, namely the periodic hill flow at $Re = 37000$, and the results are compared to the reference data and to the models mentioned in the above.

5.2. Results for flow configurations in the training and validation sets

The predictions of the discovered models Model 1 and Model 2 for the 3 test cases are presented in Figs. 9, 10, 11 and 12 for the streamwise velocity $\overline{u_1}$, the turbulent kinetic energy k, the Reynolds shear stress τ_{12} and the skin friction coefficient C_f, respectively. Their results are compared with those of the baseline k-ω SST model, the frozen-training SpaRTA model, the BSL-EARSM model and the high-fidelity data. For a more quantitative evaluation we report in Table 1 the mean squared errors on the longitudinal and wall-normal velocities, turbulent kinetic energy and Reynolds shear stress (τ_{12}) profiles, as well as for the skin friction coefficient C_f, respectively. In all cases, the errors are normalized with the mean-squared error of the baseline k-ω SST model.

Both discovered models drastically improve velocity when compared to the $k-\omega$-SST and the BSL-EARSM models, particularly within the separated regions (see Fig. 9) for the 3 flow cases. Furthermore, their accuracy is similar to the frozen-training SpaRTA. Model 2, learned using high fidelity velocity profiles and skin friction training data, provides a lower error in velocity than Model 1, learned using high fidelity Reynolds stress profiles as training data.

The discovered models improve the prediction of k profiles over the LEVM $k-\omega$ SST for the 3 learning flow configurations. As expected, Model 1 yields better prediction of turbulent kinetic energy than Model 2 for all cases. Nonetheless, Model 2 overall improves the prediction of k with respect to the baseline. Of note, the accuracy of Model 1 and the frozen-training model (also trained Reynolds stresses) in predicting k is very similar for the 3 flow cases.

The discovered models also improve the prediction of τ_{12} profiles over the LEVM (see Fig. 11 and Table 1), Model 1 having better prediction capabilities than Model 2 and its accuracy is similar to the frozen-training model.

The skin friction results are reported in Fig. 12 and Table 1. As expected, Model 2, trained on the skin friction, yields better predictions than Model 1 and the frozen-training model. Both models improve the prediction of the reattachment point location for cases PH_{10595} and $CBFS_{13700}$ with respect to the predictions of the $k-\omega$ SST model, as shown in Table 2. However, neither of the two models captures the small recirculation bubble predicted by the DNS of case CD_{12600}. The predictions of the separation and reattachment point locations for case CD_{12600} are therefore not shown in Table 2.

Table 2
Predicted separation and reattachment points (x/h) for various models.

Separation points		
	PH_{10595}	$CBFS_{13700}$
LES	0.2154	0.9429
k-ω SST	0.2696	0.7655
Model 1	0.278	1.0895
Model 2	0.278	1.0895
frozen-training	0.2845	1.0503
BSL-EARSM	0.2817	0.8181
Reattachment points		
	PH_{10595}	$CBFS_{13700}$
LES	4.7595	4.2409
k-ω SST	7.6418	6.0296
Model 1	4.860	4.3334
Model 2	4.860	4.1694
frozen-training	5.0115	4.4754
BSL-EARSM	4.5613	5.1107

Table 3
Mean-squared errors of predicted velocity, turbulent kinetic energy and Reynolds shear-stress profiles and skin friction distribution with respect to high-fidelity data for Model 1FF. The errors are normalized by the mean-squared error of the baseline k-ω SST model.

	$\overline{u}_1, \overline{u}_2$	k	τ_{12}	C_f
PH_{10595}	0.2029	0.3488	0.5524	0.6474
CD_{12600}	0.2048	0.7440	0.9388	0.5956
$CBFS_{13700}$	0.1694	0.6144	0.7209	0.3232

5.3. Model learning from full-field high fidelity data

In this Section, we investigate the effect of the amount of data required to train the model. For that reason we retrained Model 1 using the full high-fidelity fields for the Reynolds stresses, interpolated on the RANS computational grid.

The resulting learned model is hereafter referred-to as Model 1FF (FF: Full Field). The correction tensors involve only 1 out of the 36 candidate functions:

$$b_{ij}^{\Delta} = 0, \tag{28}$$

$$b_{ij}^{R} = 4.0120 \times 10^{-1} T_{ij}^{(1)}. \tag{29}$$

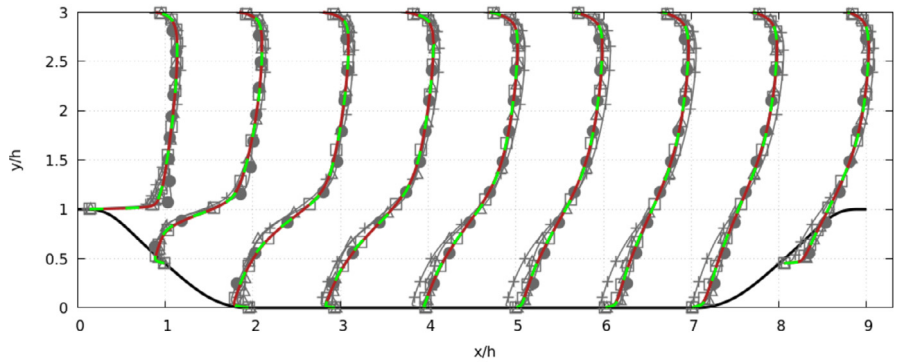
We report in Table 3 the mean squared errors on the longitudinal and wall-normal velocities, turbulent kinetic energy and Reynolds shear stress, and skin friction coefficient. We note that drastically increasing the amount of data does not improve solution accuracy.

Interestingly, the formulation of Model 1FF is very close to the formulation of model M1 of Ref. [1] (see equations (26) and (27)) trained on the same data. Moreover Model 1, Model 1FF and the full-field frozen-training model M1 have similar coefficients of the dominant term (4.6018×10^{-1}, 4.0120×10^{-1} and 3.9×10^{-1}, respectively). A possible interpretation is as follows: although more accurate than the baseline LEVM, a generalized eddy viscosity model remains an imperfect model that cannot match the data even with the best possible set of parameters (i.e. the mean squared error cannot be reduced below a given threshold). The algorithm compensates this effect by minimizing the second term in the cost function, leading to a sparser model.

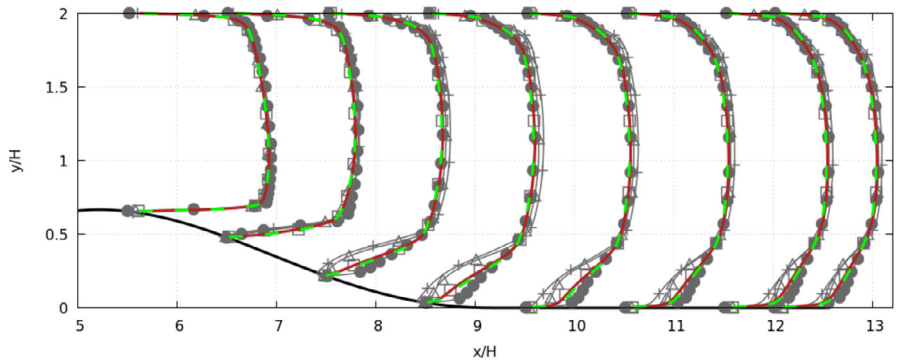
To better discriminate among the learned models, in the next Section we evaluate their extrapolation performance to a flow configuration outside the training and test sets.

5.4. Extrapolation to an unseen flow

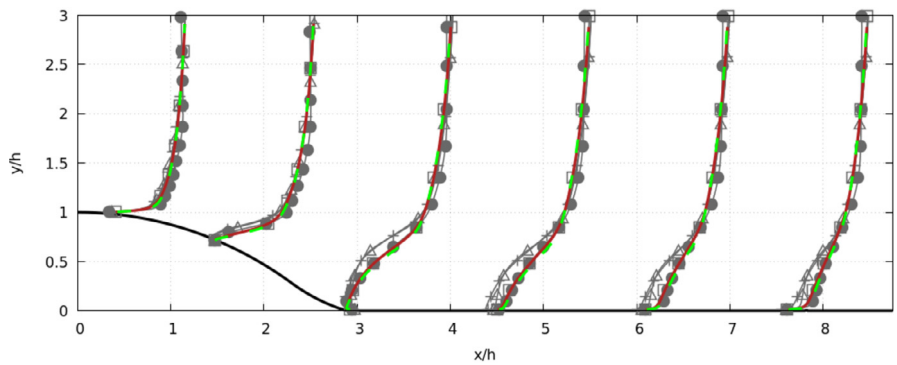
To test the ability of the CFD-driven SpaRTA to serve as a predictive model for 2D turbulent flow with separations, the two CFD-driven models learned for cases PH_{10595}, CD_{12600} and $CBFS_{13700}$ are applied to the 2D periodic hill flow case at Re=37000 (noted PH_{37000}). Model 1 and Model 2 have been learned on CD_{12600}, therefore, for both models, PH_{37000} represents an extrapolation to a higher Reynolds number and a different geometry, even if the flow still belongs to the class of 2D incompressible separated flow in a variable section channel. The results for the streamwise velocity $\overline{u_1}$, the turbulent kinetic energy k, the Reynolds shear stress τ_{12} and the skin friction coefficient C_f are compared to LES data from [47] in Figs. 13, 14, 15 and 16, respectively. The results of the baseline LEVM, of the BSL-EARSM and of the frozen-training SpaRTA model are also reported for comparison.



(a) PH_{10595}



(b) CD_{12600}

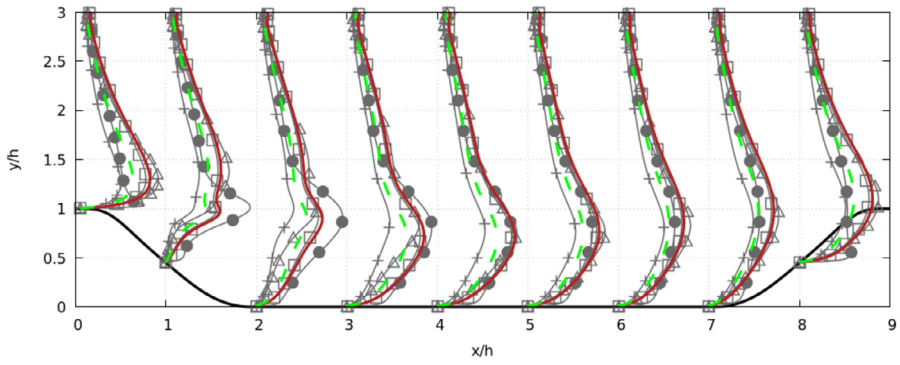


(c) CBFS_{13700}

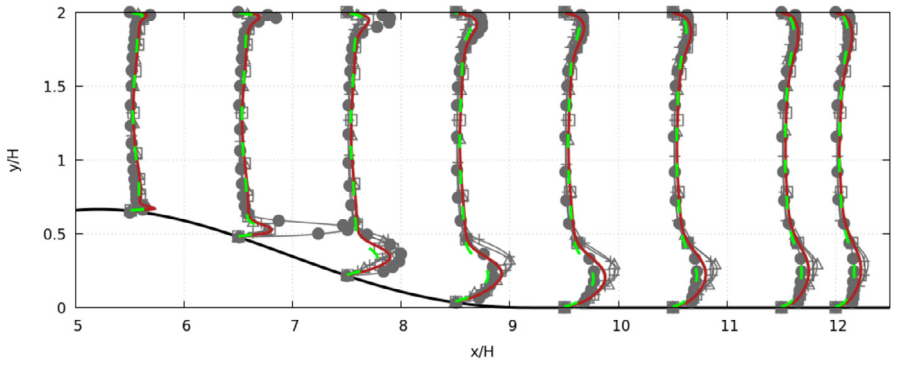
Fig. 9. Assessment of various CFD-driven, frozen-training and standard turbulence models for flows in the training and test set: streamwise velocity profiles ($\overline{u_1}/U_b + x/h$). k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

Table 4
Mean-squared errors of predicted velocity, turbulent kinetic energy and Reynolds shear-stress profiles and skin friction distribution with respect to high-fidelity data for various models applied to the PH case at Re=37000. The errors are normalized by the mean-squared error of the baseline k-ω SST model.

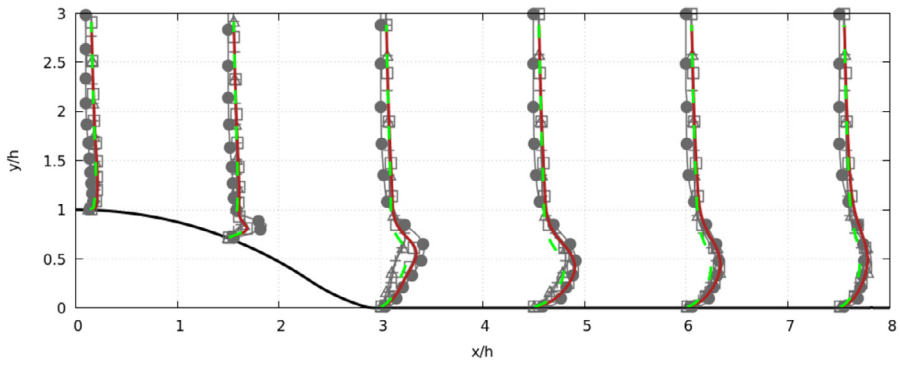
	$\overline{u}_1, \overline{u}_2$	k	τ_{12}	C_f
Model 1	0.2926	3.0006	0.6803	0.9828
Model 2	0.2726	1.1259	0.6606	0.8911
frozen-training	0.3280	2.5153	0.6841	0.9292
BSL-EARSM	0.4380	3.7959	1.4510	0.9912



(a) PH_{10595}



(b) CD_{12600}



(c) CBFS_{13700}

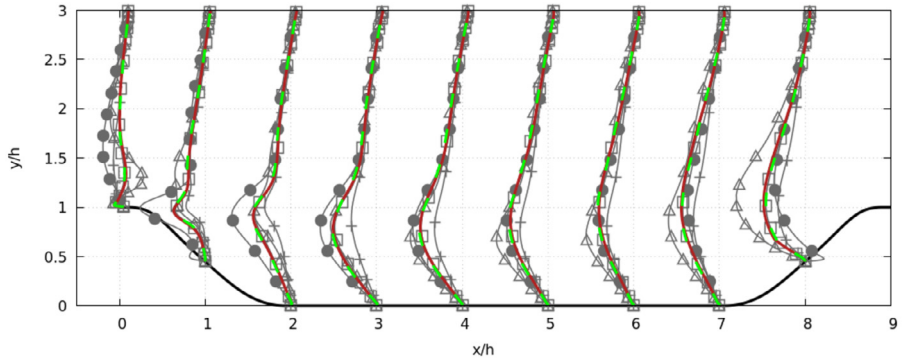
Fig. 10. Assessment of various CFD-driven, frozen-training and standard turbulence models for flows in the training and test set: turbulent kinetic energy profiles ($12k/U_b^2 + x/h$). k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

The normalized mean squared errors with respect to the high-fidelity data from [47] are reported in Table 4 and the predicted separation and reattachment points are given in Table 5.

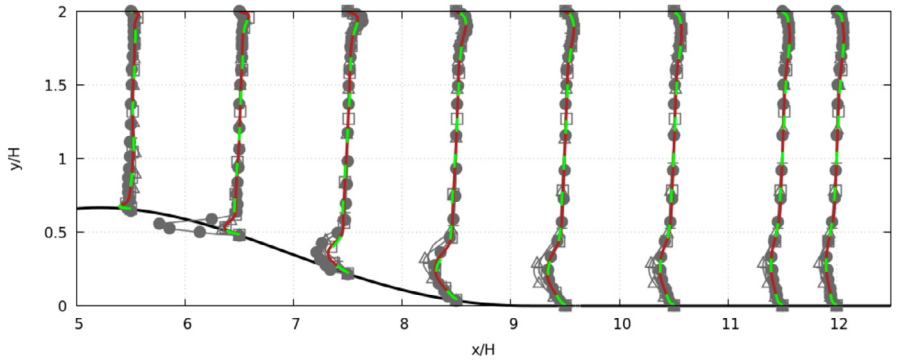
Model 1 and Model 2 improve the predictions of all quantities of interest, except the turbulence kinetic energy, showing their superior performance with respect to the baseline k-ω SST even at a higher Reynolds number and for a different geometry. Model 2 is more accurate than Model 1 and the frozen-training for each considered quantity of interest, while Model 1 yields overall the same accuracy as the frozen-training model.

BSL-EARSM provides poor predictions with respect to the learned models for PH_{37000}, degrading the accuracy of the k-ω SST model for 2 of the 4 quantities, one quantity being predicted with the same accuracy than the k-ω SST model.

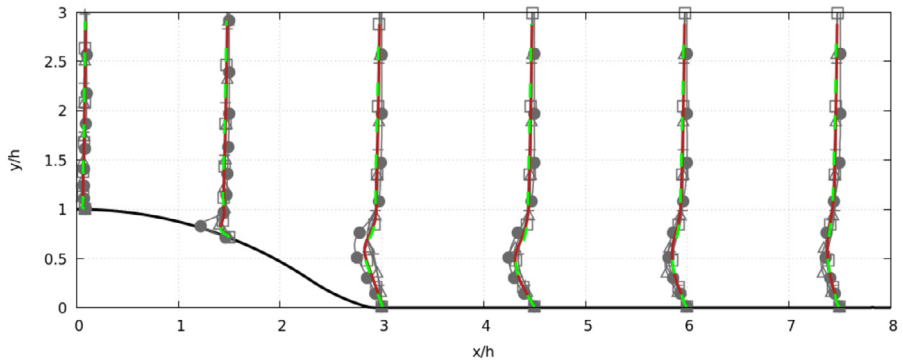
These results show that the CFD-driven models robustly improve predictive accuracy of the k-ω SST model also for a 2D separated flow outside the training set and exhibit a comparable or better extrapolation performance than the frozen-training SpaRTA.



(a) PH_{10595}



(b) CD_{12600}

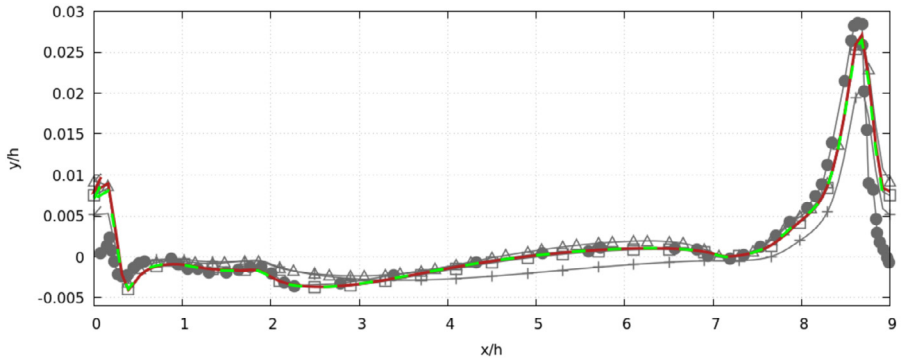


(c) CBFS_{13700}

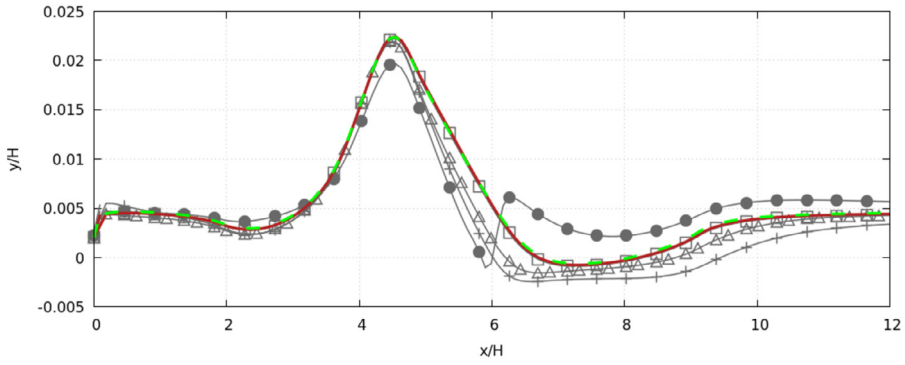
Fig. 11. Assessment of various CFD-driven, frozen-training and standard turbulence models for flows in the training and test set: Reynolds shear stress profiles ($20\tau_{12}/U_b^2 + x/h$). k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

Table 5
Predicted separation and reattachment points (x/h) for various models.

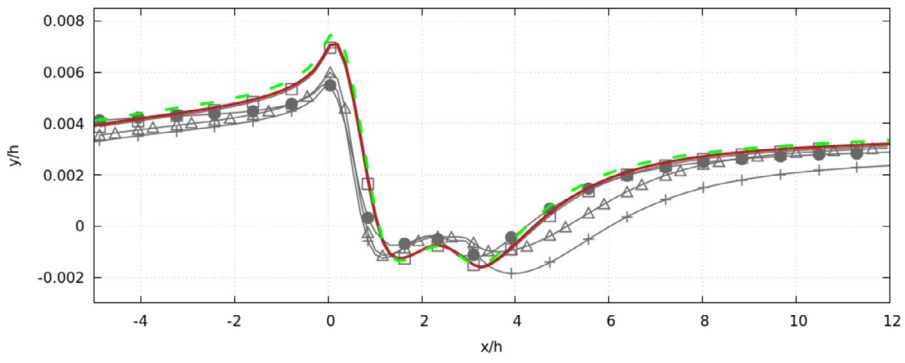
	separation	reattachment
LES	0.2552	4.0170
k-ω SST	0.2679	7.5479
Model 1	0.2780	4.4600
Model 2	0.2780	4.3799
frozen-training	0.2951	4.6303
BSL-EARSM	0.2898	4.2333



(a) PH_{10595}



(b) CD_{12600}



(c) CBFS_{13700}

Fig. 12. Assessment of various CFD-driven, frozen-training and standard turbulence models for flows in the training and test set: skin friction distributions along the bottom wall. k-ω-SST: (—+—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

6. Conclusions

A CFD-driven counterpart of the SpaRTA method initially proposed in [1] has been presented. The new method aims at discovering data-augmented EARSMs with improved performance for a given class of flows with respect to a baseline LEVM through deterministic symbolic identification of terms from a predefined library of candidate functions. As in SpaRTA, the discovered models use a generalized eddy viscosity formulation in conjunction with corrective terms for the production of the transported turbulent variables. The CFD-driven approach is intended to overcome limitations of the frozen-training model discovery in terms of numerical robustness of the learned models and to enable flexible use of any kind of data for model training. Importantly, the CFD-driven approach does not necessarily require data for second-order turbulence statistics, namely, the Reynolds stresses, which are not always easily available, e.g., in experimental data sets.

In the present work we focused on correcting the $k-\omega$ SST model for a class of 2D turbulent flows with separations caused by an enlargement of the channel section. The methodology may however be adapted to other eddy viscosity models and different flow configurations. The present test cases include the well-known periodic hill flow, a converging-diverging channel and the flow behind a curved backward facing step, for which high-fidelity DNS or LES data are available in the

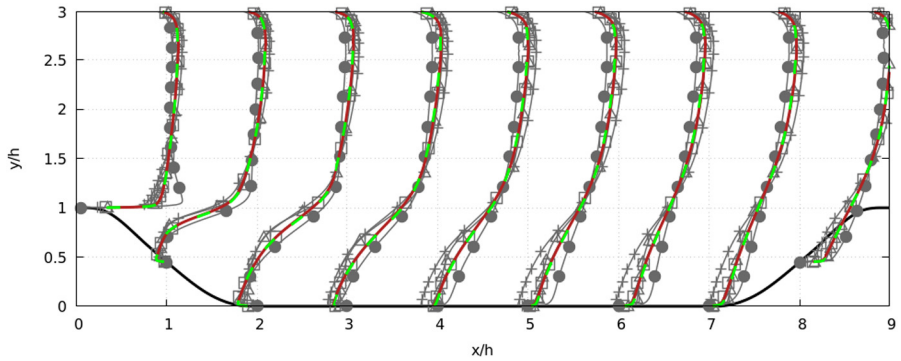


Fig. 13. Assessment of various CFD-driven, frozen-training and standard turbulence models for a flow outside the training and test set: streamwise velocity profiles normalized with bulk velocity: $\overline{u_1}/U_b + x$ for PH_{37000}. k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

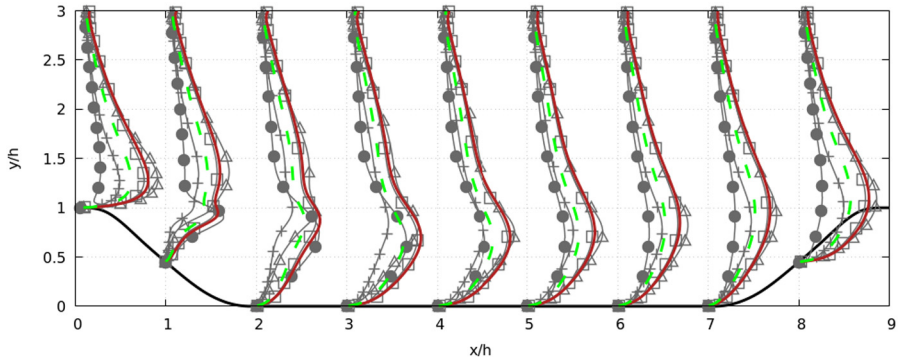


Fig. 14. Assessment of various CFD-driven, frozen-training and standard turbulence models for a flow outside the training and test set: turbulent kinetic energy profiles: $12k/U_b^2 + x$ for PH_{37000}. k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

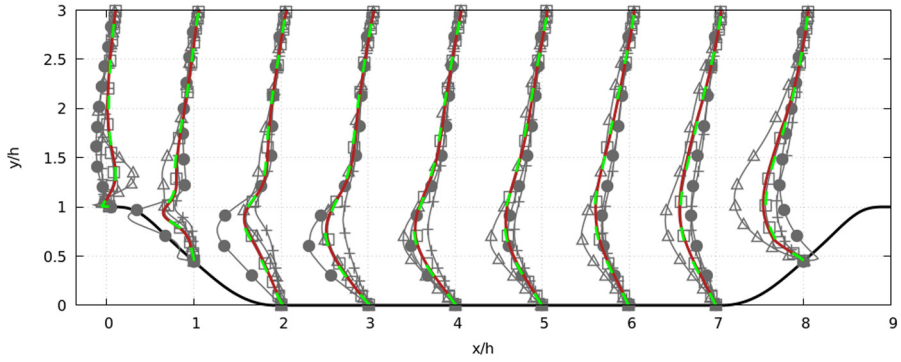


Fig. 15. Assessment of various CFD-driven, frozen-training and standard turbulence models for a flow outside the training and test set: Reynolds shear stress profiles: $20\tau_{12}/U_b^2 + x$ for PH_{37000}. k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

literature. To show the ability of the CFD-driven method to train models with any kind of data, two sets of high-fidelity data have been considered. First, models are learned using data for Reynolds stress profiles at a few streamwise locations. Second, we prove the flexibility of the proposed approach by training models against selected velocity profiles and skin friction data.

The computational cost of CFD-driven model identification is obviously greater than frozen-training approach. An important novel contribution is the introduction of a surrogate model for alleviating the cost of the CFD-driven training process. Such an approach, here coupled with the deterministic sparse identification procedure, could be also effectively coupled with different learning techniques, such as Gene Expression Programming ([27]), to lower the overall cost of CFD-driven

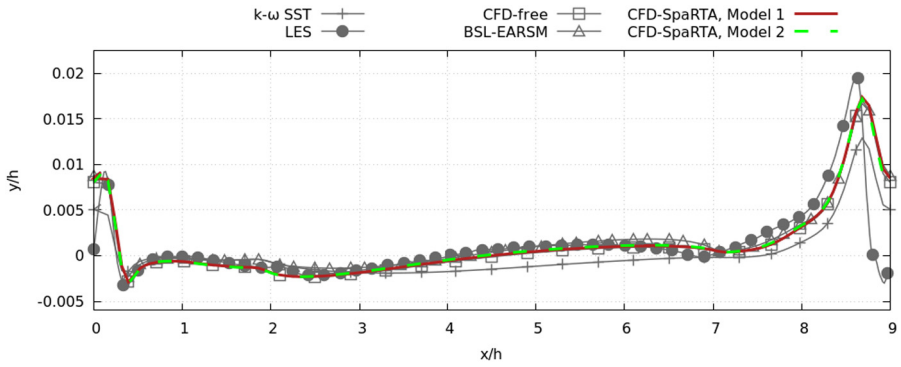


Fig. 16. Assessment of various CFD-driven, frozen-training and standard turbulence models for a flow outside the training and test set: skin friction distribution along the bottom wall. k-ω-SST: (—|—), High fidelity data: (—●—), Model 1: (——), Model 2: (- - -), frozen-training: (—□—), BSL-EARSM: (—△—).

training. By so doing, a complete model identification run, including model training and selection via cross-validation, requires approximately 5000 CFD solves, which remains affordable even on a standard workstation for 2D flow cases.

The CFD-driven discovered models are sparse and numerically robust, thanks to the regularization constraints on model parameters. They satisfy realizability conditions throughout the computational domain. For flow configurations in the training and validation sets, they exhibit improved accuracy not only over the baseline LEVM, but also over a general purpose EARSM model [32]. Furthermore, their have an accuracy comparable to a data-driven model generated with the frozen-training SpaRTA algorithm. The model learned from velocity and skin friction data is overall more accurate than the one trained on Reynolds stress data for the prediction of the velocity and skin friction. On the opposite, the model trained on the Reynolds stresses consistently provides better predictions of the turbulent kinetic energy and the Reynolds shear stress.

The effect of the amount of data required to train the model has then been investigated. Model 1 has been retrained using the full high-fidelity fields for the Reynolds stresses, interpolated on the RANS computational grid. We showed that drastically increasing the amount of data does not improve solution accuracy, but leads to a sparser model, whose formulation is very close to the frozen-training SpaRTA M1 model of Ref. [1], trained on the same data.

The CFD-driven models have then been applied to a separated flow outside the training and validation sets, characterized by a higher Reynolds number and a different geometry. For this configuration, the learned models provide significantly improved results for most flow quantities of interest, not only over the LEVM but also over the general purpose BLS-EARSM model. The CFD-driven models yield a comparable or better accuracy than the frozen-training SpaRTA.

In conclusion, the CFD-driven SpaRTA represents a promising method to generate data-driven EARSMs. The computational overcost of the training algorithm is justified by the possibility of using any kind of available data, including incomplete data sets, as those potentially obtained from experiments. Furthermore, such models are by construction numerically robust.

Code

An implementation of a general EARSM model in OpenFOAM is available here: https://github.com/shmlzr/general_earsm.git.

CRediT authorship contribution statement

Ismaïl Ben Hassan Saïdi: Conceptualization, Data curation, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Martin Schmelzer:** Conceptualization, Data curation, Methodology, Software, Writing – review & editing. **Paola Cinnella:** Conceptualization, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. **Francesco Grasso:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The present work was carried out in the frame of project PERTINENT (high-fidelity exPerimental and numERIcal study for daTa-drIven modeliNg of rEalistic turbuleNt separaTed flows), funded by Institut Carnot ARTS, France.

Appendix A. The k-ω SST model

Transport equations of k and ω write:

$$\frac{\partial k}{\partial t} + U_j \frac{\partial k}{\partial x_j} = P_k - \beta^* k\omega + \frac{\partial}{\partial x_j}[(\nu + \sigma_k \nu_t)\frac{\partial k}{\partial x_j}] \tag{A.1}$$

$$\frac{\partial \omega}{\partial t} + U_j \frac{\partial \omega}{\partial x_j} = \frac{\alpha}{\nu_t} P_k - \beta\omega^2 + \frac{\partial}{\partial x_j}[(\nu + \sigma_\omega \nu_t)\frac{\partial \omega}{\partial x_j}] + CD_{k\omega}, \tag{A.2}$$

with $P_k = -2\nu_t S_{ij}\partial_j U_i$. The corresponding eddy viscosity writes

$$\nu_t = \frac{a_1 k}{\max(a_1\omega, SF_2)}. \tag{A.3}$$

The other standard terms read

$$\begin{aligned} &CD_{k\omega} = \max(2\sigma_{\omega^2}\tfrac{1}{\omega}(\partial_i k)(\partial_i \omega), 10^{-10}),\\ &F_1 = \tanh[(\min[\max(\tfrac{\sqrt{k}}{\beta^*\omega y}, \tfrac{500\nu}{y^2\omega}), \tfrac{4\sigma_{\omega^2}k}{CD_{k\omega}y^2}])^4],\\ &F_2 = \tanh[(\max(\tfrac{2\sqrt{k}}{\beta^*\omega y}, \tfrac{500\nu}{y^2\omega})])^2],\\ &\Phi = F_1\Phi_1 + (1-F_1)\Phi_1, \end{aligned} \tag{A.4}$$

in which the latter blends the coefficients $\Phi \to (\Phi_1, \Phi_2)$

$$\alpha = (5/9, 0.44), \quad \beta = (3/40, 0.0828), \quad \sigma_k = (0.85, 1.0), \quad \sigma_\omega = (0.5, 0.856). \tag{A.5}$$

The remaining terms are $\beta^* = 0.09$, $a_1 = 0.31$.

Appendix B. The BSL-EARSMmodel

We here recall the formulation of the BSL-EARSM of Menter et al. [32]. This model is based on the EARSM formulation of Wallin and Johansson [48] (WJ model) for the stress-strain relationship. In the WJ model, the stress-strain relationship is combined with the k-ω transport equations of Wilcox [2]. In the BSL-EARSM, in order to avoid the freestream sensitivity of the Wilcox model, the WJ stress-strain relationship is combined with the BSL k-ω model of Menter [31].

Following [7], the Reynolds stress anisotropy tensor a_{ij} is projected onto a tensor basis:

$$a_{ij} = \beta_1 T_{1,ij} + \beta_2 T_{2,ij} + \beta_3 T_{3,ij} + \beta_4 T_{4,ij} + \beta_6 T_{6,ij} + \beta_9 T_{9,ij}, \tag{B.1}$$

where

$$\begin{aligned} &T_{1,ij} = S^*_{ij}; \quad T_{2,ij} = S^*_{ik}S^*_{kj} - \tfrac{1}{3}I_1\delta_{ij}; \quad T_{3,ij} = \Omega^*_{ik}\Omega^*_{kj} - \tfrac{1}{3}I_2\delta_{ij};\\ &T_{4,ij} = S^*_{ik}\Omega^*_{kj} - \Omega^*_{ik}S_{kj}; \quad T_{6,ij} = S^*_{ik}\Omega^*_{kl}\Omega^*_{lj} + \Omega^*_{ik}\Omega^*_{kl}S_{lj} - \tfrac{2}{3}I_4\delta_{ij} - I_2S^*_{ij};\\ &T_{9,ij} = \Omega^*_{ik}S^*_{kl}\Omega^*_{lm}\Omega^*_{mj} - \Omega^*_{ik}\Omega^*_{kl}S^*_{lm}\Omega^*_{mj} + \tfrac{1}{2}I_2(S^*_{ik}\Omega^*_{kj} - \Omega^*_{ik}S^*_{kj}), \end{aligned} \tag{B.2}$$

with S^*_{ij} and Ω^*_{ij}, the non-dimensional mean strain rate and rotation rate defined as follows:

$$S^*_{ij} = \frac{\tau}{2}(\frac{\partial U_i}{\partial x_j} + \frac{\partial U_j}{\partial x_i}), \quad \Omega^*_{ij} = \frac{\tau}{2}(\frac{\partial U_i}{\partial x_j} - \frac{\partial U_j}{\partial x_i}) \tag{B.3}$$

where τ is a turbulent time scale with a Kolmogorov limiter [49]:

$$\tau = \max(\frac{1}{C_\mu\omega}, 6\sqrt{\frac{\nu}{C_\mu k\omega}}). \tag{B.4}$$

The tensor invariants I_1, I_2 and I_4 read:

$$I_1 = S^*_{ij}S^*_{ji}, \quad I_2 = \Omega^*_{ij}\Omega^*_{ji}, \quad I_4 = S^*_{ik}\Omega^*_{kj}\Omega^*_{ji}. \tag{B.5}$$

The coefficients of the tensor basis β_i in (B.1) are defined as:

$$\beta_1 = -\frac{N}{Q}, \quad \beta_2 = 0, \quad \beta_3 = -\frac{2I_4}{NQ_1}, \quad \beta_4 = -\frac{1}{Q}, \quad \beta_6 = -\frac{N}{Q_1}, \quad \beta_9 = \frac{1}{Q_1}, \tag{B.6}$$

with

$$Q = \frac{(N^2 - 2I_2)}{A_1}, \quad Q_1 = \frac{Q}{6}(2N^2 - I_2) \tag{B.7}$$

where

$$N = C_1' + \frac{9}{4}\frac{\tilde{P}_k}{\epsilon} \tag{B.8}$$

and

$$A_1 = 1.2, \quad C_1' = \frac{9}{4}(C_1 - 1) \quad \text{and} \quad C_1 = 1.8. \tag{B.9}$$

N is a solution of the cubic equation:

$$N^3 - C_1' N^2 - (2.7I_1 + 2I_2)N + 2C_1' I_2 = 0 \tag{B.10}$$

which is given by:

$$\begin{cases} N = \frac{C_1'}{3} + (P_1 + \sqrt{P_2})^{1/3} + \text{sign}(P_1 - \sqrt{P2}) \mid P_1 - \sqrt{P_2} \mid^{1/3} & \text{at} \quad P_2 \geq 0 \\ N = \frac{C_1'}{3} + 2(P_1^2 - P_2)^{1/6}\cos(\frac{1}{3}\arccos(\frac{P_1}{\sqrt{P_1^2 - P_2}})) & \text{at} \quad P_2 < 0 \end{cases} \tag{B.11}$$

with

$$P_1 = C_1'(\frac{C_1'^2}{27} + \frac{9}{20}I_1 - \frac{2}{3}I_2), \quad P_2 = P_1^2 - (\frac{C_1'^2}{9} + \frac{9}{10}I_1 + \frac{2}{3}I_2)^3. \tag{B.12}$$

The BSL-EARSM constitutive equation is then supplemented by transport equations for k and ω. These are the same as in Appendix A, where a modified production term is used:

$$\tilde{P}_k = \min(-\tau_{ij}\frac{\partial U_i}{\partial x_j}, 10.\rho\beta^* k\omega). \tag{B.13}$$

The reader is referred to [32] for further details.

Appendix C. CORS algorithm

The main steps of the algorithm proceed as follows:

- The total number of function evaluations, noted N, is set from the beginning.
- Initial step:
  - An initial sampling of size n of the parameter space $S_1 = \left\{\hat{\Theta}_i | i = 1, 2, \ldots, n\right\}$ is performed using the Latin Hypercube sampling method. It must satisfy the condition $D < n < N$, where D is the dimensionality of the search space.
  - An initial response surface $\hat{f}_1$ is constructed using cubic radial basis functions (RBF) and the initial sampling S_1. The functions are of the form:

$$\hat{f}_1(\hat{\Theta}) = \sum_{i=1}^{N/2} \lambda_i \phi(\|\hat{\Theta} - \hat{\Theta}_i\|) + b^T\hat{\Theta} + a, \tag{C.1}$$

    where ϕ is a cubic function ($\phi(r) = r^3$) and λ_i, b and a are coefficients determined by interpolating the available samples.
- Steps 1 to $m = N - n$:
  - The candidate sample $\hat{\Theta}_i$ minimizing the surface response $\hat{f}_{i-1}$ is selected. In order to prevents the algorithm from being trapped in a local minimum, a ball of radius r is placed around each of previously sampled points and the candidate sample minimizing the surface response is required to be outside of any ball. The radius of the balls r decrease (all balls have the same radius) with iterations, with a given rate. The size of the balls is controlled by two parameters: their initial density ρ_0, and the rate of decay of their radius p. The density ρ is the total volume of the balls divided by the total volume of the search space. At a given iteration i ($1 \leq i \leq N/2$), the density and the radius are:

$$\rho_i = \rho_0(\frac{m - i}{m - 1})^p \quad , \tag{C.2}$$

$$r_i = (\frac{\rho_i}{(n + i - 1)v_1})^{\frac{1}{D}}, \tag{C.3}$$

where n and m are respectively the number of initial samples and the number of subsequent steps; D is the dimensionality of the search space and v_1 is a volume of a ball with radius 1, that can be expressed with the gamma function Γ:

$$v_1 = \frac{\pi^{\frac{D}{2}}}{\Gamma(\frac{D}{2}+1)} \quad . \tag{C.4}$$

– The cost function is evaluated for $\hat{\Theta}_i$.
– A new response surface $\hat{f}_i$ is constructed using the sampling $S_{i-1} \bigcup \hat{\Theta}_i$

The candidate sample $\hat{\Theta}$ corresponding to the lowest evaluated value of the cost function is the solution of the problem proposed by the algorithm.

During both the initial and subsequent steps, RANS simulations are performed using candidate models in order to evaluate the cost function of the optimization problem. To force the research towards robust models, each model preventing the convergence of the solution is eliminated and a new model is resampled by introducing a slight perturbation into the non robust model. This procedure is repeated until a candidate model allowing convergence of the RANS simulation is obtained.

References

[1] M. Schmelzer, R.P. Dwight, P. Cinnella, Discovery of algebraic reynolds-stress models using sparse symbolic regression, Flow Turbul. Combust. 104 (2) (2020) 579–603.
[2] D.C. Wilcox, et al., Turbulence Modeling for CFD, 3rd edition, DCW Industries La Canada, CA, 2006.
[3] A. Hellsten, S. Wallin, Explicit algebraic reynolds stress and non-linear eddy-viscosity models, Int. J. Comput. Fluid Dyn. 23 (4) (2009) 349–361.
[4] B.E. Launder, G.J. Reece, W. Rodi, Progress in the development of a reynolds-stress turbulence closure, J. Fluid Mech. 68 (3) (1975) 537–566.
[5] S.B. Pope, Turbulent Flows, 2001.
[6] H.K. Versteeg, W. Malalasekera, An Introduction to Computational Fluid Dynamics: the Finite Volume Method, Pearson Education, 2007.
[7] S. Pope, A more general effective-viscosity hypothesis, J. Fluid Mech. 72 (2) (1975) 331–340.
[8] T.B. Gatski, C.G. Speziale, On explicit algebraic stress models for complex turbulent flows, J. Fluid Mech. 254 (1993) 59–78, https://doi.org/10.1017/S0022112093002034.
[9] H. Xiao, P. Cinnella, Quantification of model uncertainty in rans simulations: a review, Prog. Aerosp. Sci. 108 (2019) 1–31.
[10] K. Duraisamy, G. Iaccarino, H. Xiao, Turbulence modeling in the age of data, Annu. Rev. Fluid Mech. 51 (2019) 357–377, https://doi.org/10.1146/annurev-fluid-010518-040547.
[11] E.J. Paris, K. Duraisamy, A paradigm for data-driven predictive modeling using field inversion and machine learning, J. Comput. Phys. 305 (2016) 758.
[12] J. Ling, A. Kurzawski, J. Templeton, Reynolds averaged turbulence modelling using deep neural networks with embedded invariance, J. Fluid Mech. 807 (2016) 155–166, https://doi.org/10.1017/jfm.2016.615.
[13] J.L. Wu, H. Xiao, E. Paterson, Physics-informed machine learning approach for augmenting turbulence models: a comprehensive framework, Phys. Rev. Fluids 3 (2018) 074602.
[14] J. Weatheritt, R. Sandberg, A novel evolutionary algorithm applied to algebraic modifications of the rans stress–strain relationship, J. Comput. Phys. 325 (2016) 22–37.
[15] J. Weatheritt, R. Sandberg, The development of algebraic stress models using a novel evolutionary algorithm, Int. J. Heat Fluid Flow 68 (2017) 298–318.
[16] H.D. Akolekar, J. Weatheritt, N. Hutchins, R.D. Sandberg, G. Laskowski, V. Michelassi, Development and use of machine-learnt algebraic reynolds stress models for enhanced prediction of wake mixing in low-pressure turbines, J. Turbomach. 141 (4) (2019).
[17] A. Gargiulo, T.A. Ozoroski, T. Hallock, A. Haghiri, R.D. Sandberg, M. Visonneau, G. Deng, E. Guilmineau, D. Geneau, T. Jeans, M. Bettle, M. Kerkvliet, S.L. Toxopeus, J.E. Duetsch-Patel, V. Sundarraj, A. Borgoltz, W.J. Devenport, C.J. Roy, T. Lowe, Computations of the BeVERLI Hill Three-Dimensional Separating Flow Model Validation Cases, 2022, pp. 1–33, https://doi.org/10.2514/6.2022-1034.
[18] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. 113 (15) (2016) 3932–3937.
[19] S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. 3 (4) (2017) e1602614.
[20] J.P. Huijing, R.P. Dwight, M. Schmelzer, Data-driven RANS closures for three-dimensional flows around bluff bodies, Comput. Fluids 225 (2021) 104997, https://doi.org/10.1016/j.compfluid.2021.104997, arXiv:2010.11831.
[21] J. Steiner, R.P. Dwight, A. Viré, Data-driven RANS closures for wind turbine wakes under neutral conditions, Comput. Fluids 233 (2022) 105213, https://doi.org/10.1016/j.compfluid.2021.105213, arXiv:2009.10816.
[22] Y. Zhang, R. Dwight, M. Schmelzer, J. Gómez, Z.-H. Han, S. Hickel, Customized data-driven rans closures for bi-fidelity les–rans optimization, J. Comput. Phys. 432 (2021) 110153, https://doi.org/10.1016/j.jcp.2021.110153.
[23] M. Wang, C. Chen, W. Liu, Establish algebraic data-driven constitutive models for elastic solids with a tensorial sparse symbolic regression method and a hybrid feature selection technique, J. Mech. Phys. Solids 159 (2022) 104742, https://doi.org/10.1016/j.jmps.2021.104742, https://www.sciencedirect.com/science/article/pii/S0022509621003471.
[24] S. Beetham, J. Capecelatro, Formulating turbulence closures using sparse regression with embedded form invariance, Phys. Rev. Fluids 5 (08 2020), https://doi.org/10.1103/PhysRevFluids.5.084611.
[25] K. Duraisamy, Perspectives on machine learning-augmented reynolds-averaged and large eddy simulation models of turbulence, Phys. Rev. Fluids 6 (2021) 050504.
[26] J.R. Holland, J.D. Baeder, K. Duraisamy, Towards integrated field inversion and machine learning with embedded neural networks for RANS modeling, in: AIAA Scitech 2019 Forum, 2019, p. 1884, https://arc.aiaa.org/doi/pdf/10.2514/6.2019-1884.
[27] Y. Zhao, H.D. Akolekar, J. Weatheritt, V. Michelassi, R.D. Sandberg, Rans turbulence model development using cfd-driven machine learning, J. Comput. Phys. 411 (2020) 109413.
[28] C.A. Michélen Str ofer, H. Xiao, End-to-end differentiable learning of turbulence models from indirect observations, Theor. Appl. Mech. Lett. 11 (2021) 100280.
[29] S. Banerjee, R. Krahl, F. Durst, C. Zenger, Presentation of anisotropy properties of turbulence, invariants versus eigenvalue approaches, J. Turbul. 8 (2007) N32.
[30] R.G. Regis, C.A. Shoemaker, Constrained global optimization of expensive black box functions using radial basis functions, J. Glob. Optim. 31 (1) (2005) 153–171.

[31] F.R. Menter, Two-equation eddy-viscosity turbulence model for engineering applications, AIAA J. 32 (1994) 1598–1605.
[32] F. Menter, A. Garbaruk, Y. Egorov, Explicit algebraic reynolds stress models for anisotropic wall-bounded flows, Progr. Flight Phys. 3 (2012) 89–104.
[33] H.G. Weller, G. Tabor, H. Jasak, C. Fureby, A tensorial approach to computational continuum mechanics using object-oriented techniques, Comput. Phys. 12 (6) (1998) 620–631.
[34] L. Caretto, A. Gosman, S. Patankar, D. Spalding, Two calculation procedures for steady, three-dimensional flows with recirculation, in: Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics, Springer, 1973, pp. 60–68.
[35] M. Breuer, N. Peller, C. Rapp, M. Manhart, Flow over periodic hills–numerical and experimental study in a wide range of reynolds numbers, Comput. Fluids 38 (2) (2009) 433–457.
[36] M. Marquillie, U. Ehrenstein, J.-P. Laval, et al., Instability of streaks in wall turbulence with adverse pressure gradient, J. Fluid Mech. 681 (205–240) (2011) 30.
[37] Y. Bentaleb, S. Lardeau, M.A. Leschziner, Large-eddy simulation of turbulent boundary layer separation from a rounded step, J. Turbul. 13 (2012) N4.
[38] S.L. Brunton, J.N. Kutz, Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control, Cambridge University Press, 2019.
[39] T. McConaghy, Ffx: fast, scalable, deterministic symbolic regression technology, in: Genetic Programming Theory and Practice IX, Springer, 2011, pp. 235–260.
[40] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
[41] P. Knysh, Y. Korkolis, Blackbox: a procedure for parallel optimization of expensive black-box functions, preprint, arXiv:1605.00998, 2016.
[42] C.G. Speziale, R. Abid, P.A. Durbin, On the realizability of reynolds stress turbulence closures, J. Sci. Comput. 9 (4) (1994) 369–403.
[43] M. Emory, G. Iaccarino, Visualizing Turbulence Anisotropy in the Spatial Domain with Componentality Contours, Center for Turbulence Research Annual Research Briefs, 2014, pp. 123–138.
[44] W.N. Edeling, G. Iaccarino, P. Cinnella, Data-free and data-driven rans predictions with quantified uncertainty, Flow Turbul. Combust. 100 (3) (2018) 593–616.
[45] M. Kommenda, B. Burlacu, G. Kronberger, M. Affenzeller, Parameter identification for symbolic regression using nonlinear least squares, Genet. Program. Evol. Mach. 21 (2020) 471–501, https://doi.org/10.1007/s10710-019-09371-3.
[46] Z. Yang, Z. Wang, H. Liu, Y. Eldar, T. Zhang, Sparse nonlinear regression: parameter estimation under nonconvexity, in: M.F. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, in: Proceedings of Machine, vol. 48, Learning Research, PMLR, New York, New York, USA, 2016, pp. 2472–2481, http://proceedings.mlr.press/v48/yangc16.html.
[47] X. Gloerfelt, P. Cinnella, Large eddy simulation requirements for the flow over periodic hills, Flow Turbul. Combust. 103 (1) (2019) 55–91.
[48] S. Wallin, A.V. Johansson, An explicit algebraic reynolds stress model for incompressible and compressible turbulent flows, J. Fluid Mech. 403 (2000) 89–132.
[49] P.A. Durbin, Near-wall turbulence closure modeling without “damping functions”, Theor. Comput. Fluid Dyn. 3 (1) (1991) 1–13.