Discovery of Algebraic Reynolds-Stress Models Using Sparse Symbolic Regression

Martin Schmelzer, Richard P Dwight, Paola Cinnella

Discovery of Algebraic Reynolds-Stress Models Using Sparse Symbolic Regression

Martin Schmelzer¹ [iD] **· Richard P. Dwight¹ · Paola Cinnella²**

Abstract

A novel deterministic symbolic regression method SpaRTA (Sparse Regression of Turbulent Stress Anisotropy) is introduced to infer algebraic stress models for the closure of RANS equations directly from high-fidelity LES or DNS data. The models are written as tensor polynomials and are built from a library of candidate functions. The machine-learning method is based on elastic net regularisation which promotes sparsity of the inferred models. By being data-driven the method relaxes assumptions commonly made in the process of model development. Model-discovery and cross-validation is performed for three cases of separating flows, i.e. periodic hills (Re=10595), converging-diverging channel (Re=12600) and curved backward-facing step (Re=13700). The predictions of the discovered models are significantly improved over the k-ω SST also for a *true* prediction of the flow over periodic hills at Re=37000. This study shows a systematic assessment of SpaRTA for rapid machine-learning of robust corrections for standard RANS turbulence models.

Keywords Turbulence modelling · Machine learning · Sparse symbolic regression · Explicit Algebraic Reynolds-stress models · Data-driven

1 Introduction

The capability of Computational Fluid Dynamics (CFD) to deliver reliable prediction is limited by the unsolved closure problem of turbulence modelling. The workhorse for turbulence modelling in industry are the Reynolds-Averaged Navier-Stokes (RANS) equations using linear eddy viscosity models (LEVM) [1]. The lower computational costs compared to high-fidelity approaches, e.g. Large-Eddy (LES) or Direct Numerical Simulations (DNS),

✉ Martin Schmelzer
  m.schmelzer@tudelft.nl

¹ Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 2,
  Delft, The Netherlands

² Laboratoire DynFluid, Arts et Métiers ParisTech, 151 Boulevard de l'Hopital, 75013 Paris, France

come at the price of uncertainty especially for flows with separation, adverse pressure gradients or high streamline curvature. Data-driven methods for turbulence modelling based on supervised machine learning have been introduced to leverage RANS for improved predictions [2–4]. In [5], the source terms of the Spalart-Allmaras were learnt from data using a single hidden layer neural network, which served as a first feasibility study. In [6], a factor was introduced to correct the turbulent production in the k-equation of the k-ω model. This term was found via inverse modelling and served to train a Gaussian process. While this approach has been extended and applied to industrially relevant flows such as airfoils in [7, 8] it still relies on the Boussinesq assumption. In [9], a deep neural network was trained to predict a_{ij} given input only from a baseline linear eddy viscosity simulation and thus replacing the turbulence model instead of augmenting it. The network was designed to embed Galilean invariance of the predicted a_{ij}. This concept of physics-informed machine learning was extended, e.g., in [10] using random forest regression. Despite the success of the data-driven approaches a drawback is their black box nature, which hampers the understanding of the physics of the resulting models in order to derive new modelling ideas from it.

Recently, a method has been introduced using genetic-programming (GEP) based symbolic regression to derive Explicit Algebraic Reynolds-stress Models (EARSM) directly from high-fidelity data [11, 12]. EARSM, first introduced by [13] and further developed by [14], are nonlinear extensions of LEVM and are commonly derived by projecting Reynolds-stress models (RSM) onto a set of tensorial polynomials [15, 16]. These models are numerically more robust than RSM at similar computational costs as LEVM [17], but do not show superior predictive capabilities for all kinds of flows [15]. The data-driven GEP method retains the input quantities used to derive EARSM, but replaces the commonly used projection method to find the formal structure of the model by an evolutionary process, which makes it an open-box machine learning approach. The advantage of such a data-driven method is that instead of relying on assumptions made during the development of an EARSM, a model is inferred directly from data. While such a model might not provide an universal approach for all kinds of flows as commonly aimed for in physical modelling, it serves as a pragmatic tool to correct the flow at hand. For cases exhibiting similar flow physics, e.g. separation, it has also been shown that the discovered models provide suitable corrections indicating the predictive potential of a data-driven approach.

Due to the non-deterministic nature of GEP it discovers for each run another model with a different mathematical form, e.g. other terms and/or other values for coefficients, with varying complexity. It is reported that the models using only a few nonlinear terms show a low training and prediction error as well as high numerical robustness for industrially relevant flow cases [18, 19]. Therefore, we instead introduce a new deterministic symbolic regression method SpaRTA (Sparse Regression of Turbulent Stress Anisotropy), for which we constrain the search towards sparse algebraic models using sparsity-promoting regression techniques [20, 21]. SpaRTA combines functions from a predefined library of candidates without any random recombination. It consists of four steps: (i) building a library of candidate functions, (ii) model selection using sparse-regression techniques, (iii) inference of model coefficients and (iv) cross-validation of the resulting models, see Fig. 1. The first three steps are computationally very cheap also for high-dimensional problems and allow for rapid model discovery.

The present study provides several novel concepts for data-driven modelling, which are organised as follows. In Section 2 we define additive model-form error terms within the k-ω SST LEVM model and use k-corrective-frozen-RANS, which is an extension of the method introduced in [12], to compute the model-form error from high-fidelity data. The novelty in
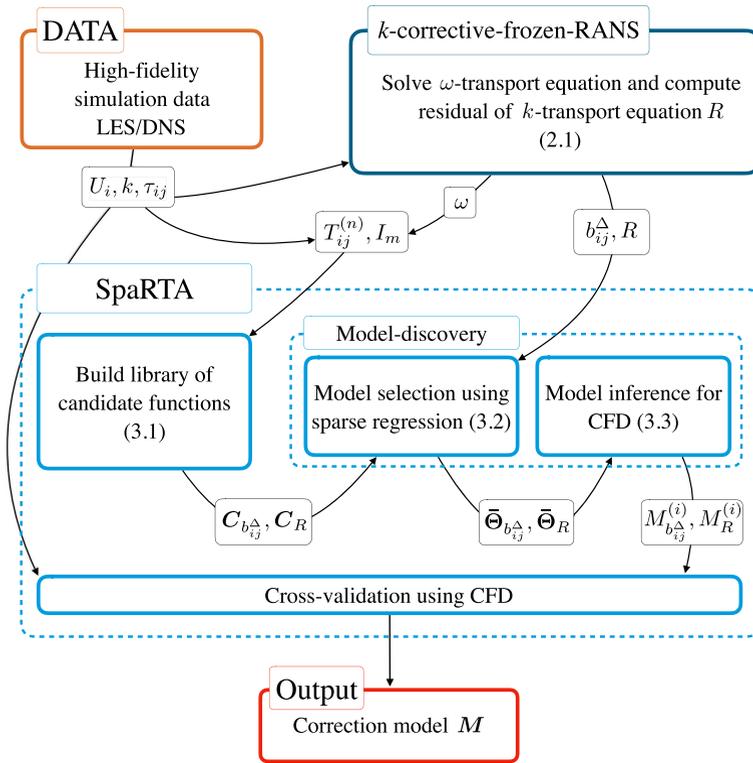
Fig. 1 Technical flow diagram of SpaRTA (Sparse Regression of Turbulent Stress Anisotropy)

this work is that we identify not only a correction of the stress-strain relation, but also one for the turbulent transport equations and thereby achieve excellent agreement with mean-fields of high-fidelity data. We also validate that the model-form error is successfully captured by adding the two terms to the solver and performing a CFD simulation. The k-corrective-frozen-RANS does not require any iterative optimisation procedure as compared to [6] and is therefore very efficient, but also limited to full-field data. In Section 3 we introduce the steps of SpaRTA. The details of the test cases, the CFD setup and the sources of the high-fidelity data are given in Section 4. In Section 5 SpaRTA is applied to the test cases, the discovered models are presented and the best models are chosen using cross-validation. Finally, conclusions are drawn in Section 6.

2 Model-Form Error of RANS Equations

In the following, we augment the baseline model, i.e. the linear eddy viscosity assumption and the turbulence transport equations of the k-ω SST, with additive terms accounting for the error due to the model-form. We introduce k-corrective-frozen-RANS, which is an extension of the method in [12], to extract these two types of error from high-fidelity data sources efficiently. Finally, we validate that the extracted terms reduce the error for given test cases.

2.1 Identification of additive model-form error from data

The incompressible and constant-density RANS equations read

$$\partial_i U_i = 0,$$
$$U_j \partial_j U_i = \partial_j \left[-\frac{1}{\rho} P + \nu \partial_j U_i - \tau_{ij} \right], \tag{1}$$

where U_i is the mean velocity, ρ is the constant density, P is the mean pressure and ν is the kinematic viscosity. The Reynolds-stress τ_{ij} is the subject of modelling. This symmetric, second-order tensor field can be decomposed into an anisotropic $a_{ij} = 2k b_{ij}$ and isotropic part $\frac{2}{3} k \delta_{ij}$

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

in which the baseline model, $b_{ij}^o = -\frac{\nu_t}{k} S_{ij}$, forms a linear relation between anisotropy and the mean-strain rate tensor S_{ij} via the scalar eddy viscosity ν_t. Commonly, ν_t is computed using a transport model such as k-ω SST [15], in which k is the turbulent kinetic energy and ω the specific dissipation rate.

In order to extract the model-form error in these models from high-fidelity data sources, we compute the residuals of the baseline turbulence model given the data. The residual for the constitutive relation is equivalent to an additive term b_{ij}^A leading to an augmented constitutive relation

$$b_{ij} = -\frac{\nu_t}{k} S_{ij} + b_{ij}^A. \tag{3}$$

To evaluate b_{ij}^A it is necessary to estimate ν_t, therefore also ω needs to be specified. In [12, 22], ω was efficiently obtained by passively solving the ω transport equation given high-fidelity data for U_i, k and b_{ij}. The associated ν_t was then used to compute b_{ij}^A with Eq. 3. This method is named frozen-RANS as only one equation is solved iteratively while the remaining variables are frozen [23, 24]. Despite the fact that b_{ij}^A also alters the production of turbulent kinetic energy P_k, it is not evident that solving the k equation given the data and the frozen ω should lead to the same k as present in the data. Therefore, we introduce k-corrective-frozen-RANS for which we also compute the residual of the k equation alongside the computation of the frozen ω. The residual is equivalent to an additive correction term, which we define as R, leading to an augmented k-ω SST model

$$\partial_t k + U_j \partial_j k = P_k + R - \beta^* \omega k + \partial_j \left[(\nu + \sigma_k \nu_t) \partial_j k \right], \tag{4}$$
$$\partial_t \omega + U_j \partial_j \omega = \frac{\gamma}{\nu_t} (P_k + R) - \beta \omega^2 + \partial_j \left[(\nu + \sigma_\omega \nu_t) \partial_j \omega \right] + C D_{k\omega}, \tag{5}$$

in which the production of turbulent kinetic energy is augmented by b_{ij}^A and bounded following Menter's limiter [25]

$$P_k = \min \left(-2k(b_{ij}^o + b_{ij}^A) \partial_j U_i, \ 10\beta^* \omega k \right). \tag{6}$$

The corresponding eddy viscosity is $\nu_t = \frac{a_1 k}{\max(a_1\omega, SF_2)}$. The other standard terms of k-ω SST read

$$CD_{k\omega} = \max\left(2\sigma_{\omega 2}\frac{1}{\omega}(\partial_i k)(\partial_i \omega),\, 10^{-10}\right),$$

$$F_1 = \tanh\left[\left(\min\left[\max\left(\frac{\sqrt{k}}{\beta^* \omega y},\frac{500\nu}{y^2\omega}\right),\frac{4\sigma_{\omega 2}k}{CD_{k\omega}y^2}\right]\right)^4\right],$$

$$F_2 = \tanh\left[\left(\max\left(\frac{2\sqrt{k}}{\beta^* \omega y},\frac{500\nu}{y^2\omega}\right)\right)^2\right],$$

$$\Phi = F_1\Phi_1 + (1-F_1)\Phi_2, \qquad (7)$$

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

$$\alpha = (5/9, 0.44),\ \beta = (3/40, 0.0828),\ \sigma_k = (0.85, 1.0),\ \sigma_\omega = (0.5, 0.856). \qquad (8)$$

The remaining terms are $\beta^* = 0.09$, $a_1 = 0.31$ and $S = \sqrt{2S_{ij}S_{ij}}$. For the iterative computation of the frozen ω the variables U_i, k and b_{ij} are kept frozen in Eqs. 4 and 5. At each iteration the term R is computed as the residual of Eq. 4 and fed back into Eq. 5. The equations are discretized using linear upwinding for the divergence terms and 2nd order central differencing for diffusion. The fields predicted by a baseline k-ω SST were used as initial condition for ω. For the cases studied the solver reaches convergence after a few hundred iterations. In order to validate that the resulting fields compensate the model-form error, b_{ij}^Δ and R are added as static fields to a modified OpenFOAM solver [26] and a CFD simulation is performed starting from the baseline solution for the flow configurations described in Section 4, for which high-quality data is available. The mean-squared error between the high-fidelity data and the reconstructed velocity U_i as well as the Reynolds-stress τ_{ij} is low, see Table 1. Also the stream-wise velocity profiles shown in Fig. 2 demonstrate that the high-fidelity mean-flow data is essentially reproduced given b_{ij}^Δ and R.

In [6, 8, 27] an inversion procedure was introduced to infer correction factors such as R from data. For this method the data can be scarce, e.g. single profiles of experiments. However, this method is built upon a computationally-intensive optimisation problem, whereas k-corrective-frozen-RANS only requires a single equation to be solved. This makes k-corrective-frozen-RANS a cost-efficient way to extract the model-form error, if full-field data of high-fidelity simulations is available.

Table 1 Mean-squared error ϵ of reconstructed velocity U_i and Reynolds-stress τ_{ij} for different test cases with b_{ij}^Δ and R added as static fields to the solver

Case	$\epsilon(U_i) \cdot 10^{-5}$	$\epsilon(U_i)/\epsilon(U_i^o)$	$\epsilon(\tau_{ij}) \cdot 10^{-6}$	$\epsilon(\tau_{ij})/\epsilon(\tau_{ij}^o)$
PH$_{10595}$	1.74	0.00165	36.7	0.1495
CD$_{12600}$	31.4	0.0229	7.21	0.4781
CBFS$_{13700}$	59.6	0.22703	1.34	0.4949

Normalisation with ϵ of the baseline k-ω SST results U_i^o and τ_{ij}^o. Description of cases in Section 4

(a) PH_{10595}


(b) CD_{12600}

(c) $CBFS_{13700}$
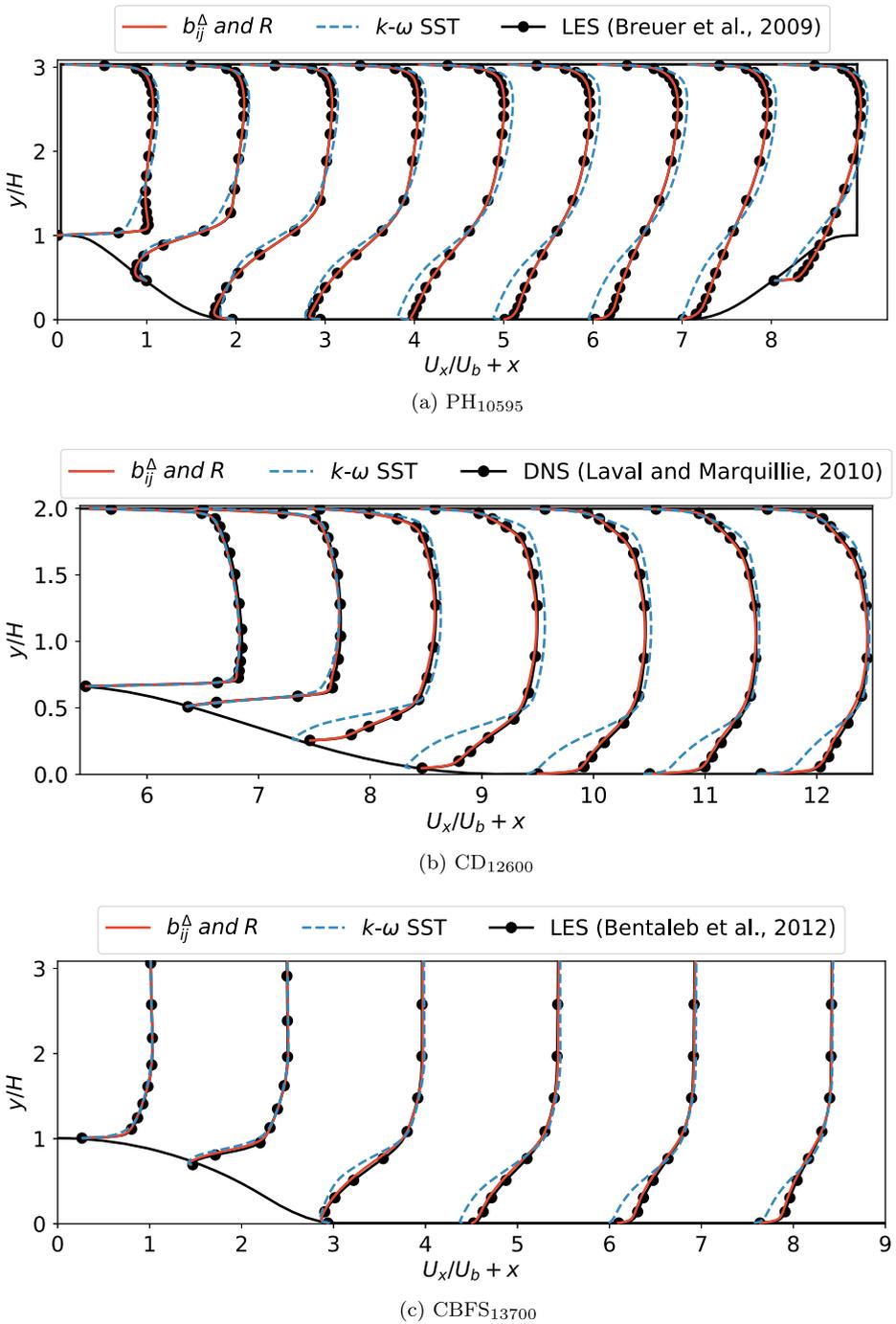
Fig. 2 Stream-wise velocity component for propagated model-form error acquired using k-corrective-frozen-RANS

2.2 Nonlinear eddy-viscosity models for b_{ij}^{Δ} and R

In order to discover corrections for the model-form error b_{ij}^{Δ} and R, we need to decide on a modelling ansatz. Within this mathematical framework the symbolic regression targets to find specific expressions as corrections models. In [13], a nonlinear generalisation of the linear eddy viscosity concept was proposed. This concept has been used in several works on data-driven turbulence modelling [2, 3]. The fundamental assumption is made that the anisotropy of the Reynolds-stress b_{ij} not only depends on the strain rate tensor $S_{ij} = \tau \frac{1}{2}(\partial_j U_i + \partial_i U_j)$ but also on the rotation rate tensor $\Omega_{ij} = \tau \frac{1}{2}(\partial_j U_i - \partial_i U_j)$ with the timescale $\tau = 1/\omega$. The Cayley-Hamilton theorem then dictates that the most general form of the anisotropic part of the Reynolds-stress can be expressed as

$$b_{ij}(S_{ij}, \Omega_{ij}) = \sum_{n=1}^{N} T_{ij}^{(n)} \alpha_n (I_1, ..., I_5), \tag{9}$$

with ten nonlinear base tensors $T_{ij}^{(n)}$ and five corresponding invariants I_m. In the following, we only consider two-dimensional flow cases, for which the first three base tensors form a linear independent basis and only the first two invariants are nonzero [28]. Our set of base tensors and invariants reads

$$T_{ij}^{(1)} = S_{ij}, \ T_{ij}^{(2)} = S_{ik}\Omega_{kj} - \Omega_{ik}S_{kj},$$
$$T_{ij}^{(3)} = S_{ik}S_{kj} - \frac{1}{3}\delta_{ij}S_{mn}S_{nm} \tag{10}$$
$$I_1 = S_{mn}S_{nm}, \ I_2 = \Omega_{mn}\Omega_{nm}. \tag{11}$$

Using this set for Eq. 9 we have an ansatz, which only requires functional expressions for the coefficients α_n, to model b_{ij}^{Δ}. However, computing b_{ij}^{Δ} using Eq. 3 requires a correct k as discussed in Section 2.1. This aspect is taken into account in the modelling ansatz for R, for which we take a closer look at the eddy viscosity concept.

Both linear and nonlinear eddy viscosity models provide expressions for the anisotropy b_{ij} based on a local relation between stress and strain. Due to the restriction of this local closure only the normal stresses $\frac{2}{3}k\delta_{ij}$ can account for nonlocal effects by transport equations for the turbulent quantities using convection and diffusion terms [15, 29]. The term R provides local information to correct the transport equations. Depending on the local sign of R it either increases or decreases the net production P_k locally. Hence, it acts as an additional production or dissipation term, which can overcome the error in k. We model it in a similar way to the turbulent production

$$R = 2kb_{ij}^{R}\partial_j U_i, \tag{12}$$

which has the additional benefit that we can also use the framework of nonlinear eddy viscosity models to model R. Given the polynomial model (9) and the set of base tensors (10) and invariants (11) we are now left with the task of providing suitable expressions for $\alpha_n(I_1, I_2)$ for $n = 1, ..., 3$ to overcome the model-form error.

In [12] models identified using genetic programming were modified such that any additional contribution of the first base tensor $T_{ij}^{(1)}$ in Eq. 9 was added with a positive sign for the computation of P_k. This ad-hoc correction was established based on physical reasoning to avoid very low production close to walls and led to significantly improved predictions. However, in contrast to [12] we have extracted two target terms b_{ij}^{Δ} and R using k-corrective-frozen-RANS, which also make it possible to systematically study (i) how to

obtain corrections models for each target individually and (ii) their combined effect on the predictions. Treating the two targets separately has the disadvantage that energy is not conserved, because Eq. 12 has no corresponding part in the momentum equation [30–32]. Also in literature several classical [33, 34] as well as more recent data-driven [6, 12] approaches exist, which violate conservation of energy. It has been shown that overcoming this inconsistency improves the predictive performance of the models [31]. Thus, a natural next step would be to combine both targets in order to find a single model accounting for the sources of model-form error on the level of the constitutive relation as well as within the turbulent transport equations. This requires a multi-objective version of the deterministic symbolic regression detailed below, which is beyond the scope of this paper.

3 Model Discovery Methodology

Deterministic symbolic regression constructs a large library of nonlinear candidate functions to regress data. It identifies the relevant candidates by adopting a sparsity constraint. Two fundamental methods have been proposed: Sparse identification of nonlinear dynamics (SINDy) [20, 35] and fast function extraction (FFX) [36]. Both methods were applied in several areas of physical modelling. In the following, we introduce the steps of the model discovery methodology SpaRTA based on FFX, for which a library is constructed using a set of raw input variables and mathematical operations. The model selection uses elastic net regression. Finally, for the inference of the model coefficients the stability requirements of a CFD solver are considered. An overview of SpaRTA is given in Fig. 1.

3.1 Building a library of candidate functions

The deterministic symbolic regression requires a library of candidate functions, from which a model is deduced by building a linear combination of the candidates. Hence, the library is an essential element of the entire methodology and needs to accommodate relevant candidates explaining the data. We rely on the nonlinear eddy viscosity concept and aim to find models for α_n in Eq. 9 given as primitive input features the invariants I_1 and I_2. For the present work we focus on a library, in which the primitive input features are squared and the resulting candidates are multiplied by each other leading to a maximum degree of 6. In addition to the two invariants we also include a constant function c to the set of raw input features. The resulting vector $\mathcal{B}$ reads

$$\mathcal{B} = \Big[1, I_1, I_2, I_1^2, I_2^2, I_1^2 I_2^3, I_1^4 I_2^2, I_1 I_2^2, I_1 I_2^3, $$
$$I_1 I_2^4, I_1^3 I_2, I_1^2 I_2^4, I_1^2 I_2, I_1 I_2, I_1^3 I_2^2, I_1^2 I_2^2 \Big]^T \tag{13}$$

with the cardinality of $\mathcal{B}$, $|\mathcal{B}| = 16$.

For the library to regress models for b_{ij}^Δ each function of $\mathcal{B}$ is multiplied with each base tensor $T_{ij}^{(n)}$, leading to the library of tensorial candidate functions

$$\mathcal{C}_{b_{ij}^\Delta} = \Big[T_{ij}^{(1)}, T_{ij}^{(2)}, \ldots, I_1^2 I_2^2 T_{ij}^{(3)} \Big]^T. \tag{14}$$

In order to regress models for R the double dot product of each function in $\mathcal{C}_{b_{ij}^\Delta}$ with the mean velocity gradient tensor $\partial_j U_i$ is computed, leading to

$$\mathcal{C}_R = \left[T_{ij}^{(1)} \partial_j U_i, \dots, I_1^2 I_2^2 T_{ij}^{(3)} \partial_j U_i \right]^T. \tag{15}$$

The two libraries $\mathcal{C}_{b_{ij}^\Delta}$ and $\mathcal{C}_R$ are evaluated given the high-fidelity validation data for each test case and stored column-wise in matrices $C_{b_{ij}^\Delta}$ and C_R. In order to avoid large values, a candidate function is discarded from the library, if it contains values with a magnitude larger than 10^5. Finally, the target data b_{ij}^Δ and R are stacked to vectors.

3.2 Model selection using sparsity-promoting regression

Given the above defined libraries the task is to form a linear model to regress the target data $\Delta = b^\Delta$ or R by finding the coefficient vector $\boldsymbol{\Theta}$

$$\Delta = C_\Delta \boldsymbol{\Theta}, \tag{16}$$

which represents a large, overdetermined system of equations. When using ordinary least-squares regression a dense coefficient vector $\boldsymbol{\Theta}$ is obtained, resulting in overly complex models, which are potentially overfitting the data given the large libraries Eqs. 14 and 15. Due to multi-collinearity between the candidates, C_Δ can be ill-conditioned, so that the coefficients may also display large differences in magnitude expressed in a large l_1-norm of $\boldsymbol{\Theta}$. Such models are unsuitable to be implemented in a CFD solver as they increase the numerical stiffness of the problem and impede convergence of the solution.

Following the idea of parsimonious models we constrain the search to models which optimally balance error and complexity and are not overfitting the data [35]. In principle, given a library a combinatoric study can be carried out, by performing an ordinary least-squares regression for each possible subset of candidates. Starting from each single candidate function individually, proceeding with all possible pairs up to more complex combinations. As the number of possible models grows exponentially with the number of candidates $I = 2^{|\mathcal{C}_\Delta|} - 2$ this approach becomes already infeasible for the simple libraries Eqs. 14 and 15 with $|\mathcal{C}_\Delta| \approx 48$.

Hence, we follow [35, 36] and engage sparsity-promoting regularisation of the underlying least-squares optimisation problem. The model-discovery procedure is divided into two parts: (i) model selection and (ii) model inference, see Fig. 1. For the first step, the model selection, we use the elastic net formulation

$$\begin{aligned} \boldsymbol{\Theta} = \underset{\hat{\boldsymbol{\Theta}}}{\arg\min} \ & \left\| C_\Delta \hat{\boldsymbol{\Theta}} - \Delta \right\|_2^2 + \lambda \rho \left\| \hat{\boldsymbol{\Theta}} \right\|_1 \\ & + 0.5\lambda(1 - \rho) \left\| \hat{\boldsymbol{\Theta}} \right\|_2^2, \end{aligned} \tag{17}$$

which blends the l_1- and l_2-norm regularisation given the mixing parameter $\rho \in [0, 1]$ and the regularisation weight λ, to promote the sparsity of $\boldsymbol{\Theta}$ [36, 37]. On its own, the l_1-norm, known as Lasso-regression, promotes sparsity by allowing only a few nonzero coefficients while shrinking the rest to zero. The l_2-norm, known as Ridge-regression, enforces relatively small coefficients without setting them to zero, but is able to identify also correlated candidate functions instead of picking a single one. By combining both methods, the elastic net can find sparse models with a good predictive performance. Besides the mixing parameter, also the regularisation parameter λ shapes the form of the model: For a very large λ

the vector $\boldsymbol{\Theta}$ will only contain zeros independent of ρ. The amount of nonzero coefficients increases for smaller λ values making the discovery of sparse models possible.

Given the elastic net regularisation method we need to specify suitable combinations of the weight λ and type of the regularisation ρ, for which the optimisation problem Eq. 17 is solved. Most commonly the optimal (λ, ρ) combination is found based on a strategy to avoid overfitting of the resulting models, e.g. using cross-validation [35], for which the data is split into a training and a test set. While the optimisation problem given a grid $(\boldsymbol{\lambda}, \boldsymbol{\rho})$ is solved on the former, only the model with the best performance evaluated on the latter survives. For the purpose of CFD a true validation of the models can only be performed once they are implemented in a solver and applied to a test case. In order to not overcharge the role of the training data from k-corrective-frozen-RANS at this stage of the methodology, we select a wide spectrum of models varying in accuracy and complexity using Eq. 17 instead of a single one. The validation task will be performed later using a CFD solver. Following [36] we use

$$\boldsymbol{\rho} = [0.01, 0.1, 0.2, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0]^T, \tag{18}$$

which ensures that we cover a substantial range of different regularisation types. The upper limit of the regularisation weight is defined as $\lambda_{\max} = \max(|\boldsymbol{C}_\Delta^T \boldsymbol{\Delta}|)/(K\rho)$, because for any $\lambda > \lambda_{\max}$ all elements in $\boldsymbol{\Theta}$ will be equal to zero. The entire vector

$$\boldsymbol{\lambda} = [\lambda_0, ..., \lambda_{\max}]^T \tag{19}$$

is defined of having 100 entries between $\lambda_0 = \xi \lambda_{\max}$ with $\xi = 10^{-3}$ uniformly spaced using a log-scale as defined in [36]. This provides a search space $(\boldsymbol{\lambda}, \boldsymbol{\rho})$, the elastic net, which is large enough and has an appropriate resolution. At each grid point (λ_i, ρ_j) a vector $\boldsymbol{\Theta}_\Delta^{(i,j)}$ as a solution of Eq. 17 is found using the coordinate descent algorithm. The duration for the model selection step given the number of data points $K \sim 15000$ is of the order of a minute on a standard consumer laptop.

Solving Eq. 17 for different (λ_i, ρ_j) might produce $\boldsymbol{\Theta}_\Delta^{(i,j)}$ with the same abstract model form $\bar{\boldsymbol{\Theta}}$, which means that the same entries are equal to zero. As the specific values of the coefficients will be defined in the next step, the selection step of SpaRTA concludes with filtering out the set of D unique abstract model forms $\mathcal{D}_\Delta = \left\{ \bar{\boldsymbol{\Theta}}_\Delta^d \,|\, d = 1, ..., D \right\}$.

3.3 Model inference for CFD

The abstract models $\mathcal{D}_\Delta$ are found using standardised candidates, because the relevance of each candidate should not be determined by its magnitude during the model selection step. With the aim of defining a model with the correct units, we need to perform an additional regression using the unstandardised candidate functions for each subset determined by the abstract model forms in $\mathcal{D}_\Delta$, which is the purpose of the model inference step outlined in the following.

In [35, 38, 39] this was done using ordinary least-squares regression for problems in the domains of dynamical systems and biological networks. As mentioned above, the ability of the CFD solver, in which the models will be implemented, to produce a converged solution is sensitive to large coefficients, which has been reported in [11, 12, 22]. We take this additional constraint into account by performing a Ridge regression

$$\boldsymbol{\Theta}_\Delta^{s,d} = \underset{\hat{\boldsymbol{\Theta}}_\Delta^{s,d}}{\arg\min} \left\| \boldsymbol{C}_\Delta^s \hat{\boldsymbol{\Theta}}_\Delta^{s,d} - \boldsymbol{\Delta} \right\|_2^2 + \lambda_r \left\| \hat{\boldsymbol{\Theta}}_\Delta^{s,d} \right\|_2^2, \tag{20}$$

in which λ_r is the Tikhonov-regularisation parameter. The index s denotes the submatrix of C_Δ and the subvector of $\boldsymbol{\Theta}_\Delta^d$ consisting of the selected columns or elements respectively as defined in $\mathcal{D}_\Delta$. The elements of $\boldsymbol{\Theta}_\Delta^d$ associated with the inactive candidates are zero and are not modified during this step.

By using the l_2-norm regularisation the magnitude of the nonzero coefficients is shrunk [35, 40]. In general, low values for λ_r reduce the bias introduced through regularisation, but lead to larger coefficient values, and vice versa. Since shrinkage of the coefficients also reduces the influence of candidate functions with a lower magnitude compared to others, we need to find a trade-off between error of the model on the target data $\boldsymbol{\Delta}$ and the likelihood that the model will deliver converged solutions when used in a CFD solver. The problem of finding such an optimum is that the latter aspect can only be answered retrospectively. Recently, this problem has been addressed in [41] by embedding CFD simulations in the search for correction models guided by genetic programming. While this increases the costs of the model search drastically, it also significantly increases the chance of delivering models with better convergence properties. Even though this procedure provides a strong indication, the identified models are also not guaranteed to converge *a priori* for any other test case outside the training set. Via testing using the cases in Section 4, we have identified $0.01 < \lambda_r < 0.1$ able to deliver coefficients in a range balancing the error on the target data Δ and the likelihood to produce converged CFD solutions. Our efforts are based on an empirical observation, but do not guarantee a well-behaving numerical setup under all conditions. We have identified corrections of b_{ij}^Δ as the only contribution which can do harm to the convergence properties for the given test cases. However, at this stage of the process we exclude models, if they are not converging on a given test case.

Finally, the resulting coefficient vector $\boldsymbol{\Theta}_\Delta^d$ is used to retrieve the symbolic expression of the models by a dot product with the library of candidate functions $\mathcal{C}_\Delta$ in Eqs. 14 and 15

$$M_\Delta^d := \mathcal{C}_\Delta^T \boldsymbol{\Theta}_\Delta^d, \tag{21}$$

which are implemented in the open-source finite-volume code OpenFOAM [26]. The divergence terms of the equations are discretised with linear upwinding and turbulent diffusion with 2nd order central differencing. In summary, the model discovery step of SpaRTA selects models utilising elastic net regression in Eq. 17 and further infers the coefficients of the selected models in Eq. 20. The latter process is guided by the aim to discover models complying with the restrictions of a CFD solver.

4 Test Cases and High-Fidelity Data

In order to apply SpaRTA we need full-field data of U_i, k and τ_{ij}, which we take from LES and DNS studies conducted by other researchers. We have selected three test cases of separating flows over curved surfaces in two-dimensions with similar Reynolds-numbers. For each case fine meshes are selected, which ensure that the discretisation error is much smaller compared to the error due to turbulence modelling.

Periodic hills (PH) Flow over a series of hills in a channel. Initially proposed by [42] this case has been studied both experimentally as well as numerically in detail. We use LES data from [43] for $Re = 10595$ (PH_{10595}) to apply SpaRTA and test the performance of the resulting models. In addition, we also use experimental data from [44] at a much larger $Re = 37000$ (PH_{37000}) in order to test the models outside the range of the training data. The numerical mesh consists of 120×130 cells. Cyclic boundary conditions are used at the

inlet and outlet. The flow is driven by a volume forcing defined to produce a constant bulk velocity.

Converging-diverging channel (CD) A DNS study of the flow within a channel, in which an asymmetric bump is placed, exposed to an adverse pressure gradient was performed by [45] for $Re = 12600$ (CD$_{12600}$). The flow shows a small separation bubble on the lee-side of the bump, which is challenging for RANS to predict. The numerical mesh consists of 140×100 cells. The inlet profile was obtained from a channel-flow simulation at equivalent Re.

Curved backward-facing step (CBFS) In [46] a LES simulation of a flow over a gently-curved backward-facing step was performed at $Re = 13700$ (CBFS$_{13700}$). Similar to PH also for this flow the mean effect of separation and reattachment dynamics is the objective. The numerical mesh consists of 140×150 cells. The inlet was obtained from a fully-developed boundary layer simulation.

Despite the simple geometries, the mean effect of the separation and reattachment dynamics of a flow on a curved surface is a challenging problem for steady-RANS approaches. Especially, PH serves as an important testbed for classical and data-driven approaches for turbulence modelling, e.g. [2, 47], but also the other two have been introduced with the purpose of closure investigation.

5 Discovered Models and Cross-Validation Using CFD

In the following, the method SpaRTA introduced in Section 3 is applied to the three test cases of Section 4. The models resulting from the model-discovery are presented and their mean-squared error on the training data is evaluated. In order to identify the models with the best predictive capabilities, we carry out cross-validation of the resulting models using CFD [40]: Models identified given training data of one case are used for CFD simulations of the remaining two case. For each case a single model is chosen as the best-performing one. Finally, the three resulting models are tested in a *true* prediction for the flow over periodic hills at $Re = 37000$.

The goal of the model-discovery is to identify an ensemble of diverse models with small coefficients, varying in model-structure (complexity) and accuracy. Such an ensemble is better-suited for the cross-validation on unseen test cases, than a selection of the best models given only the training data. The sparse-regression for b_{ij}^A applied to the three test cases resulted in 7 distinct models for PH$_{10595}$ and 8 for CD$_{12600}$ and CBFS$_{13700}$. For R the model discovery resulted in 1, 3 and 4 distinct model forms for CBFS$_{13700}$, PH$_{10595}$ and CD$_{12600}$ respectively. We identify $T_{ij}^{(1)}$, $I_1 T_{ij}^{(1)}$ and $I_2 T_{ij}^{(1)}$ as the relevant candidates to regress R, and models combining all three give the lowest error per test case. The ensembles of discovered models are shown in Figs. 3 and 4. Each row represents a model structure, the colours indicate the nonzero value of the coefficient corresponding to the chosen candidate function. If a candidate function is not selected the corresponding field is left blank. The result is a hierarchical spectrum of models regressing the training data varying in complexity and error.

Cross-validation tests how well models identified on training data perform on unseen test cases [40]. This assessment allows to determine the best-predictive models from a set. As stated above, the role of the frozen training data should not be overcharged, so that we cross-validate using CFD. By doing so, we can assess the validity of SpaRTA as a tool
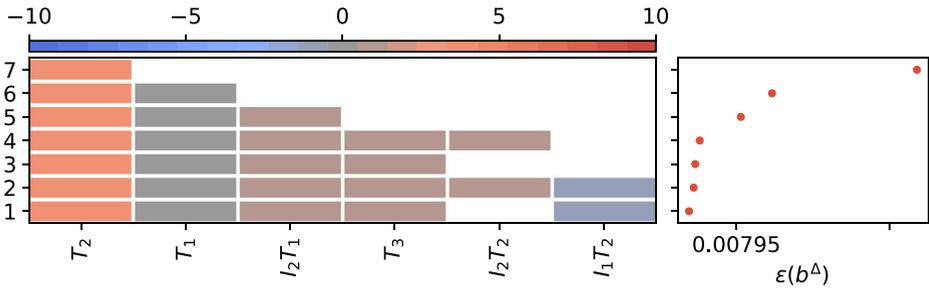
(a) PH$_{10595}$
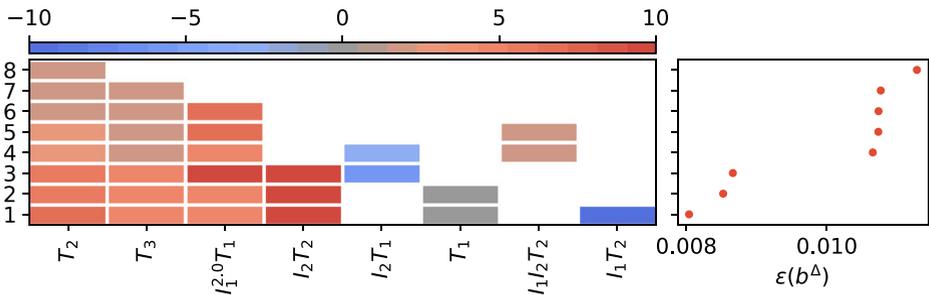
(b) CD$_{12600}$
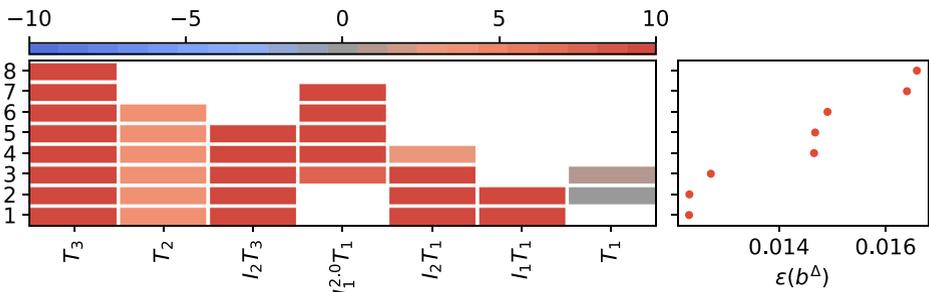
(c) CBFS$_{13700}$

Fig. 3 Model-structure of all discovered models using SpaRTA and mean-squared error on training data. The matrix (l.) shows the values of the active (coloured) candidate functions (x-axis) for each model M_i with model index i (y-axis). The mean-squared error between the frozen data b_{ij}^Δ and the model is also shown (r.)

for model discovery as well as the predictive performance of the identified models outside of their training set. The found correction models regress b_{ij}^Δ and R individually and can also be applied individually for predictions when implemented in the solver, i.e. a model correcting b_{ij}^Δ can be used without a correction of R and vice-versa. We can also study their combined effect. For the cross-validation in the following, all models are applied to the three test cases, which requires 61, 48 and 75 simulations for the cases PH$_{10595}$, CD$_{12600}$ and CBFS$_{13700}$ respectively.
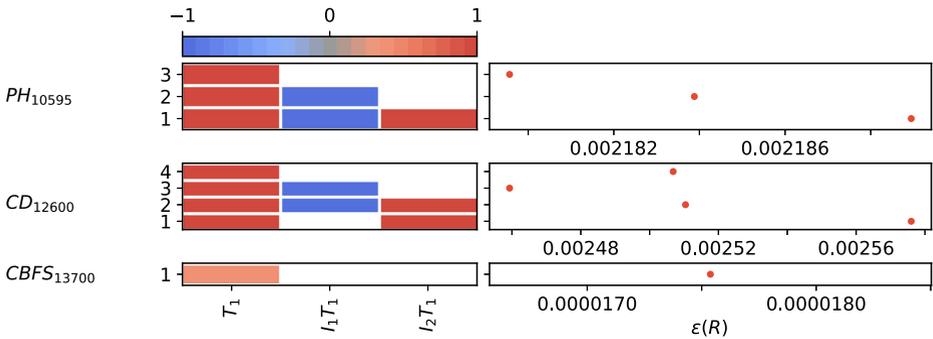
Fig. 4 Selected models and their mean-squared error evaluated on the training data R

In Fig. 5 the mean-squared error of each model on the velocity field $\epsilon(U)$ normalised with the mean-squared error of the baseline $\epsilon(U^o)$ is shown. The type of model, whether it is providing a correction both for b_{ij}^{Δ} and R or for each one individually, and from which training data it originated, is emphasised by a unique marker form and color combination. Most of the models show a good or even substantial improvement over the baseline. But, for the set of models, only providing a correction for b_{ij}^{Δ}, not all lead to an improvement of the resulting velocity field. In contrast to that, if only a correction for R is deployed, the result is a consistent, substantial improvement across all test cases. Using both a model for b_{ij}^{Δ} and R only provides a minor additional improvement for some cases. For the test case $CBFS_{13700}$ using both corrections leads to a detrimental effect, the error increases for models identified on PH_{10595} and CD_{12600} data. Surprisingly, the best model per test case is not always identified on the associated training data. While this expectation holds for the cases $CBFS_{13700}$ and CD_{12600} it is not true for PH_{10595}, for which the other two training sets deliver significantly better performing models. In general, the data of CD_{12600} and $CBFS_{13700}$ provide models, which are well performing on all test cases presented.
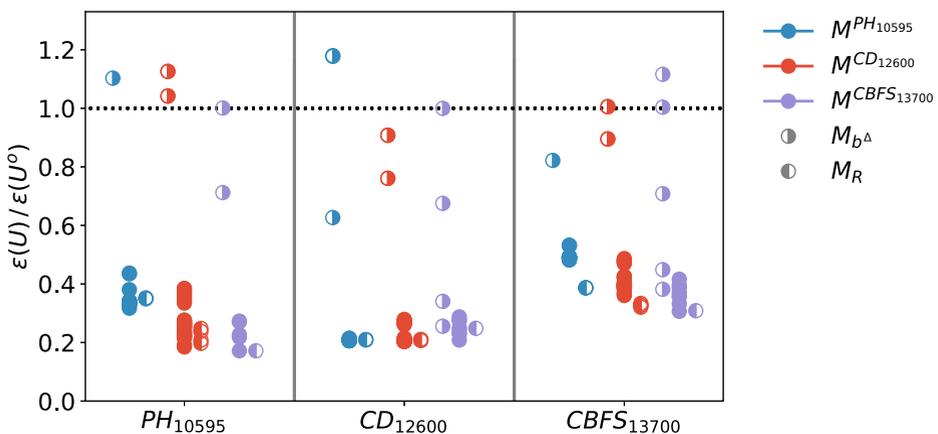

Fig. 5 Mean-squared error of velocity vector of each correction model normalised by the mean-squared error of the baseline k-ω SST. The colour indicates on which high-fidelity data the models have been identified. Full circles represent simulations using both corrections, while left-/right-filled circles represent simulations using only correction for R or b_{ij}^{Δ} respectively
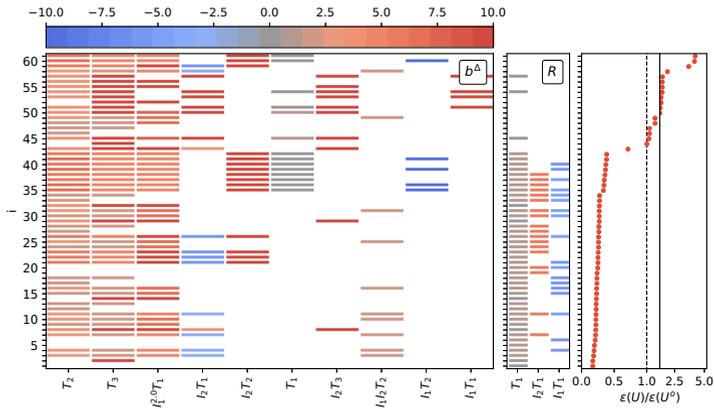
(a) PH$_{10595}$
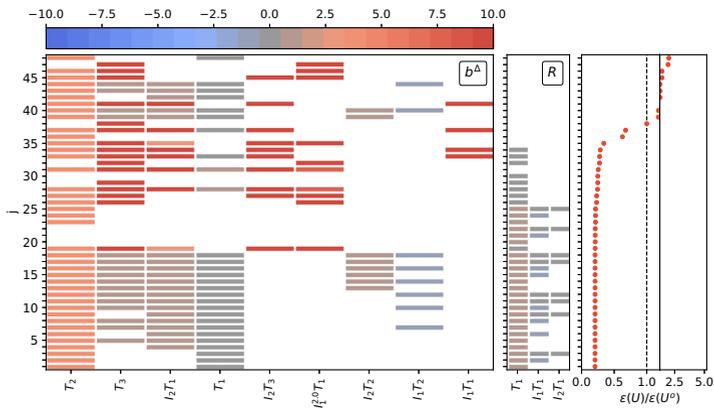
(b) CD$_{12600}$
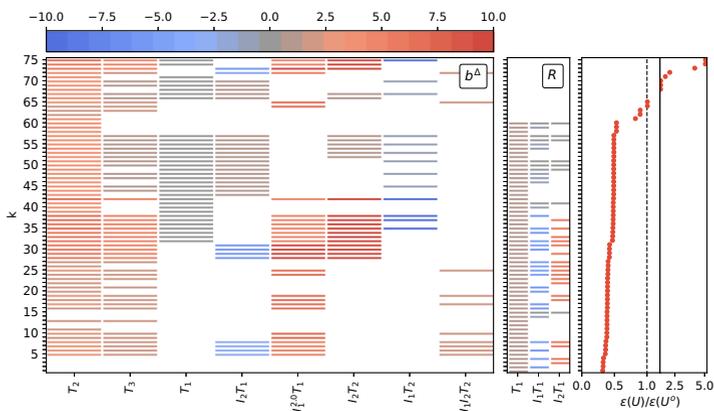
(c) CBFS$_{13700}$

Fig. 6 The two matrices (l.) show the models M_i for b_{ij}^{Δ} and R. The mean-squared error in velocity U normalised by the mean-squared error of the baseline k-ω SST model is also shown (right)

In Fig. 6, both the error and the model structure for the correction of b_{ij}^{Δ} as well as for R is shown. The models are ordered according to the mean-squared error on the stream-wise velocity U. In line of the discussion of Fig. 5 three groups can be identified: a few models, which lead to an increased error compared to the baseline; a small group of models per test case, which are equal or similar to the baseline; and the great majority of models, which result in an improvement. It can be observed how the error in the velocity is significantly reduced once a correction of R is used. The best models correct the velocity up to 5 times better in mean-squared error than the k-ω SST baseline model. This leaves still room for further improvement compared to the error using the frozen data sets, see Table 1. But, especially for case CBFS$_{13700}$ the result is already very close to the possible correction provided by the frozen data at least for U.

Given this cross-validation assessment we select models $\boldsymbol{M}^{(i)} = (M_{b^{\Delta}}^{(i)}, M_R^{(i)})^T$ based on the lowest $\epsilon(U)$ per case

$$
\begin{aligned}
M_{b^{\Delta}}^{(1)} &= 0, \\
M_R^{(1)} &= 0.39 \, T_{ij}^{(1)},
\end{aligned}
\tag{22}
$$

$$
\begin{aligned}
M_{b^{\Delta}}^{(2)} &= 0.1 \, T_{ij}^{(1)} + 4.09 \, T_{ij}^{(2)}, \\
M_R^{(2)} &= 1.39 \, T_{ij}^{(1)},
\end{aligned}
\tag{23}
$$

$$
\begin{aligned}
M_{b^{\Delta}}^{(3)} &= 0 \\
M^{(3)} R &= 0.93 \, T_{ij}^{(1)},
\end{aligned}
\tag{24}
$$

for which further details on the corresponding training data and the rank of the model on each test case are given in Table 2. The models $M^{(1)}$ and $M^{(3)}$ provide very simple correction of R and none for b^{Δ}. Also for $M^{(2)}$ an equivalent model with a similar error using no correction for b^{Δ} could be found, see row number 20 in Fig. 6. Especially model $M^{(3)}$ performs very well both on CBFS$_{13700}$ (rank 1.) and PH$_{10595}$ (5.). While the rank of the others varies more between the test cases, they are still within the set of well-performing models with $\epsilon(U)/\epsilon(U^o) < 0.5$. Their predictions of stream-wise velocity U, k, the Reynolds-stress component τ_{xy} and the skin-friction coefficient C_f are shown in Figs. 7, 8, 9 and 10 for the three test cases. As already stated for the error evaluated on the entire domain discussed above, these three models show an improvement of the spatial distribution of the predicted quantities in comparison to the baseline prediction of k-ω SST. Especially the velocity is well-captured for all three. While k is better identified compared to the baseline, we still observe a discrepancy between the predictions and the data. For PH$_{10595}$ the three models

Table 2 Best-predictive models with rank (index i, j, k in Fig. 6) and normalised error on velocity $\epsilon(U)/\epsilon(U^o)$ for different cases

Model	PH$_{10595}$		CD$_{12600}$		CBFS$_{13700}$	
	index i	$\epsilon(U)/\epsilon(U^o)$	index j	$\epsilon(U)/\epsilon(U^o)$	index k	$\epsilon(U)/\epsilon(U^o)$
$M^{(1)}$	(1.)	0.17166	(30.)	0.24839	–	0.30861
$M^{(2)}$	–	0.32683	(1.)	0.20612	(39.)	0.48244
$M^{(3)}$	(5.)	0.19737	–	0.20975	(1.)	0.32062

(a) PH_{10595}


(b) CD_{12600}

(c) CBFS_{13700}

**Fig. 7** Predicted stream-wise velocity

(a) PH$_{10595}$

(b) CD$_{12600}$

(c) CBFS$_{13700}$
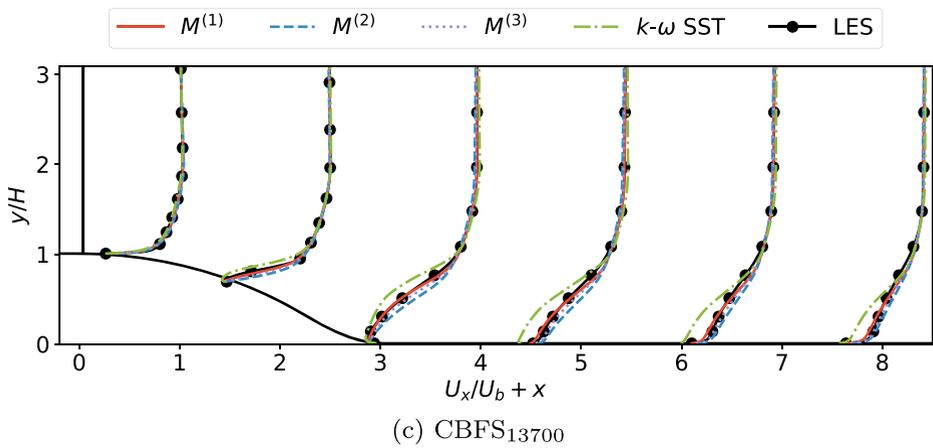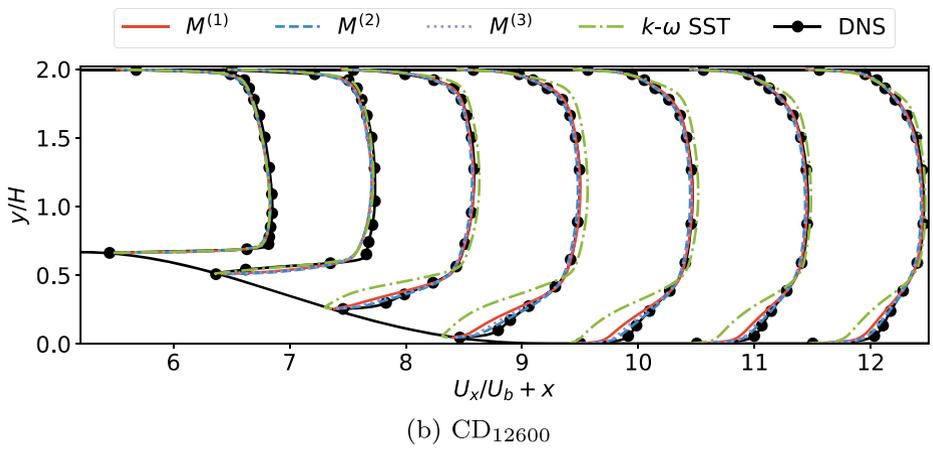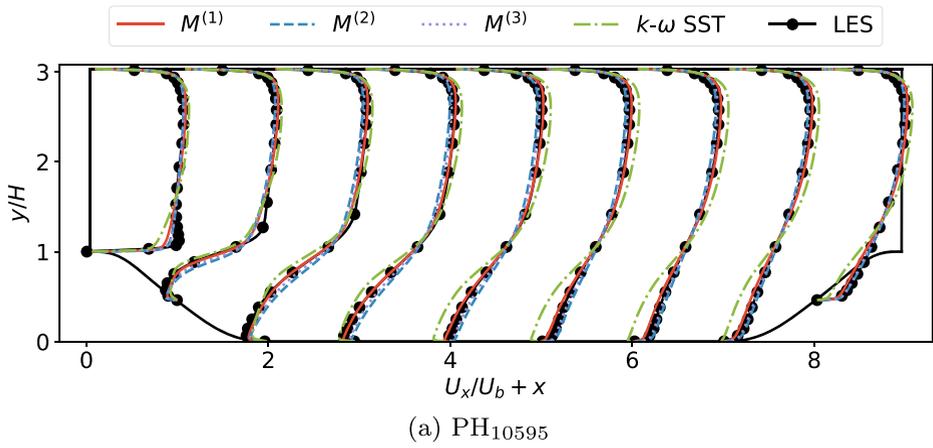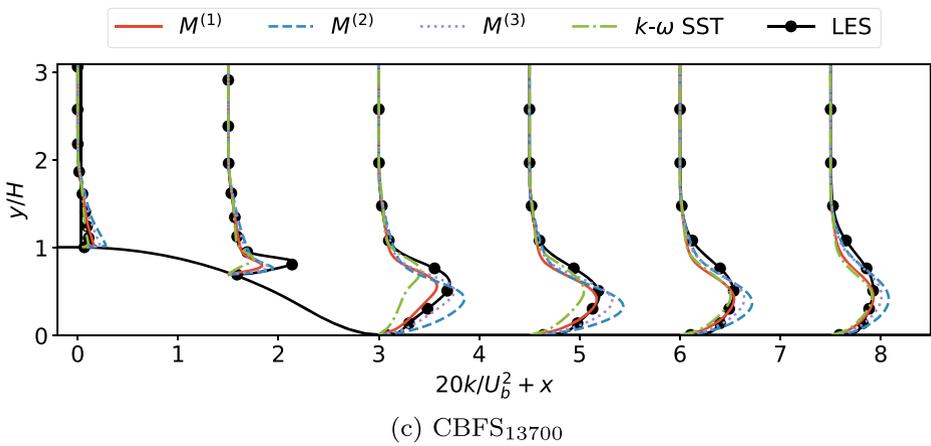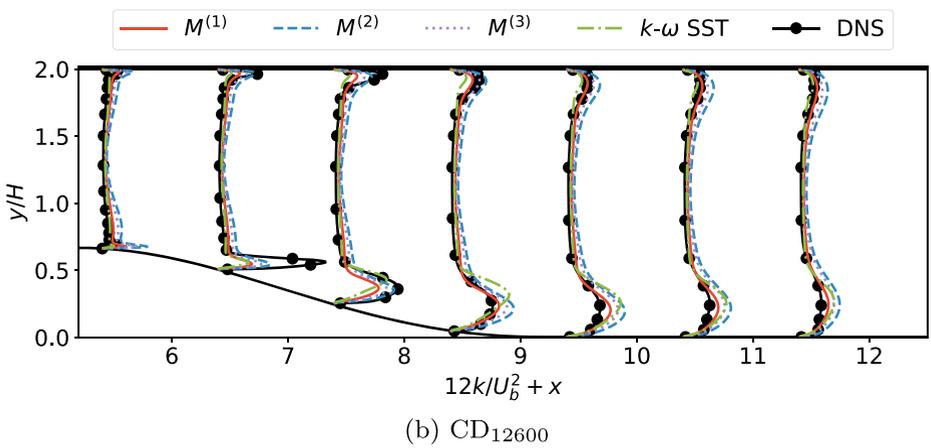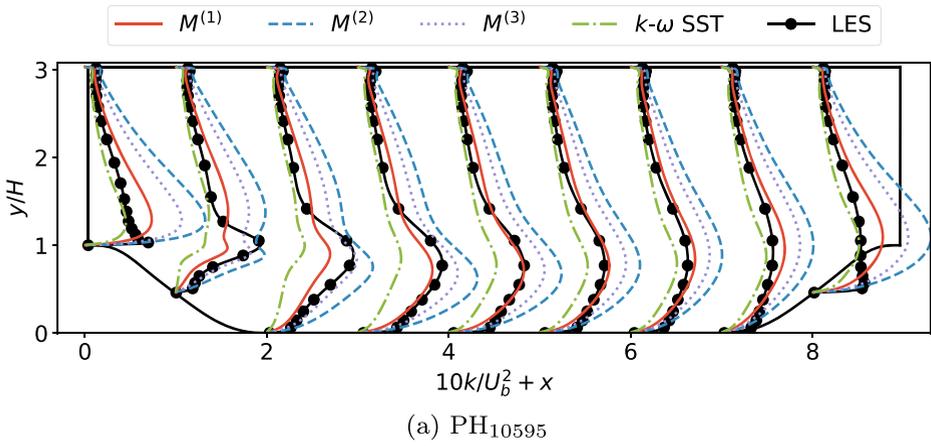
Fig. 8 Predicted turbulent kinetic energy

(a) PH_{10595}

(b) CD_{12600}

(c) CBFS_{13700}
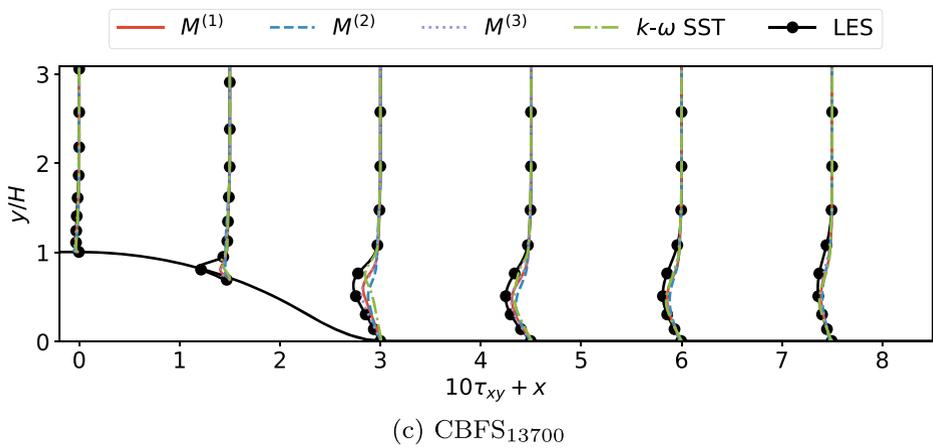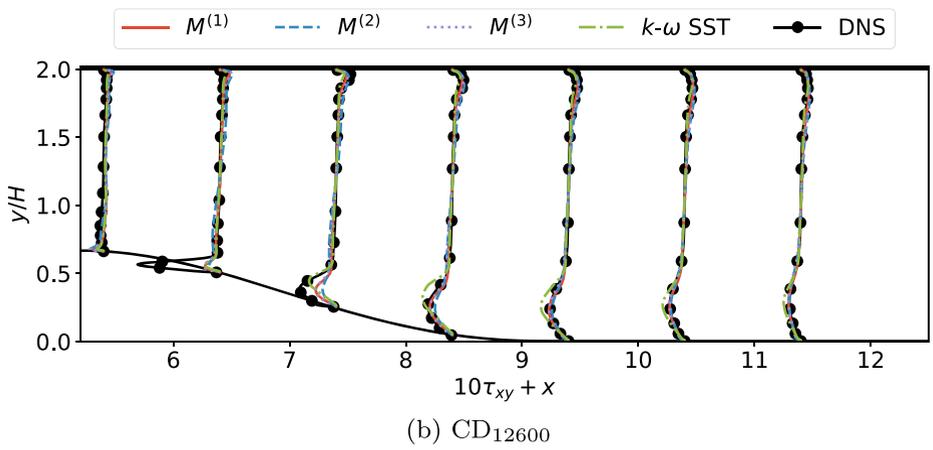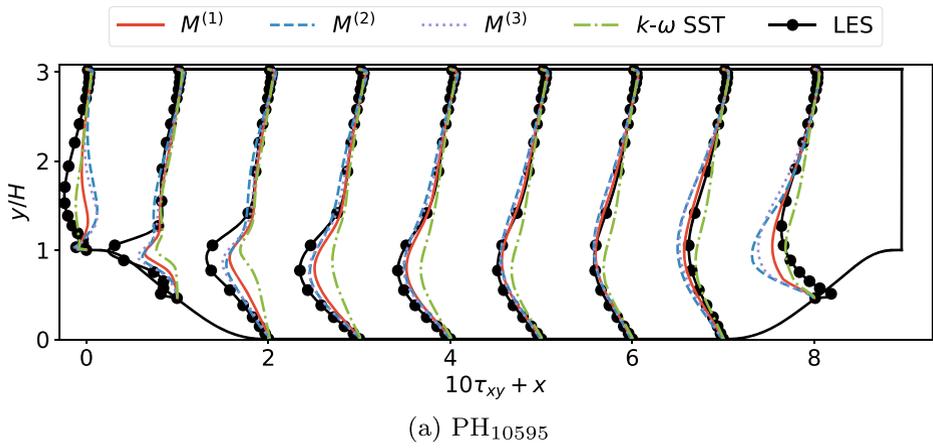
Fig. 9 Predicted shear stress

(a) PH_{10595}

(b) CD_{12600}

(c) $CBFS_{13700}$
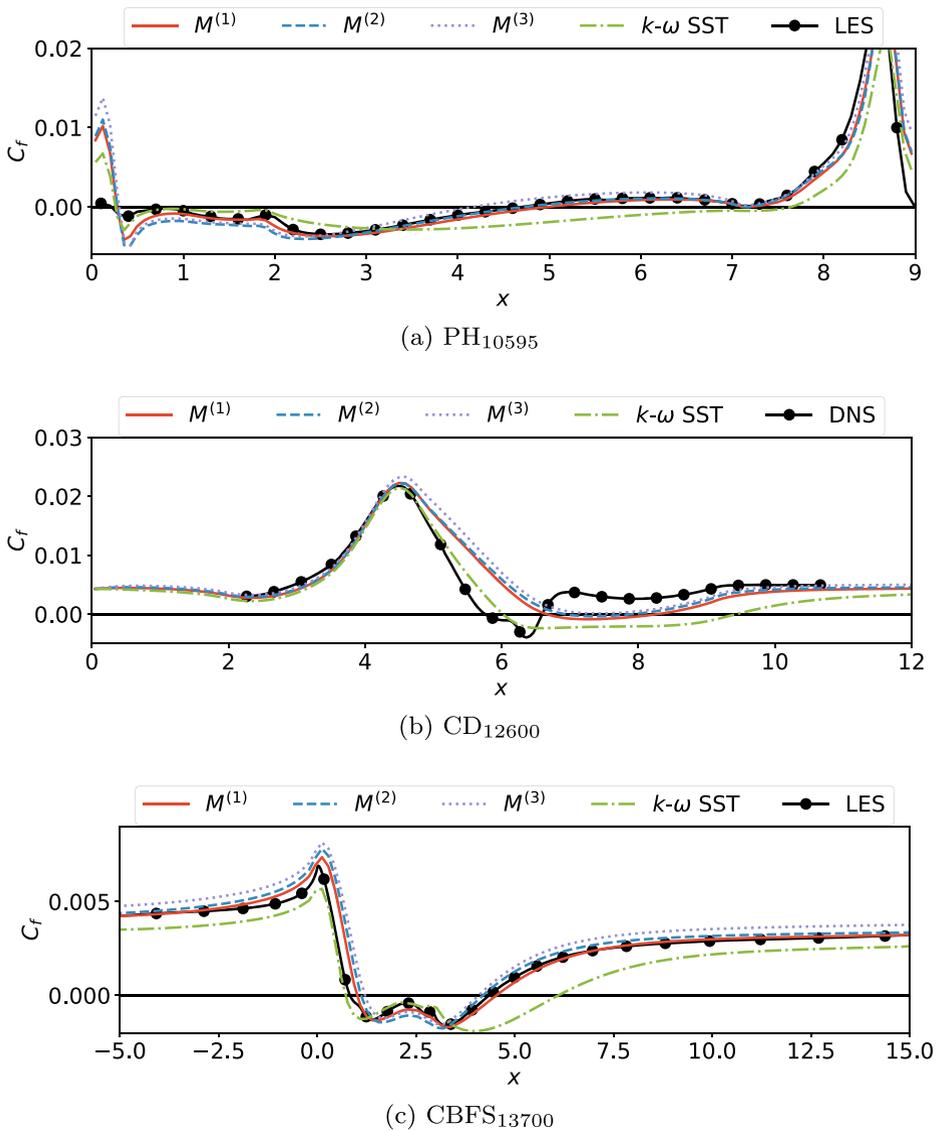
Fig. 10 Predicted skin friction coefficient

do not fit the complex spatial structure especially in the shear-layer, but together encapsulate the data for most of the profiles. For CD_{12600} the models are underestimating k for $x < 7$ and overestimate it further downstream. For $CBFS_{13700}$ the models also underestimate on the curved surface, but fit the data better than the baseline for $3 < x < 5$. The magnitude of the Reynolds-stress component τ_{xy} is underestimated on the curved surfaces of all test cases. For PH_{10595} the models fail to fit the complex spatial structure especially within the separated shear-layer behind the hill and on the hill itself. The skin friction coefficient C_f and the associated separation and reattachment points are better captured compared to the
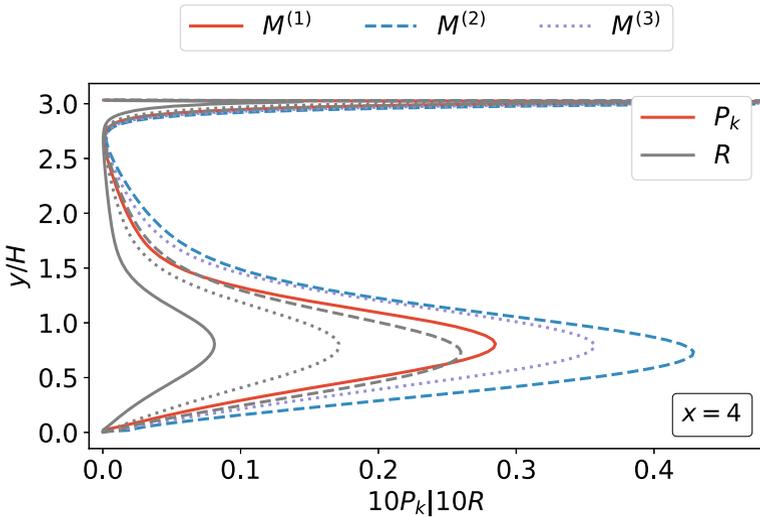
Fig. 11 P_k (coloured) and R (grey) at $x = 4.0$ for flow over periodic hills at $Re = 10595$ using correction models

baseline k-ω SST for PH$_{10595}$ and CBFS$_{13700}$. For CD$_{12600}$, we observe a small recirculation zone as reported in the literature using $M^{(1)}$, but too far down-stream. However, while the baseline k-ω SST drastically over-predicts this zone, $M^{(2)}$ and $M^{(3)}$ ignore it entirely.

Overall, the models agree well with the spatial data, which is in line with the global error on U in Table 2. To correct the velocity prediction sufficiently only a slight modification of the baseline k-ω SST model is necessary. A model for R using a scalar times $T_{ij}^{(1)}$ is sufficient for the given test cases. The resulting magnitudes of R and P_k are shown in
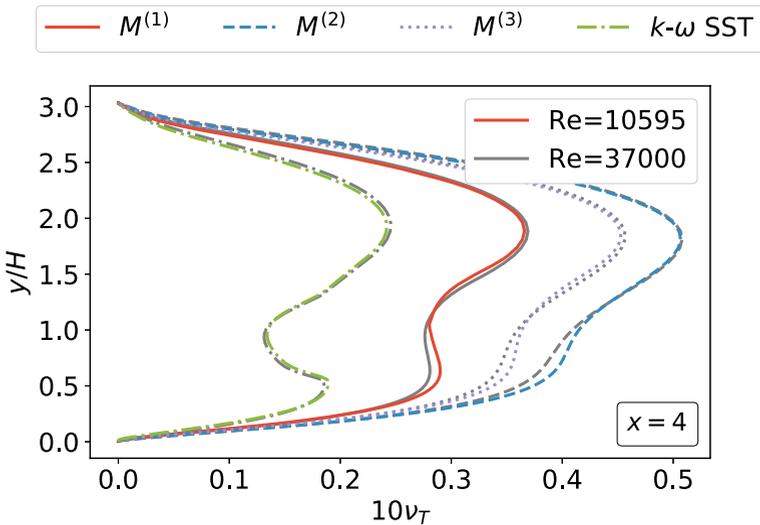

Fig. 12 Eddy viscosity profile at $x = 4.0$ for flow over periodic hills at $Re = 10595$ (coloured) and $Re = 37000$ (grey) using correction models compared to baseline k-ω SST
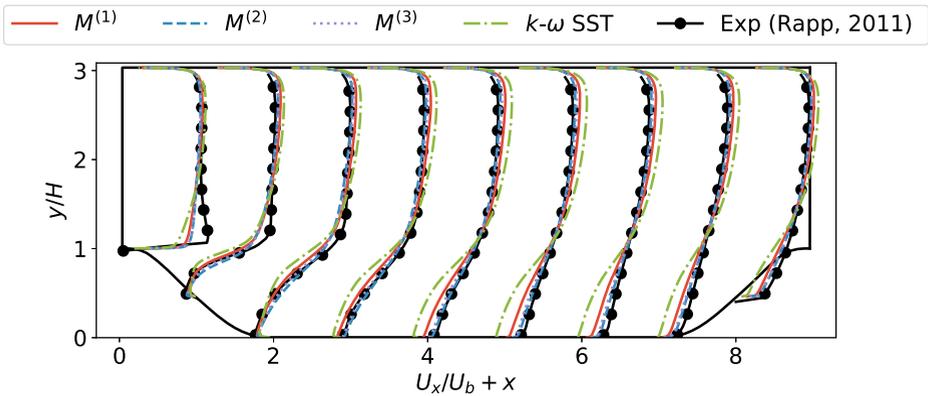
Fig. 13 Predicted stream-wise velocity for flow over periodic hills at $Re = 37000$ using correction models compared to baseline k-ω SST and experimental data of [44]

Fig. 11. The enhanced production P_k due to a positively-signed R induces an increase of eddy viscosity, which makes the corrected model more dissipative compared to the baseline k-ω SST model. This is shown in Fig. 12 for the periodic hills at stream-wise position $x = 4.0$. The result is an increase of shear-stress, which leads to shortening of the recirculation bubble [47]. Consequently, the increase of eddy viscosity follows the magnitude of the coefficients of the models, i.e. $M^{(1)} < M^{(3)} < M^{(2)}$.

In order to test how the models extrapolate to cases of larger Re, we predict the flow over periodic hills at $Re = 37000$, see Fig. 13. Due to an increase of turbulence this case has a significantly shorter recirculation zone. For this *true* prediction throughout the domain the three models improve significantly compared to the baseline. Interestingly, the models $M^{(2)}$ and $M^{(3)}$ are providing a better fit of the data than $M^{(1)}$, which was performing better on the lower Re case. Also for this case, the eddy viscosity is significantly larger compared to the baseline model, see Fig. 12, inducing the same pattern as discussed above for PH_{10595}.

6 Conclusion and Extension

In this work SpaRTA was introduced to discover algebraic models in order to correct the model-form error within the k-ω SST. For this novel machine learning method two additive terms, on the level of the stress-strain relation b_{ij}^Δ and within the turbulent transport equations R, were identified by means of k-corrective-frozen-RANS, for which the governing equations are evaluated given high-fidelity data of three cases of separating flows. It was validated that the computed terms are compensating the model-form error and reproduce the high-fidelity LES or DNS mean-flow data. Hence, k-corrective-frozen-RANS is a cost-efficient way to distill useful information directly from full-field data without the need of an inversion procedure.

Cross-validation of the discovered models using CFD was carried out to rank the models. While using both corrections for R as well es for b_{ij}^Δ leads to an improvement of the predictions over the baseline, a correction only for R is already enough to achieve sufficient results for the velocity field. This also shows the potential for a multi-objective optimisation approach aiming to find models, which are correcting R and b_{ij}^Δ together instead of separately. For the best performing models on each case both the global error on U as well as

the spatial structure on U, k and τ_{xy} was coherent. The models also performed well for the periodic hills flow at a much larger Re-number ($Re = 37000$). As the sparse regression is computationally inexpensive, SpaRTA allows for rapid discovery of robust and pragmatic models, i.e. a model trained for one flow may perform well for flows outside of the training range, but with similar features.

The necessary modification in order to improve the predictions of U is simple and in-line with common knowledge on eddy viscosity turbulence models for separating flows in 2D. But, the corrections are learnt directly from data without the aid of *a priori* knowledge. In other work it was shown that using more complex function approximators from the machine learning toolbox, e.g. neural networks or random forest, more details of the flow can be captured, e.g. on the hill's crest of PH$_{10595}$, which are missed by SpaRTA. However, the present systematic study has shown the capabilities of SpaRTA to discover effective corrections to k-ω SST at low costs. Further work will focus on the uncertainty of the inference step of SpaRTA. We will also apply SpaRTA to a larger variety of flow cases in order to show its potential for rapid model discovery of corrections for industrial purposes.

Compliance with Ethical Standards

Conflict of interests   The authors declare that they have no conflict of interest.

References

1. Slotnick, J., Khodadoust, A., Alonso, J., Darmofal, D., Gropp, W., Lurie, E., Mavriplis, D.: CFD vision 2030 study: A path to revolutionary computational aerosciences. https://doi.org/10.1017/CBO9781107415324.004 (2014)
2. Xiao, H., Cinnella, P.: Quantification of model uncertainty in RANS simulations: A review. Progress in Aerosp. Sci., **108**, 1–31. https://doi.org/10.1016/j.paerosci.2018.10.001 (2019)
3. Duraisamy, K., Iaccarino, G., Xiao, H.: Turbulence modeling in the age of data. Annu. Rev. Fluid Mech., **51**(1). https://doi.org/10.1146/annurev-fluid-010518-040547 (2019)
4. Durbin, P.A.: Some recent developments in turbulence closure modeling. Annu. Rev. Fluid Mech., **50**(1). https://doi.org/10.1146/annurev-fluid-122316-045020 (2018)
5. Tracey, B.D., Duraisamy, K., Alonso, J.J.: A machine learning strategy to assist turbulence model development. In: AIAA SciTech Forum 53rd AIAA Aerospace Sciences Meeting, pp. 1–23 (2015). https://doi.org/10.2514/6.2015-1287
6. Parish, E.J., Duraisamy, K.: A paradigm for data-driven predictive modeling using field inversion and machine learning. J. Comput. Phys., **305**. https://doi.org/10.1016/j.jcp.2015.11.012 (2016)
7. Singh, A.P., Duraisamy, K.: Using field inversion to quantify functional errors in turbulence closures. Phys. Fluids, **28**(045110). https://doi.org/10.1063/1.4947045 (2016)
8. Singh, A.P., Duraisamy, K., Zhang, Z.J.: Augmentation of turbulence models using field inversion and machine learning. In: 55th AIAA Aerospace Sciences Meeting (2017). https://doi.org/10.2514/6.2017-0993

9. Ling, J., Kurzawski, A., Templeton, J.: Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. J. Fluid Mech., **807**. https://doi.org/10.1017/jfm.2016.615 (2016)

10. Wu, J.L., Xiao, H., Paterson, E.: Physicsinformed machine learning approach for augmenting turbulence models: A comprehensive framework. Phys. Rev. Fluids **3**, 074602 (2018)

11. Weatheritt, J., Sandberg, R.: A novel evolutionary algorithm applied to algebraic modifications of the RANS stressstrain relationship. J. Comput. Phys. **325**, 22 (2016). https://doi.org/10.1016/j.jcp.2016.08.015

12. Weatheritt, J., Sandberg, R.D.: The development of algebraic stress models using a novel evolutionary algorithm. Int. J. Heat Fluid, **68**. https://doi.org/10.1016/j.ijheatfluidflow.2017.09.017 (2017)

13. Pope, S.B.: A more general effective-viscosity hypothesis. J. Fluid Mech. **72**(2), 331 (1975). https://doi.org/10.1017/S0022112075003382

14. Gatski, T.B., Speziale, C.G.: On explicit algebric stress models for complex turbulent flows. J. Fluid Mech. **254**, 59 (1993). https://doi.org/10.1017/S0022112093002034

15. Leschziner, M.: Statistical turbulence modelling for fluid dynamics - demystified: an introductory text for graduate engineering students (Imperial College Press) (2015)

16. Pope, S.B.: Turblent flows. Cambridge University Press, Cambridge (2000)

17. Wallin, S.: Engineering turbulence modelling for CFD with a focus on explicit algebraic Reynolds stress models by. PhD thesis Royal Institute of Technology Stockholm (2000)

18. Weatheritt, J., Sandberg, R.D.: The development of algebraic stress models using a novel evolutionary algorithm. In: 11th International ERCOFTAC Symposium on Engineering Turbulence Modelling and Measurements, vol. 2, pp. 2–7 (2017)

19. Akolekar, H.D., Weatheritt, J., Hutchins, N., Sandberg, R.D., Laskowski, G., Michelassi, V.: Development and use of machine-learnt algebraic reynolds stress models for enhanced prediction of wake mixing in LPTs. In: Proceedings of ASME Turbo Expo 2018 (Oslo, Norway), pp. 1–13 (2018)

20. Brunton, S.L., Proctor, J.L., Kutz, J.N.: Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc. Natl. Acad. Sci., **113**(15). https://doi.org/10.1073/pnas.1517384113 (2016)

21. Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Supplementary materials for data-driven discovery of partial differential equations. Sci. Adv., **3**. https://doi.org/10.1126/sciadv.1602614 (2017)

22. Weatheritt, J., Sandberg, R.D.: Improved junction body flow modeling through data-driven symbolic regression. Journal of Ship Research. https://doi.org/10.5957/josr.09180053 (2019)

23. Parneix, S., Laurence, D., Durbin, P.A.: A procedure for using DNS databases. J. Fluids Eng., Trans. ASME **120**(1), 40 (1998). https://doi.org/10.1115/1.2819658

24. Raiesi, H., Piomelli, U., Pollard, A.: Evaluation of turbulence models using direct numerical and Large-Eddy simulation data. J. Fluids Eng. **133**(2), 021203 (2011). https://doi.org/10.1115/1.4003425. http://fluidsengineering.asmedigitalcollection.asme.org/article.aspx?articleid=1435161

25. Menter, F.R., Kuntz, M., Langtry, R.: Ten years of industrial experience with the SST turbulence model. In: K. Hanjalic, Y. Nagano, M. Tummers (eds.) Turbulence, Heat and Mass Transfer 4, pp. 625–632. Begell House, Inc. (2003)

26. Weller, H.G., Tabor, G., Jasak, H., Fureby, C.: A tensorial approach to computational continuum mechanics using object-oriented techniques. Comput. Phys. **12**(6), 620 (1998). https://doi.org/10.1063/1.168744

27. Matai, R., Durbin, P.A.: Zonal Eddy viscosity models based on machine learning. Flow Turbul. Combust., **1**. https://doi.org/10.1007/s10494-019-00011-5 (2019)

28. Gatski, T., Jongen, T.: Turbulence modeling in the age of data. Prog. AEROSP Sci. **36**, 655 (2000)

29. Wilcox, D.C.: Turbulence modeling for CFD, 3rd edn. (DCW Industries, Inc.) (2006)

30. Durbin, P.A.: On the k-3 stagnation point anomaly. Int. J. Heat Fluid Flow **17**(1), 89 (1996). https://doi.org/10.1016/0142-727X(95)00073-Y

31. Durbin, P., Pettersson-Reif, B.: Statistical theoy and modeling for turbulent flow, 2nd edn. Wiley (2011)

32. Arolla, S.K.: Modeling and Eddy simulation of rotating and curved turbulent flows. Graduate theses and dissertations, Iowa State University (2013)

33. Kato, M., Launder, B.: The modeling of turbulent flow around stationary and vibrating square cylinders. https://doi.org/10.1007/s13398-014-0173-7.2 (1993)

34. Smirnov, P.E., Menter, F.R.: Sensitization of the SST turbulence model to rotation and curvature by applying the Spalart-Shur correction term. J. Turbomach. **131**(4), 1 (2009). https://doi.org/10.1115/1.3070573

35. Brunton, S.L., Kutz, J.N.: Data-driven science and engineering machine learning, dynamical systems, and control. Cambridge University Press, Cambridge (2019). https://doi.org/10.1017/9781108380690

36. McConaghy, T.: FFX: Fast, scalable, deterministic symbolic regression technology. In: Genetic Programming Theory and Practice IX. Genetic and Evolutionary Computation. Springer, New York (2011)
37. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Series B Stat. Methodol. **67**(2), 301 (2005). https://doi.org/10.1111/j.1467-9868.2005.00503.x
38. Quade, M., Abel, M., Kutz, J.N., Brunton, S.L.: Sparse identification of nonlinear dynamics for rapid model recovery. Chaos, **28**(6). https://doi.org/10.1063/1.5027470 (2018)
39. Mangan, N.M., Kutz, J.N., Brunton, S.L., Proctor, J.L.: Model selection for dynamical systems via sparse regression and information criteria. Proc. Royal Soc. A Math. Phys. Eng. Sci., **473**(2203). https://doi.org/10.1098/rspa.2017.0009 (2017)
40. Bishop, C.M.: Pattern recognition and machine learning. Springer (2006)
41. Zhao, Y., Akolekar, H.D., Weatheritt, J., Michelassi, V., Sandberg, R.D. Turbulence model development using CFD-Driven machine learning. arXiv:1902.09075 (2019)
42. Mellen, C.P., Fröhlich, J., Rodi, W.: Large Eddy simulation of the flow over periodic hills. In: 16th IMACS World Congress (2000)
43. Breuer, M., Peller, N., Rapp, C., Manhart, M.: Flow over periodic hills - numerical and experimental study in a wide range of Reynolds numbers. Comput. Fluids **38**(2), 433 (2009). https://doi.org/10.1016/j.compfluid.2008.05.002
44. Rapp, C., Manhart, M.: Flow over periodic hills: An experimental study. Exp. Fluids **51**(1), 247 (2011). https://doi.org/10.1007/s00348-011-1045-y
45. Laval, J.P., Marquillie, M.: Direct numerical simulations of converging diverging channel flow. In: Stanislas, M., Jimenez, J., Marusic, I. (eds.) Progress in Wall Turbulence: Understanding and Modeling, vol. 14 (ERCOFTAC Series), pp. 203–209 (2011). https://doi.org/10.1007/978-90-481-9603-6
46. Bentaleb, Y., Lardeau, S., Leschziner, M.A.: Large-Eddy simulation of turbulent boundary-layer separation from a rounded step. J. Turbul., **13**(4). https://doi.org/10.1080/14685248.2011.637923 (2012)
47. Jakirlic, S.: Extended excerpt related to the test case: flow over a periodical arrangement of 2D hills. Tech. Rep. (2012)