Data-Driven Deterministic Symbolic Regression of Nonlinear Stress-Strain Relation for RANS Turbulence Modelling

Martin Schmelzer, Richard Dwight, Paola Cinnella

Data-Driven Deterministic Symbolic Regression of Nonlinear Stress-Strain Relation for RANS Turbulence Modelling

Martin Schmelzer* and Richard Dwight†
Faculty of Aerospace Engineering, Section of Aerodynamics, Delft University of Technology, The Netherlands

Paola Cinnella‡
Laboratoire DynFluid, Arts et Métiers ParisTech, Paris, France

This work presents developments towards a deterministic symbolic regression method to derive algebraic Reynolds-stress models for the Reynolds-Averaged Navier-Stokes (RANS) equations. The models are written as tensor polynomials, for which optimal coefficients are found using Bayesian inversion. These coefficient fields are the targets for the symbolic regression. A method is presented based on a regularisation strategy in order to promote sparsity of the inferred models and is applied to high-fidelity data. By being data-driven the method reduces the assumptions commonly made in the process of model development in order to increase the predictive fidelity of algebraic models.

I. Introduction

Turbulence modelling is a key-challenge for computational fluid dynamics (CFD) especially in industry. Despite considerable progress made in the field of high-fidelity turbulence modelling, such as large-eddy simulation (LES) and direct numerical simulation (DNS), RANS continues to be the standard approach used to predict a wide range of flows [1]. However, using the less-computationally demanding RANS approach comes at the price of uncertainty due to approximate physical modelling of the turbulence closure.

Despite Reynolds-stress models (RSM), based on transport equations for the Reynolds-stress tensor components, offer the more elaborate second-moment closures for RANS, modelling of industrial flows is mainly done using linear eddy viscosity models (LEVM) [2–4]. This is due to the fact that RSM involve a more complex mathematical structure, which requires additional modelling and introduces computational challenges, whilst the method does not offer a superior predictive capability for all flows [4]. However, for flows with streamline curvature, adverse pressure gradients, flow separation or rotation LEVM do not deliver reliable predictions due to their inherent inability to predict the anisotropy of the Reynolds-stress correctly. Explicit Algebraic Reynolds-stress Models (EARSM), first introduced by Pope [5] and further developed by Gatski and Speziale [6], have the potential to fill the gap by offering higher predictive fidelity compared to LEVM and being numerically more robust than RSM at similar computational costs as LEVM [7]. EARSM are derived from a projection of (simplified) RSM onto a set of tensorial polynomials (see below) [3, 4], leading to a non-linear stress-strain relation for the Reynolds-stress tensor. As such, these models can be seen as higher-order extensions of the linear eddy viscosity concept. The main simplification is the omission of anisotropy transport in order to enhance the numerical robustness, which however also reduces the predictive potential of this modelling strategy [4].

Recently a new approach based on symbolic regression utilising genetic programming (GP) was introduced to learn the non-linear stress-strain relationship for the anisotropy tensor based on high-fidelity data from DNS and LES [8, 9]. This data-driven 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 based on model fitness. In that way it produces models similar to EARSM but with a mathematical form proven to reproduce the data it was trained on. This method has the potential to generate numerically robust models with a high predictive fidelity. Even though the method is non-deterministic it discovers similar expressions for different runs. However, it is not clear if this variability comes from the data or is due to the inherent randomness of GP.

To overcome this characteristic of GP non-evolutionary methods for symbolic regression have been introduced, e.g. Fast Function Extraction (FFX) [10], Elite Bases Regression (EBR) [11], Sparse identification of nonlinear

*PhD Candidate, AWEP Department, TU Delft, Kluyverweg 2, 2629 HS Delft, The Netherlands.
†Associate Professor, AWEP Department, TU Delft, Kluyverweg 2, 2629 HS Delft, The Netherlands.
‡Professor, Laboratoire DynFluid, 151 Boulevard de l'Hopital, 75013 Paris, France.

dynamics (SINDy) [12] or PDE functional identification of nonlinear dynamics (PDE-FIND) [13]. These methods are all based on sparsity-promoting linear regression and show for a couple of cases similar or better performance and higher convergence rates for high-dimensional problems than GP. Due to their deterministic nature, they discover always the same model given the input quantities and parameters. Additionally, by varying the input parameters of the method, a hierarchy of models of different complexity and predictive fidelity are discovered, which can be used to assess overfitting for prediction.

In this work a deterministic version of symbolic regression is introduced to learn models for the nonlinear stress-strain relation. Our aim is to identify models given a set of different input features without restricting the search by too many modelling assumptions.

II. Nonlinear constitutive stress-strain relation

The challenge in RANS-based turbulence modelling is the closure of the RANS equations by a model for the Reynolds-stress tensor τ_{ij}. This symmetric second order tensor can be decomposed into

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

in which a_{ij} represents the non-dimensionalised anisotropic part and k the turbulent kinetic energy the isotropic part. Based on the decomposition of the mean velocity gradient tensor into the symmetric and the anti-symmetric part

$$\partial_j U_i = s_{ij} + \omega_{ij}, \tag{2}$$

with the mean strain-rate tensor $s_{ij} = \frac{1}{2}(\partial_j U_i + \partial_i U_j)$ and the rotation rate tensor $\omega_{ij} = \frac{1}{2}(\partial_j U_i - \partial_i U_j)$, Pope was the first to use the Cayley-Hamilton theorem in order to derive an integrity basis of ten nonlinear base tensors and five corresponding invariants [5]. Given a turbulent time scale τ to non-dimensionalise $S_{ij} = \tau s_{ij}$ and $\Omega_{ij} = \tau\omega_{ij}$ the base tensors can be combined linearly as a functional form for a_{ij}:

$$a_{ij}(S_{ij},\Omega_{ij}) = \sum_{n=1}^{N} T_{ij}^{(n)}\alpha_n. \tag{3}$$

The base tensors are

$$\begin{aligned}
T_{ij}^1 &= S_{ij},\ T_{ij}^2 = S_{ik}\Omega_{kj} = \Omega_{ik}S_{kj},\\
T_{ij}^3 &= S_{ik}S_{kj} - \frac{1}{3}\delta_{ij}S_{mn}S_{nm},\\
T_{ij}^4 &= \Omega_{ik}\Omega_{kj} - \frac{1}{3}\delta_{ij}\Omega_{mn}\Omega_{nm},\\
T_{ij}^5 &= \Omega_{ik}S_{kl}S_{lj} - S_{ik}S_{kl}\Omega_{lj},\\
T_{ij}^6 &= \Omega_{ik}\Omega_{kl}S_{lj} + S_{ik}\Omega_{kl}\Omega_{lj} - \frac{2}{3}S_{kl}\Omega_{lm}\Omega_{mk},\\
T_{ij}^7 &= \Omega_{ik}S_{km}\Omega_{mn}\Omega_{nj} - \Omega_{ik}\Omega_{km}S_{mn}\Omega_{nj},\\
T_{ij}^8 &= S_{ik}\Omega_{km}S_{mn}S_{nj} - S_{ik}S_{km}\Omega_{mn}S_{nj},\\
T_{ij}^9 &= \Omega_{ik}\Omega_{km}S_{mn}S_{nj} + S_{ik}S_{km}\Omega_{mn}\Omega_{nj} - \frac{2}{3}S_{km}S_{mn}\Omega_{np}\Omega_{pk},\\
T_{ij}^{10} &= \Omega_{ik}S_{km}S_{mn}\Omega_{np}\Omega_{pj} - \Omega_{ik}\Omega_{km}S_{mn}S_{np}\Omega_{pj}
\end{aligned} \tag{4}$$

and the corresponding invariants read

$$\begin{aligned}
I_1 &= S_{mn}S_{nm},\ I_2 = \Omega_{mn}\Omega_{nm},\ I_3 = S_{km}S_{mn}S_{nk}\\
I_4 &= \Omega_{km}\Omega_{mn}S_{nk},\ I_5 = \Omega_{km}\Omega_{mn}S_{np}S_{pk}.
\end{aligned} \tag{5}$$

While the combination of base tensors is linear, according to Pope's analysis the scalar coefficient fields α_n are nonlinear functions of the invariants, i.e. $\alpha_n = \alpha_n(I_1,...,I_5)$. The identification of the functional form of the coefficients

is the essential step to build a nonlinear eddy-viscosity model. Classical methods to identify the functional forms are given in [5, 14]. As mentioned above these methods are based on modelling assumptions such as the weak equilibrium hypothesis, which omits the transport of anisotropy [15]. In the following, optimal scalar fields α_n are identified for the given test cases and further used to learn a functional form for α_n using symbolic regression.

III. Test case, high-fidelity data and solver

The test case, for which we want to infer and test correction models, is the flow over a series of periodic hills in 2D [16], the geometry is shown in Fig. 1. The flow exhibits separation on the leeward side of the hills and reattachment in between. Both features are challenging to predict by LEV models [17]. We consider two different Reynolds-numbers $Re_H = 2800$ and 10595 based on the bulk-velocity at the hill's crest and the hill height. At both the lower as well as the upper wall a no-slip boundary condition is applied. The flow setup is periodic in the stream-wise direction and driven with a volume forcing. The CFD solver used is `simpleFoam` from the OpenFOAM toolbox [18]. The inferred correction models resulting from the proposed learning method are implemented in a new turbulence model based on the $k-\omega$ model. The high-fidelity data used for model-learning and validation are mean velocity fields and Reynolds-stress fields from Breuer et al. [19]. More precisely, DNS data for the lower Reynolds number case and LES data for the higher Reynolds number one.

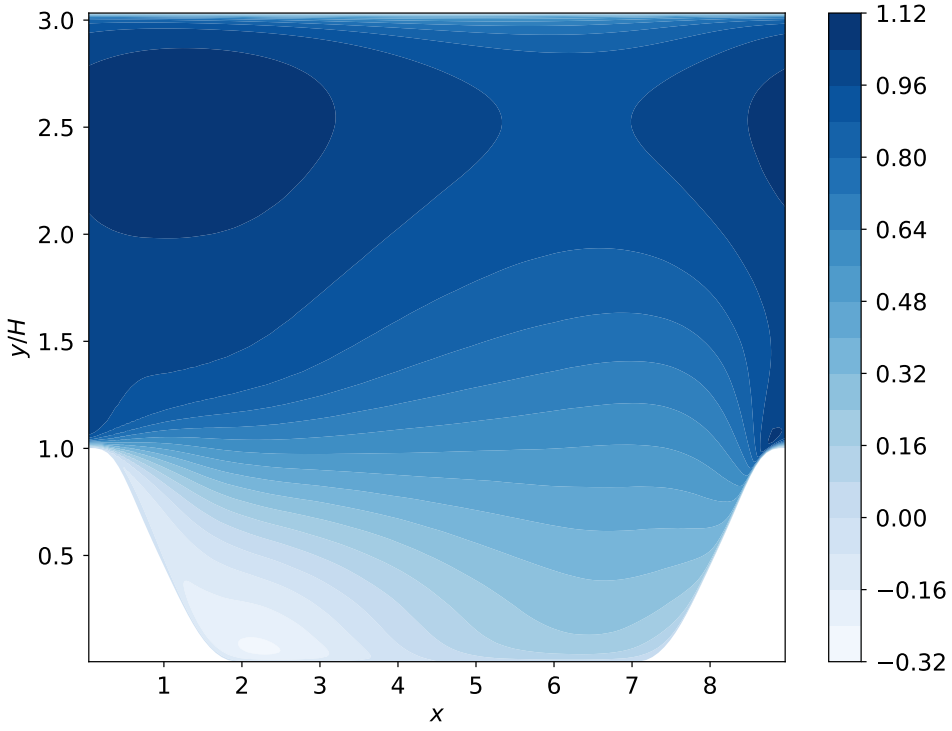


Fig. 1 Stream-wise velocity of flow over periodic hills in 2D at $Re_H = 10595$. LES data from [19].

IV. Optimal coefficients for the nonlinear constitutive relation

As the modelling of the anisotropy of the Reynolds-stress tensor using the linear stress-strain relation is insufficient, we introduce an additive term b_{ij}^{Δ} to account for the model-form error

$$\tau_{ij} = -2\nu_t S_{ij} + \frac{2}{3}k\delta_{ij} + 2kb_{ij}^{\Delta}, \tag{6}$$

which is a non-dimensional symmetric tensor with zero trace. The mean strain-rate S_{ij}, the Reynolds-stress tensor τ_{ij} and the turbulent kinetic energy k are directly available from high-fidelity simulations (LES or DNS) or from experiments. The eddy viscosity ν_t can be obtained by passively solving the turbulent transport equations of a turbulence model, e.g. $k-\omega$, for k and the mean velocity field U_i (frozen approach) [9]. The model-form error b_{ij}^{Δ} is then readily computable as a residual given the equation above.

The obtained model-form error is case-dependent, meaning that it only corrects the stress-strain relation for a given setup (geometry, boundary conditions, Reynolds number, etc.). The key is to distill information from it to identify a suitable mathematical expression as a corrective term, which can be used for prediction of other flow cases. For that we identify the coefficient fields of the base tensor series given input data b_{ij}^{Δ}. The relation between the data and the base tensor series is expressed by a statistical model

$$\boldsymbol{b}^{\boldsymbol{\Delta}} = \boldsymbol{H}\boldsymbol{a} + \boldsymbol{\epsilon}, \tag{7}$$

in which the sparse block-diagonal matrix $\boldsymbol{H} \in \mathbb{R}^{6K\times M}$ with $M = N \times K$ contains the base tensors ordered per component and spatial location $k = 1,...,K$

$$\boldsymbol{H} = \begin{bmatrix}
T^{(1)}_{xx,1} & \cdots & T^{(N)}_{xx,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
T^{(1)}_{xy,1} & \cdots & T^{(N)}_{xy,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
T^{(1)}_{xz,1} & \cdots & T^{(N)}_{xz,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
T^{(1)}_{yy,1} & \cdots & T^{(N)}_{yy,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
T^{(1)}_{yz,1} & \cdots & T^{(N)}_{yz,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
T^{(1)}_{zz,1} & \cdots & T^{(N)}_{zz,1} & 0 & \cdots & 0 & \dots & 0 & \cdots & 0 \\
0 & \cdots & 0 & T^{(1)}_{xx,2} & \cdots & T^{(N)}_{xx,2} & \dots & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\
0 & \cdots & 0 & T^{(1)}_{zz,2} & \cdots & T^{(N)}_{zz,2} & \dots & 0 & \cdots & 0 \\
\vdots & & \vdots & \vdots & & \vdots & \ddots & \vdots & & \vdots \\
0 & \dots & 0 & 0 & \dots & 0 & \dots & T^{(1)}_{zz,K} & \cdots & T^{(N)}_{zz,K}
\end{bmatrix} \tag{8}$$

and the vectors $\boldsymbol{a} \in \mathbb{R}^{M}$ and $\boldsymbol{b}^{\boldsymbol{\Delta}} \in \mathbb{R}^{6K}$ contain the stacked coefficient and tensor fields as well ordered per component and spatial location, respectively:

$$\begin{aligned}
\boldsymbol{a} &= (\alpha_{1,1},\dots,\alpha_{N,1},\alpha_{1,2},\dots,\alpha_{N,K})^T, \\
\boldsymbol{b}^{\boldsymbol{\Delta}} &= (b^{\Delta}_{xx,1}, b^{\Delta}_{xy,1}, b^{\Delta}_{xz,1}, b^{\Delta}_{yy,1}, b^{\Delta}_{yz,1}, b^{\Delta}_{zz,1}, b^{\Delta}_{xx,2},\dots, b^{\Delta}_{zz,K})^T.
\end{aligned} \tag{9}$$

The random variable $\boldsymbol{\epsilon}$ serves as an additive error term expressing the fact, that given the best set of coefficients $\alpha_{n,k}$ the base tensor series might not fit the data perfectly. As the input data (Reynolds-stress and mean velocity field) is relatively smooth spatially, we also want to infer smooth coefficients fields. In a preliminary study utilising ordinary least-square regression, we discovered rough coefficient fields, in which the values next to each other differed several orders of magnitude, which is a common issue for inverse problems [20]. While these coefficients lead to a very low reconstruction error of the anisotropic Reynolds-stress, they are discarded as they are not useful to serve as a target for the symbolic regression. A lack of smoothness leads to unreasonably complex models, with an increased risk of overfitting the rough coefficient fields. Furthermore, more complex correction models also increase the numerical stiffness of the CFD solver leading to a reduction or loss of numerical stability. Even though the smoothness constraint increases the a priori reconstruction error, it enables the discovery of sparse correction models.

The inference of the coefficients given the data $\boldsymbol{b}^{\boldsymbol{\Delta}}$ is done via Bayesian inversion [21, 22]

$$p(\boldsymbol{a}|\boldsymbol{b}^{\boldsymbol{\Delta}}) \propto L(\boldsymbol{b}^{\boldsymbol{\Delta}}|\boldsymbol{a})\, p(\boldsymbol{a}) \tag{10}$$

in which the posterior probability $p(\boldsymbol{a}|\boldsymbol{b}^{\boldsymbol{\Delta}})$ is the product of the likelihood $L(\boldsymbol{b}^{\boldsymbol{\Delta}}|\boldsymbol{a})$ based on the statistical model defined above and the prior probability $p(\boldsymbol{a})$, which embodies the smoothness constraint. As the dimension of this problem depends on the mesh size, which is already large for the considered 2D flow case (18000 cells), we focus on the inference of the maximum a priori estimation (MAP) of the posterior

$$\boldsymbol{a}_{MAP} = \arg\max_{\boldsymbol{a}} p(\boldsymbol{a}|\boldsymbol{b}^{\boldsymbol{\Delta}}) \tag{11}$$

and not the full posterior distribution. We further simplify the problem by assuming that both the likelihood function and the prior are normally distributed

$$L(\boldsymbol{b}^{\boldsymbol{\Delta}}|\boldsymbol{a}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{b}^{\boldsymbol{\Delta}} - \boldsymbol{H}\boldsymbol{a})^T \boldsymbol{R}^{-1}(\boldsymbol{b}^{\boldsymbol{\Delta}} - \boldsymbol{H}\boldsymbol{a})\right], \tag{12}$$

$$p(\boldsymbol{a}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{a}_o)^T \boldsymbol{P}^{-1}(\boldsymbol{a} - \boldsymbol{a}_o)\right]. \tag{13}$$

Thus, also the posterior is normally distributed and, under the assumption of a prior zero mean $\boldsymbol{a}_o = 0$, it reads

$$p(\boldsymbol{a}|\boldsymbol{b}^{\boldsymbol{\Delta}}) \propto \exp\left[-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{a}_{MAP})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{a}_{MAP})\right] \tag{14}$$

with mean $\boldsymbol{a}_{MAP}$ and covariance matrix $\boldsymbol{\Sigma}$

$$\boldsymbol{a}_{MAP} = K\boldsymbol{b}^{\boldsymbol{\Delta}}, \tag{15}$$

$$\boldsymbol{\Sigma} = (\boldsymbol{I} - \boldsymbol{K}\boldsymbol{H})\boldsymbol{P}, \tag{16}$$

$$K = \boldsymbol{P}\boldsymbol{H}^T(\boldsymbol{R} + \boldsymbol{H}\boldsymbol{P}\boldsymbol{H}^T)^{-1} \tag{17}$$

the latter is known as the Kalman gain [21]. In the present study the observation covariance matrix $\boldsymbol{R} \in \mathbb{R}^{6K}$ is assumed to be diagonal $\boldsymbol{R} = \sigma_R \boldsymbol{I}$ implying that the underlying spatial correlation of the data $\boldsymbol{b}^{\boldsymbol{\Delta}}$ is omitted. Further work will also use other covariance structures. The prior covariance $\boldsymbol{P} \in \mathbb{R}^M$ acts to regularise the inversion problem in order to obey the smoothness constraint on the coefficient fields and is defined as

$$\boldsymbol{P}(\boldsymbol{x}, \boldsymbol{x}') = \begin{cases} \sigma_P \exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{x}')^2}{L^2}\right], & \text{if } \exp\left[-\frac{(\boldsymbol{x}-\boldsymbol{x}')^2}{L^2}\right] \geq 0.001 \\ 0, & \text{otherwise,} \end{cases} \tag{18}$$

in which σ_P is the scalar variance, $\boldsymbol{x}$ represents the coordinates of location within the mesh and L is the correlation length. The Gaussian kernel is a common choice resulting in a smooth correlation function. With a correlation length shorter then the smallest cell size the MAP-inversion framework reduces to a regularised least-squares problem, while a larger correlations length increases both the smoothness of the coefficients and the reconstruction error.

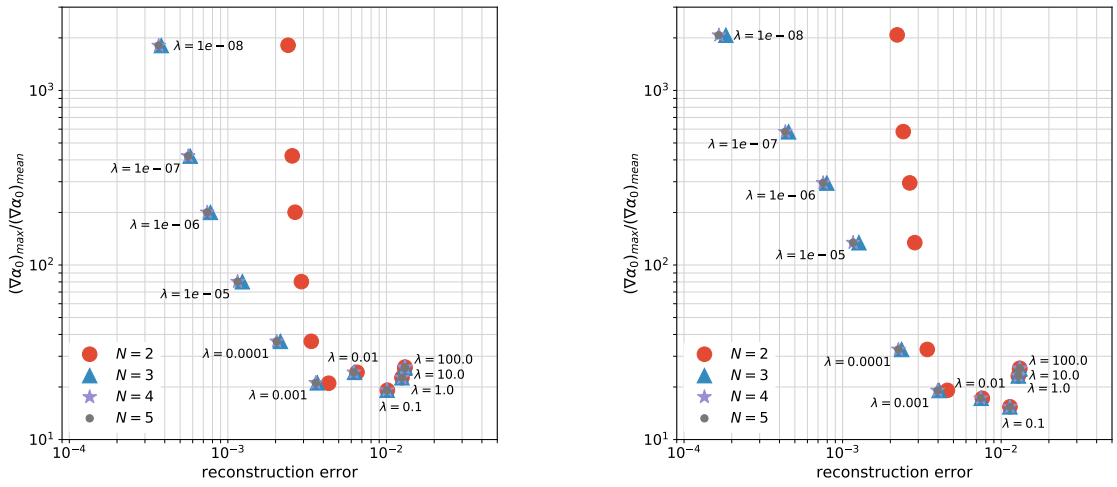
Given values for σ_R, σ_P and L we are now able to compute $\boldsymbol{a}_{MAP}$ with (15). As the size of the problem and therefore the size of the matrices present in the Kalman gain in (17) scales with the number of mesh points of the flow case at hand, we need to exploit the sparsity of the matrices in order to store them efficiently and make the computations tractable. Due to their structure defined above especially the memory requirements for $\boldsymbol{H}$ and $\boldsymbol{R}$ can be drastically reduced. The density of $\boldsymbol{P}$ given (18) is controlled by a cut-off for small-sized correlations. As it depends on the correlation length L, also its memory requirements increase with a larger correlations length. For the periodic hill test case we have tested $L = 0.01$ and $L = 0.001$, see Fig. 2a and 2b respectively. In these figures, the reconstruction error $\epsilon = \left[(\boldsymbol{b}^{\boldsymbol{\Delta}} - \boldsymbol{H}\hat{\boldsymbol{a}})^T(\boldsymbol{b}^{\boldsymbol{\Delta}} - \boldsymbol{H}\hat{\boldsymbol{a}})\right]^{0.5}$ for an inferred set of coefficients $\hat{\boldsymbol{a}}$ is shown against an empirical roughness criterion

$$s = \frac{(|\nabla \hat{\boldsymbol{a}}_0|)_{\max}}{(|\nabla \hat{\boldsymbol{a}}_0|)_{\text{mean}}}, \tag{19}$$

corresponding to the maximum of the gradient of $\hat{\boldsymbol{a}}_0 = \hat{\alpha}_{0,k}$ normalised by its mean value. Unregularised inversion is prone to small differences in the data leading to large discrepancies in the inferred coefficients [20] , i.e. localised large gradients. The rationale of the criterion is therefore to identify if unphysically large gradients are present for a given regularisation strategy. In general, the lower s is, the lower are the highest gradients, which contributes to smoother coefficient fields. The coloured marker-lines represent a different number of employed base tensors ranging from $N = 2$ to 5, which is a common choice for EARSM for 2D [4], and the scalar variances of $\boldsymbol{P}$ and $\boldsymbol{R}$ are summarised to

$$\lambda = \frac{\sigma_R}{\sigma_P}, \tag{20}$$

equivalent to a Tikhonov regularisation parameter in case of regularised least-square regression [20].



(a) $L = 0.01$

(b) $L = 0.001$

Fig. 2 Reconstruction error versus roughness metric s for two different correlation length L. The number of base tensors N are indicated by coloured marker lines. The regularisation parameter λ are associated to the points next to the mentioned value.

In general, the lower the regularisation parameter λ, the lower the reconstruction error and the larger s, and vice versa. For $\lambda \leq 0.001$ the reconstruction error decreases while the roughness increases more than one order of magnitude. By increasing the number of base tensors from $N = 2$ to $N = 3$ the reduction of the reconstruction error is significant for a given λ, but only minor for a further expansion. This implies that for a correction model at least $N = 3$ should be considered. For $\lambda \geq 0.001$ the reconstruction error increases but the smoothness settles, also the number of base tensors is not important anymore in terms of the reconstruction error, meaning that the problem is dominated by the regularisation.

Changing the correlation length ($L = 0.01$ in Fig.2a, $L = 0.001$ in Fig.2b) decreases the roughness of the coefficient fields slightly, but has a minor effect overall. We have observed for even larger correlation lengths a further decrease in roughness for a given λ, but also an increase in the reconstruction error. Smoothing the coefficient fields by increasing L is therefore detrimental to the endeavour of designing a prior to steer the inversion towards low roughness and low reconstruction error (lower-left corner of the Fig. 2).

Finally, the preceding parametric study helps us in identifying suitable coefficient fields to serve as targets for the symbolic regression. We identify coefficient fields given $L = 0.01$ and $\lambda = 0.0001$ corresponding to a reasonable

compromise between low bias according to a low reconstruction error and low roughness s. Further research will focus on a rigorous optimisation of the parameters λ and L to find an optimum of low reconstruction error and low roughness.

V. Deterministic Symbolic Regression of Correction Models

This Section deals with the methodology of deterministic symbolic regression. First, it is explained how a library of candidate functions is build automatically given raw input features and a set of mathematical operations. Second, details are given on the algorithm to identify the best linear combination of as few as possible candidate functions to fit the target coefficient fields, which were identified in the preceding section.

A. Building a library of candidate functions

The deterministic symbolic regression constructs an over-complete library of possible nonlinear candidate functions $\mathcal{B}$ approximating given data and identifies the relevant ones by adopting a sparsity constrain. An important step is the design of the library of candidate functions. Because the correction models we want to identify are based on the nonlinear eddy viscosity concept, we use the invariants I_1 and I_2 of the nonlinear base tensor expansion as our primitive input features, which are shown for the periodic hill test case calculated from the mean velocity of the LES data of [19] at $Re_H = 10,595$ in Fig. 3a and 3b respectively. Further candidate functions are constructed by applying mathematical operations to the already existing ones present in the library of candidate functions, see Table 1. The operations can be applied in a different order and also repetitively. For example a library containing only polynomials and products of the resulting set of candidates is encoded in

$$\mathcal{B} = \{\texttt{R} \quad | \quad \texttt{P} \quad | \quad \texttt{M}\}, \tag{21}$$

in which the vertical line | indicates that the candidate functions resulting from the operation to the left are passed on to the operation to the right. The resulting new functions after each operation are appended to the library. This procedure makes the construction of the library automatic similar to FFX [10].

Symbol	Operation	Details
R	$(\cdot)$	raw input features
P	$(\cdot)^p$	$p = [\pm 0.5, \pm 2, ...]$
F	$f(\cdot)$	$f = [\sin, \cos, \exp, \log, ...]$
A	$(\cdot) + (\cdot)$	summation
M	$(\cdot) \times (\cdot)$	product

Table 1 Operations to build a library of candidate functions

Since the procedure can produce duplicates at different stages of the process, we check at the end for equivalent expressions and retrieve only a unique set of candidate functions. In addition for every candidate function it is checked, if it can be evaluated for the flow case, for which we want to find a correction model. Functions, which are not defined for the input domain provided by the data, are deleted from the library. Obviously the nonlinearity of the resulting library $\mathcal{B}$ can be determined by the type and order of the operations and by their frequency. This is illustrated in Table 2.

Order	example of resulting most complex candidates
R | P | M	$I_1^{0.5} I_2^2$
R | P | F | M	$\exp(I_1^{0.5}) \log(I_2^2)$
R | P | A | F	$\exp(I_1^{0.5} + I_2^2)$

Table 2 Illustration of resulting most complex candidate functions for a given order of operations.

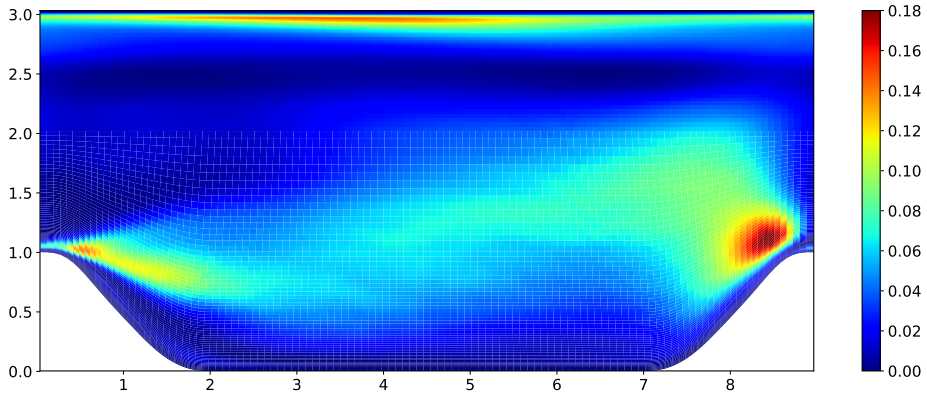
As a starting point for the present work we focus on a weakly nonlinear library using the rule described in (21) with $p = \{1.0, 2.0\}$. In addition to the two invariants I_1 and I_2 we also include a constant function c to the set of raw input

features. This helps to identify very sparse models, for which the form of a base tensors only needs to be scaled by a factor and not altered by a spatially-dependent function. The resulting library $\mathcal{B}$ has 16 candidate functions and reads

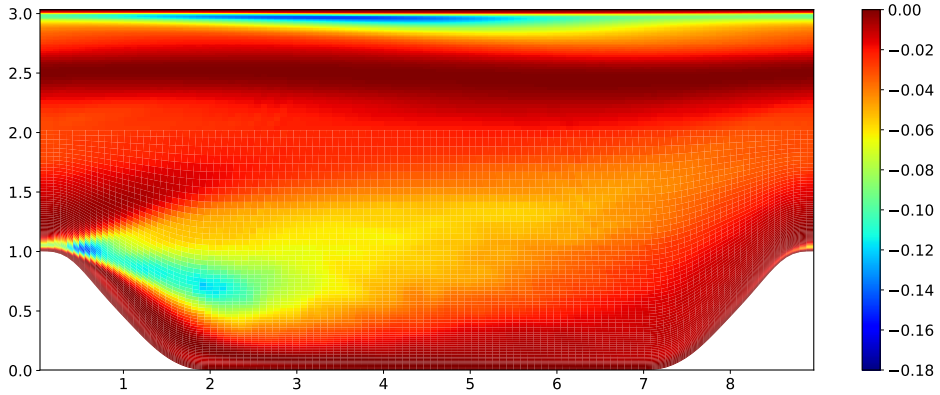
$$\mathcal{B} = \left[c, I_1, I_2, I_1^2, I_2^2, I_1^2 I_2^3, I_1^4 I_2^2, I_1 I_2^2, I_1 I_2^3, I_1 I_2^4, I_1^3 I_2, I_1^2 I_2^4, I_1^2 I_2, I_1 I_2, I_1^3 I_2^2, I_1^2 I_2^2\right] \tag{22}$$

which are evaluated using the reference data. Thus, by using $p = \{1.0, 2.0\}$ and allowing products of candidate function with the same base, we effectively achieve exponents $p = \{1.0, 2.0, 3.0, 4.0\}$. The evaluated candidate functions are stored column-wise in the library matrix

$$\boldsymbol{B} = \begin{bmatrix} c|_{k=0} & I_1|_{k=0} & \dots & I_1^2 I_2^2|_{k=0} \\ \vdots & \vdots & & \vdots \\ c|_{k=K} & I_1|_{k=K} & \dots & I_1^2 I_2^2|_{k=K} \end{bmatrix}. \tag{23}$$



(a) I_1



(b) I_2

Fig. 3 Two invariants of the nonlinear base tensor series for flow over periodics hills at $Re = 10{,}595$. Data from [19].

B. Deterministic Symbolic Regression

The deterministic symbolic regression identifies the most relevant candidate functions by solving the optimisation problem

$$\mathbf{\Theta}^{(n)} = \underset{\hat{\mathbf{\Theta}}^{(n)}}{\arg\min} \left\| \boldsymbol{B}\hat{\mathbf{\Theta}}^{(n)} - \boldsymbol{a}_n \right\|_2^2 + \lambda_r \left\| \hat{\mathbf{\Theta}}^{(n)} \right\|_q, \tag{24}$$

in which the vector of coefficients $\mathbf{\Theta}^{(n)}$ needs to be identified. The target is a specific coefficient field $\boldsymbol{a}_n$ for a given base tensor n. The regularisation term using norm $q = 1$ (lasso regression) or $q = 2$ (ridge regression) acts to increase the sparsity of $\mathbf{\Theta}^{(n)}$, i.e. increasing the number of zeros in order to turn off the corresponding candidate functions [12, 13], proportional to the magnitude of the Tikhonov parameter λ_r.

We use STRidge (Sequential Threshold Ridge Regression), which solves the optimisation problem in Eq. 24 using ridge regression $q = 2$ iteratively [12]. After each step sparsity is achieved by setting coefficients to zero, which have a smaller magnitude than a given threshold $t_{\max}$, and the regression is repeated for the non-zero coefficients until a suitable model is identified. For the resulting set of coefficients an ordinary least-square regression, i.e. $\lambda_r = 0$, is performed to achieve the unbiased values for the coefficients. Once $\mathbf{\Theta}^{(n)}$ has been identified for every base tensor n, the mathematical expression of the discovered correction model can be retrieved from

$$M := b_{ij}^{\Delta} = \sum_{n=1}^{N} \boldsymbol{\mathcal{B}}\mathbf{\Theta}^{(n)} T_{ij}^{(n)}. \tag{25}$$

Using data the flow over periodic hills at $Re_H = 10595$ [19] only one step of STRidge with $\lambda_r = 0.02$ was performed and the threshold was defined as $t_{\max} = \xi \max(|\hat{\theta}|)$ with $0 < \xi < 1$ depending on the largest absolute coefficient $\max(|\hat{\theta}|)$. This variation of STRidge allows to determine a hierarchical set of models given different values of ξ. Give $\xi = 0.9$ the expression for the resulting correction model reads

$$M^{(1)} = 7.0815\ I_1\ T_{ij}^{(1)} + 4.2099\ T_{ij}^{(2)} + 3.6909\ I_1\ T_{ij}^{(3)} - 0.0509\ T_{ij}^{(4)} \tag{26}$$

and by applying the same setup for the symbolic regression to the data of $Re_H = 2800$ we obtain

$$M^{(2)} = -6.4069\ I_2\ T_{ij}^{(1)} + 4.1292\ T_{ij}^{(2)} - 17.4622\ I_2\ T_{ij}^{(3)} + 18.1459\ I_2\ T_{ij}^{(4)}. \tag{27}$$

The models have a very sparse structure as only the raw input features and the constant have been identified from $\boldsymbol{\mathcal{B}}$ to serve as functions for the coefficients. This is a result of a large $\xi = 0.9$ value. Interestingly, for $M^{(1)}$ only I_1 with a positive coefficient and for $M^{(2)}$ only I_2 with a negative coefficient have been chosen for $T_{ij}^{(1)}$ and $T_{ij}^{(3)}$. This choice is reasonable, as the invariants exhibit a similar spatial structure with opposite signs shown in Fig. 3. While the coefficient for $T_{ij}^{(3)}$ is almost the same for both models, the identified function for $T_{ij}^{(4)}$ is different.

The a priori error is calculated using two metrics, the root-mean squared error ϵ and the tensor alignment ρ between the training data and the model

$$\epsilon(M) = \sqrt{\frac{1}{6K} \sum_{k=1}^{K} \sum_{i=1}^{3} \sum_{j=1}^{i} \left(M_{ij,k} - b_{ij,k}^{\Delta} \right)^2} \tag{28}$$

$$\rho(M) = \sum_{k=1}^{K} \frac{M_{ij,k} b_{ij,k}^{\Delta}}{M_{mn,k} M_{nm,k} b_{pq,k}^{\Delta} b_{qp,k}^{\Delta}}, \tag{29}$$

in which the tensors components ij are evaluated at each grid point k. The results are given in Table 3. Given the l2-norm of the distance between the model and the data in eq. 24 the metric ϵ quantifies the achieved goodness-of-fit of the symbolic regression. In addition, the metric ρ calculates how well the resulting tensor shape is aligned with the data, and therefore does not include magnitude information. In comparison to the literature the achieved a priori error is similar to other approaches using deep learning [23] or genetic programming based symbolic regression [8].

Model	Training case	λ	L	λ_r	ξ	ϵ	ρ
$M^{(1)}$	$Re_H = 10595$	0.0001	0.01	0.02	0.9	0.0875	0.8197
$M^{(2)}$	$Re_H = 2800$	0.0001	0.01	0.02	0.9	0.0768	0.7981

Table 3 Parameters in derivation and a priori error of the models.

VI. Prediction

Fig. 4 shows predictions of the stream-wise velocity for Reynolds number $Re_H = 10595$ using the correction models $M^{(1)}$ and $M^{(2)}$. While $M^{(1)}$ has been derived from the same dataset, using it to predict the same flow serves as a verification. In general, both models show an improvement compared to the baseline LEVM $k-\omega$ all over the domain. The most challenging quantity of this flow case is the reattachment point x_a between the hills. This is reflected by a vanishing wall-shear stress at the lower wall, shown in Fig. 5 and summarised in Tab. 4. Again the predictions of the models are closer to the reference data than the baseline LEVM, which expresses the fact that the symbolic regression successfully identified corrections of the baseline LEVM. Also for all shown quantities $M^{(1)}$ is closer to the reference data than $M^{(2)}$ consistently. However, for the lower Reynolds-number case ($Re_H = 2800$) shown in Fig. 6, $M^{(1)}$ performs similarly well and predicts the reattachment closer to the reference than $M^{(2)}$. This behaviour is not unexpected as the models are derived to be sparse corrections of the underlying LEVM in order to avoid over-fitting. Further work needs to be done in order to study the dependency of the sparsity-controlling cut-off parameter ξ on the predictive performance of the resulting models in relation to the Reynolds-number.

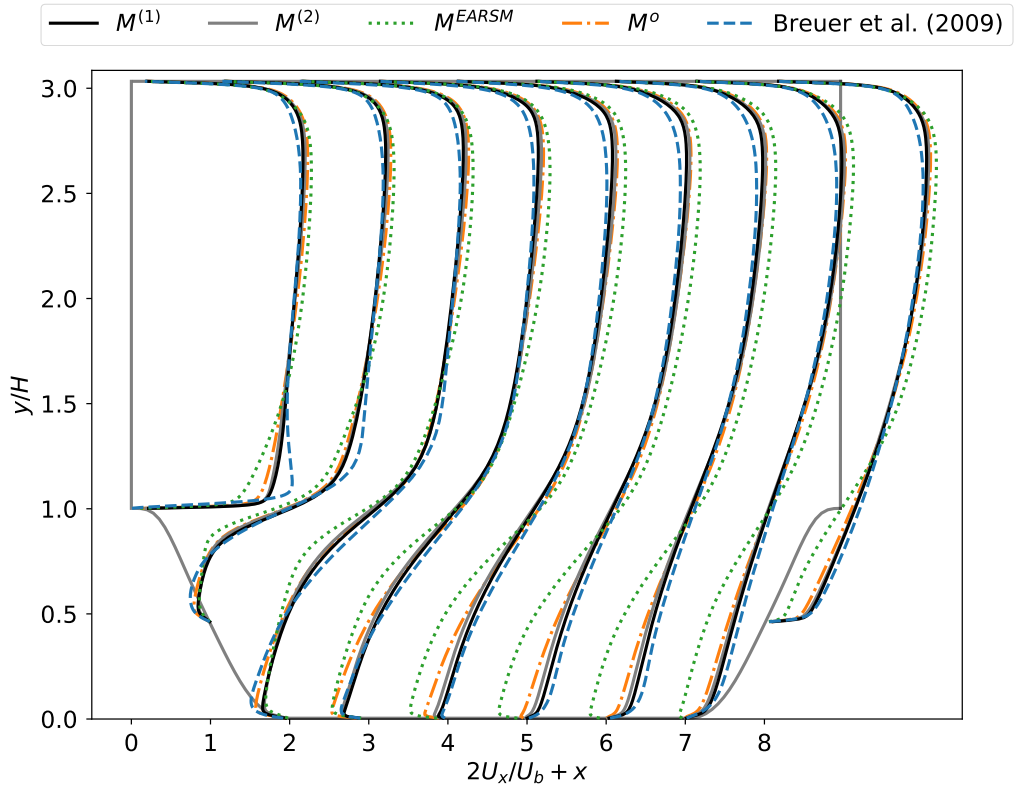


Fig. 4 Predictions of the stream-wise velocity for $Re_H = 10595$ using correction model $M^{(1)}$, $k-\omega$ LEVM M^o and EARSM M^{EARSM} from literature [24]. Validation data from [19].

For the present case, the Explicit-Algebraic Reynolds-stress model (EARSM) from literature [24] performs worse than the linear $k-\omega$ model all over the domain at $Re_H = 10595$, see Fig. 4. In a benchmark study [17], in which several EARSM and RSM were applied to this test case, it was also observed that not necessarily more complex models

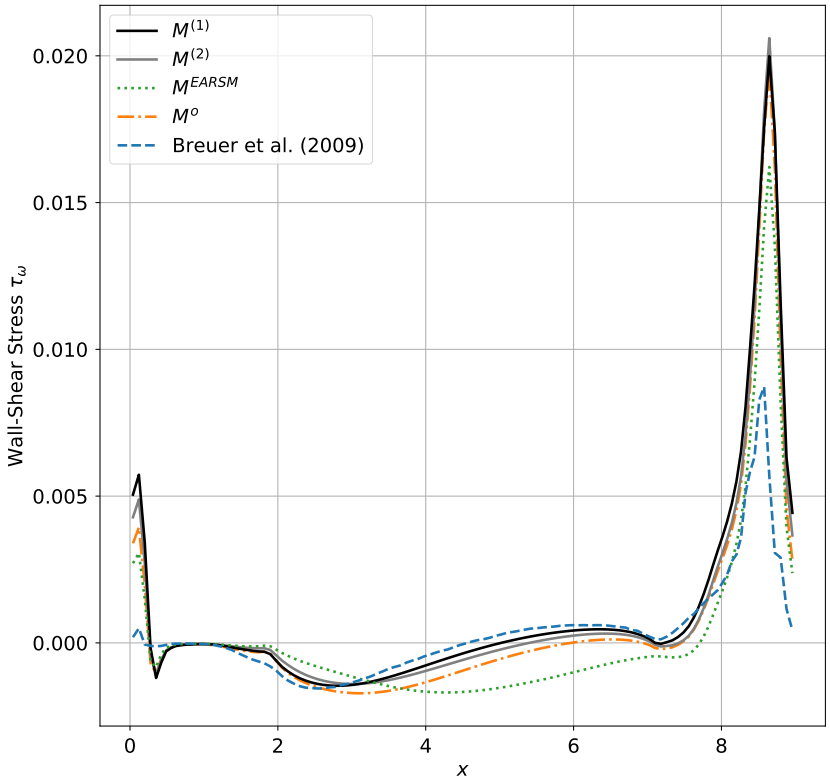


Fig. 5 Prediction of the wall-shear stress for $Re_H = 10595$ using correction model $M^{(1)}$, $k-\omega$ LEVM M^o and EARSM M^{EARSM} from literature [24]. Validation data from [19].

perform better.

	Re_H	$M^{(1)}$	$M^{(2)}$	M^o	LES[19]
x_a	2800	5.69	6.17	7.61	5.37
x_a	10595	5.02	5.42	5.96	4.54

Table 4 Predictions of reattachment point x_a for $Re_H = 10595$ using correction model $M^{(1)}$, $M^{(2)}$ and $k-\omega$ LEVM M^o in comparison to the LES data from [19].

VII. Conclusion

In this contribution we show that symbolic regression is a powerful machine learning method to learn the mathematical form of models from data for the purpose of RANS-based turbulence modelling. The goal is to identify concise models ready to be implemented in existing codes as used in industry proven in [8, 9] but with a deterministic framework, which can be applied to high-dimensional regression problems.

We introduced a framework for the inference of optimal coefficients fields of the base tensor series used in the nonlinear eddy viscosity concept. By utilising a maximum a posteriori (MAP) based Bayesian inversion we conduced a parameter study in order to achieve coefficients with low roughness and low reconstruction error by using data from high-fidelity simulations of the flow over periodic hills at $Re_H = 2800$ and 10595.

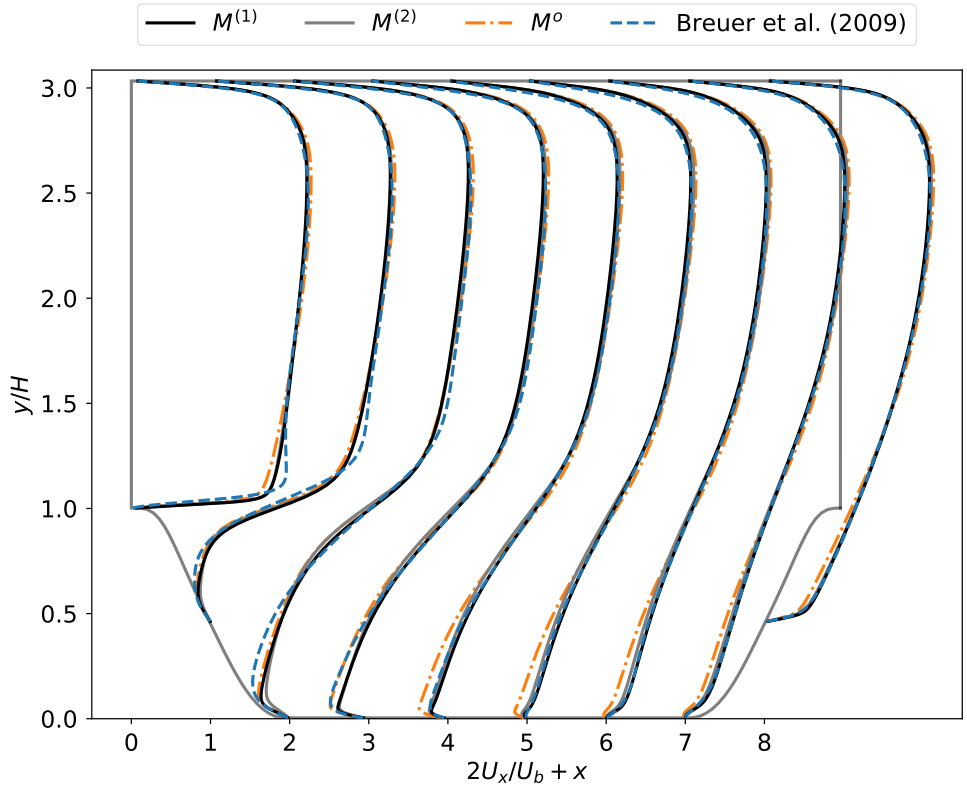


Fig. 6 Predictions of the stream-wise velocity for Reynolds number $Re_H = 2800$ using two correction models $M^{(1)}$ and $M^{(2)}$ and $k-\omega$ LEVM M^o. Validation data from [19].

The resulting coefficient fields were used as targets of the symbolic regression based on sparsity-promoting regularised least square regression. We have successfully identified models and presented the a priori error and the predictive performance of both of them.

The method shows potential for data-driven correction of RANS-based turbulence modelling. Next steps will be the application to flow cases with a different geometry and establishing a relation between the a priori error and the predictive performance of a model. In addition, the two step procedure of inferring the targets and conducting the symbolic regression separately will be merged into a single optimisation framework.

References

[1] Slotnick, J., Khodadoust, A., Alonso, J., Darmofal, D., Gropp, W., Lurie, E., and Mavriplis, D., "CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences," Tech. Rep. March, 2014. doi:10.1017/CBO9781107415324.004.

[2] Basara, B., and Jakirlic, S., "A new hybrid turbulence modelling strategy for industrial CFD," *International Journal for Numerical Methods in Fluids*, Vol. 42, No. 1, 2003, pp. 89–116. doi:10.1002/fld.492.

[3] Pope, S. B., *Turblent Flows*, Cambridge University Press, 2000.

[4] Leschziner, M., *Statistical Turbulence Modelling for Fluid Dynamics - Demystified: An Introductory Text for Graduate Engineering Students*, 2015.

[5] Pope, S. B., "A more general effective-viscosity hypothesis," *Journal of Fluid Mechanics*, Vol. 72, No. 2, 1975, pp. 331–340. doi:10.1017/S0022112075003382.

[6] Gatski, T. B., and Speziale, C. G., "On explicit algebric stress models for complex turbulent flows," *Journal of Fluid Mechanics*, Vol. 254, 1993, pp. 59–78. doi:10.1017/S0022112093002034.

[7] 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.

[8] Weatheritt, J., and Sandberg, R., "A novel evolutionary algorithm applied to algebraic modifications of the RANS stress–strain relationship," *Journal of Computational Physics*, Vol. 325, 2016, pp. 22–37. doi:10.1016/j.jcp.2016.08.015.

[9] Weatheritt, J., and Sandberg, R. D., "The development of algebraic stress models using a novel evolutionary algorithm," *Flow, Turbulence and Combustion*, 2017. doi:10.1016/j.ijheatfluidflow.2017.09.017.

[10] McConaghy, T., "FFX: Fast, Scalable, Deterministic Symbolic Regression Technology," *Genetic Programming Theory and Practice IX. Genetic and Evolutionary Computation.*, Springer, New Zork, NY, 2011. doi:10.1007/978-1-4614-1770-5_13.

[11] Chen, C., Luo, C., and Jiang, Z., "Elite Bases Regression: A Real-time Algorithm for Symbolic Regression," Vol. 1, No. 2, 2017.

[12] Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Discovering governing equations from data: Sparse identification of nonlinear dynamical systems," Vol. 113, No. 15, 2015, pp. 3932–3937. doi:10.1073/pnas.1517384113.

[13] Rudy, S. H., Brunton, S. L., Proctor, J. L., and Kutz, J. N., "Data-driven discovery of partial differential equations," *Science Advances*, Vol. 3, 2017. doi:10.1126/sciadv.1602614.

[14] Apsley, A. D., and Leschziner, M. A., "A New Low-Re Non-Linear Two-Equation Turbulence Model for Complex Flows," *International Journal of Heat and Fluid Flow*, Vol. 19, No. 3, 1998, pp. 209–222.

[15] Gatski, T., and Jongen, T., "Nonlinear eddy viscosity and algebraic stress models for solving complex turbulent flows," *Progress in Aerospace SciencesAerospace Sciences*, Vol. 36, 2000, pp. 655–682.

[16] Mellen, C. P., Fröhlich, J., and Rodi, W., "Large Eddy Simulation of the flow over periodic hills," *16th IMACS World Congress,*, 2000.

[17] Jakirlic, S., "Extended excerpt related to the test case: "Flow over a periodical arrangement of 2D hills"," Tech. Rep. June, 2012.

[18] Weller, H. G., Tabor, G., Jasak, H., and Fureby, C., "A tensorial approach to computational continuum mechanics using object-oriented techniques," *Computers in Physics*, Vol. 12, No. 6, 1998, p. 620. doi:10.1063/1.168744.

[19] Breuer, M., Peller, N., Rapp, C., and Manhart, M., "Flow over periodic hills - Numerical and experimental study in a wide range of Reynolds numbers," *Computers and Fluids*, Vol. 38, No. 2, 2009, pp. 433–457. doi:10.1016/j.compfluid.2008.05.002.

[20] Hansen, P. C., *Rank-Deficient and Discrete Ill-Posed Problems*, Society for Industrial and Applied Mathematics, 1998.

[21] Bishop, and Christopher M, *Pattern Recognition and Machine Learning*, Springer, 2006.

[22] Tarantola, A., *Inverse Problem Theory*, Vol. 120, SIAM - Society for Industrial and Applied Mathematics, 2005. doi: 10.1137/1.9780898717921.

[23] Ling, J., Kurzawski, A., and Templeton, J., "Reynolds averaged turbulence modelling using deep neural networks with embedded invariance," *Journal of Fluid Mechanics*, Vol. 807, 2016, pp. 155–166. doi:10.1017/jfm.2016.615.

[24] Hellsten, A. K., "New Advanced k-w Turbulence Model for High-Lift Aerodynamics," *AIAA Journal*, Vol. 43, No. 9, 2005, pp. 1857–1869. doi:10.2514/1.13754.