Classifying Regions of High Model Error Within a Data-Driven RANS Closure Application to Wind Turbine Wakes

Steiner, Julia; Viré, Axelle; Dwight, Richard P.

DOI
10.1007/s10494-022-00346-6

Publication date
2022

Document Version
Final published version

Published in
Flow, Turbulence and Combustion

Citation (APA)
Steiner, J., Viré, A., & Dwight, R. P. (2022). Classifying Regions of High Model Error Within a Data-Driven RANS Closure: Application to Wind Turbine Wakes. *Flow, Turbulence and Combustion*, *109*(3), 545-570. https://doi.org/10.1007/s10494-022-00346-6

Flow, Turbulence and Combustion
https://doi.org/10.1007/s10494-022-00346-6

Classifying Regions of High Model Error Within a Data-Driven RANS Closure: Application to Wind Turbine Wakes

Julia Steiner[1] · **Axelle Viré**[1] · **Richard P. Dwight**[1]

Received: 1 July 2021 / Accepted: 6 July 2022

Abstract
Data-driven Reynolds-averaged Navier–Stokes (RANS) turbulence closures are increasing seen as a viable alternative to general-purpose RANS closures, when LES reference data is available—also in wind-energy. Parsimonious closures with few, simple terms have advantages in terms of stability, interpret-ability, and execution speed. However experience suggests that closure model corrections need be made only in limited regions—e.g. in the near-wake of wind turbines and not in the majority of the flow. A parsimonious model therefore must find a middle ground between precise corrections in the wake, and zero corrections elsewhere. We attempt to resolve this impasse by introducing a classifier to identify regions needing correction, and only fit and apply our model correction there. We observe that such classifier-based models are significantly simpler (with fewer terms) than models without a classifier, and have similar accuracy, but are more prone to instability. We apply our framework to three flows consisting of multiple wind-turbines in neutral conditions with interacting wakes.

1 Introduction

Aerodynamic models for wind farms are essential to optimize energy yield and turbine loading both during the design and the operational stage of wind farms. The simplest models are algebraic engineering models, and the most complex large-eddy simulations (LES) (Stevens and Meneveau 2017). While the former do not give meaningful results if strong wake interaction is present, the latter are too expensive to be used for engineering purposes (Breton et al. 2017; Ghaisas et al. 2017). Reynolds-averaged Navier–Stokes (RANS) offers a possible middle path: they require about two orders-of-magnitude less computational time than LES, while incorporating much of the physics, but they have significant accuracy deficits caused by turbulence closure modelling.

✉ Richard P. Dwight
r.p.dwight@tudelft.nl

Julia Steiner
j.steiner@tudelft.nl

Axelle Viré
a.c.vire@tudelft.nl

[1] Aerodynamics & Wind Energy, Faculty of Aerospace Engineering, TU Delft, Kluyverweg 2, 2629 HT Delft, The Netherlands

Published online: 09 August 2022

For quasi-steady atmospheric conditions around wind farms, the most commonly used RANS model is the $k-\varepsilon$ model (van der Laan et al. 2014; Rethore 2009), which has significant structural shortcomings. It over-predicts the eddy viscosity in the near wake which leads to an accelerated wake recovery; it fails to account for the effects of turbulence anisotropy (Luan and Dwight 2020; Sanderse et al. 2011); and the direct effect of the turbine on the turbulence mean quantities is not modeled (Rethore 2009). Modifications have been proposed in literature, notably anisotropy corrections (Gómez-Elvira et al. 2005); and corrections to the dissipation rate (ε) equation in the near-wake (El Kasmi and Masson 2008); and modelling the effect of the turbine with a volume forcing not local to the physical turbine (Cabezón et al. 2011). Prospathopoulos et al. (2010) compare a four separate model corrections in the context of turbines in series. In all these works improvements in mean-flow predictions are made, but these are test-case specific, require non-general tuning parameters, the models are not numerically robust, and atmospheric stratification is not considered (van der Laan and Andersen 2018). Additionally, most of these models do not directly consider the effect of actuator forcing on the turbulence equations. In our view the most successful modification proposed so far is the $k-\varepsilon-f_P$ model of van der Laan and Andersen (2018), who use an eddy-viscosity limiter that is only active in regions with high-velocity gradients. A summary of possible modifications can be found in our previous publication (Steiner et al. 2022). The perspective for further improvements by expert modellers is in our opinion limited, given the complexity of the modelling task—especially when stratification is considered.

1.1 Data-Driven Turbulence Modelling

It is for this reason that we propose data-driven modelling as an tool for devising more accurate closures. In our previous work (Steiner et al. 2020, 2022) we extended the data-driven framework *Sparse Regression of Turbulent Stress Anisotropy (SpaRTA)* first introduced by Schmelzer et al. (2019). The framework introduces two nonlinear corrections to the baseline $k-\varepsilon$ or $k-\omega$ equations: (i) an anisotropy correction and (ii) an additive correction to the transport equation for the turbulent kinetic energy. This has the benefit of separately correcting both the directionality and the magnitude of the Reynolds stress tensor (RST), as well as accounting for model-form errors in the transport equation for k. To model these corrections SpaRTA uses deterministic symbolic regression, for which the search space is constrained towards parsimonious algebraic models using sparsity-promoting regression techniques (Brunton et al. 2016; Rudy et al. 2017). The idea of obtaining simple model expressions follows the work of Weatheritt and Sandberg (2016, 2017) who pioneered Gene-Expression Programming (GEP) as a means of obtaining similar concise models.

While the SpaRTA framework worked well for our turbine wake-interaction problems (Steiner et al. 2020), leading to significantly improved wake-evolution predictions, the experience highlighted two effects. Firstly, that closure model corrections were necessary only in limited spatial regions. In particular, while the majority of the domain consisted of an undisturbed Atmospheric Boundary Layer (ABL), corrections to the baseline $k-\varepsilon$ model were needed only in the turbine wakes, and especially the near-wakes. Our SpaRTA model therefore was forced to find a middle ground between precise corrections in the wake, and zero corrections elsewhere. This lead to the second effect: the model became quite complex, involving 25 terms to represent the anisotropy correction alone. Many of these terms were observed to cancel outside the wakes, in which way the model achieved an effect localized to the wakes.

1.2 Classifiers in Turbulence Models

To resolve this issue—and the novelty of this work—we introduce a logistic classifier to the framework. A classifier is a function that yields values between zero and one, and is used here to switch the closure corrections off and on, so they are active where needed, and elsewhere the unmodified $k-\varepsilon$ model is used. This is analogous to sensors in traditional closures, which detect specific physical effects and active relevant terms locally. Our classification target is based on the magnitude of the required model correction (a quantity derived from the LES data), rather than being a manually defined target, or being based on metrics estimating the significance of RANS modelling assumptions as in Ling and Templeton (2015); Gorlé et al. (2014). As such the classifiers trained to predict this target directly pertain to the need-for-correction within the SpaRTA framework. The logistic classifier by its nature gives a smooth transition between "off" and "on", reducing spurious numerical effects due to switching. A side-effect of using a classifier is also reduced computational costs for both training and prediction (as a result of reduced data and simpler models respectively). The classifiers, and the regression models for the corrections, are based on the feature-set detailed in Wang et al. (2017); Wu et al. (2018), with the addition of a feature describing the actuator forcing used to model the wind turbines.

While classifiers have been explored in the context of RANS modelling before, e.g. Ling and Templeton (2015), this work is—to our knowledge—the first example of a learned classifier forming an integral part of a RANS closure. The classifier is itself parsimonious (based on sparse regression), straightforward to construct, and is shown to generalize well. The correction models obtained in combination with the classifier are significantly simpler (with fewer terms) than correction models based on the full field, but have similar or better accuracy in a predictive setting. One significant issue arising with our classifier-based models is that—in limited regions of high velocity-gradients, they tend to remove too much turbulence energy. We analyze the mechanism behind this instability, and introduce simple limiters to prevent these small areas from destabilizing the predictive simulations.

While there are a multitude of publications on data-driven turbulence modeling with various approaches as summarized in Duraisamy et al. (2019), only few use classifiers or markers to identify regions in the flow field with high uncertainty due to the turbulence model. Gorlé et al. (2014) developed a simple nonlinear marker for RANS simulations to identify regions in which the flow field deviates from parallel shear flow. Ling and Templeton (2015) defined three separate markers that pertain to different ways in which the Boussinesq hypothesis fails: (i) the negativity of the eddy viscosity, (ii) turbulence anisotropy, and (iii) the difference between a linear and nonlinear eddy-viscosity model prediction. The markers were derived by solving a classification problem using different supervised machine-learning approaches, namely support vector machines (SVMs), Adaboost decision trees, and random forests (RFs).

However, none of these publications integrate these markers with either a turbulence correction or more accurate turbulence models in regions with positive indication. The authors are aware of one publication publication by Longo et al. (2017) where the marker from Gorlé et al. is used to blend a LEVM model with a NLEVM in regions of non-parallel shear flow around buildings. A blending function is used to further smooth the marker properties, because the marker itself can have very sharp gradients. The general approach of selectively modifying the closure model depending on the local flow properties has parallels in the Generalized $k-\omega$ (GEKO) models of Menter et al. (2019). Both these are hand-designed models, with some similarities with the approach presented in this paper, whereas our models result from a data-driven approach targeted at a specific class of flows.

1.3 Overview

This publication is structured as follows. In Sect. 2 we specify the methodology. Additive model-form error terms within the $k-\varepsilon$ LEVM model are identified via the introduction of corrections to the stress-strain relation and the turbulence transport equations. The k-corrective-frozen-RANS approach to identify the optimal model correction is explained, and the target for the classifier is defined. The modelling of both the correction term and the classifier using an elastic net is introduced. In Sect. 3, the results of the frozen approach, the training, and cross-validation of the classifier and the correction terms, as well as the inclusion of the weighted correction terms in the flow solver are displayed. Some thoughts on numerical stability are also presented. A comparison between models derived with and without the classifier is also shown. Finally, conclusions are drawn in Sect. 4.

2 Methodology

2.1 Ground-Truth Data Generation (LES)

Our database comprised three cases: A, B and C shown in Fig. 1. All cases use the same surface roughness and an inflow velocity profile modelling a neutrally stable atmospheric boundary-layer. The cases vary in the turbine constellations. The turbine and inflow properties were taken from the wind-tunnel experiment of Chamorro and Porté-Agel (2010). Table 1 presents the parameters related to the inflow profiles, turbine dimension and domain size.

For both LES and RANS OpenFOAM-6.0 was used in conjunction with the SOWFA-6 toolbox (Churchfield and Lee 2022). For the RANS solver, a modified $k-\varepsilon$ model is the baseline closure; for the LES solver, the WALE model was used to model the unresolved

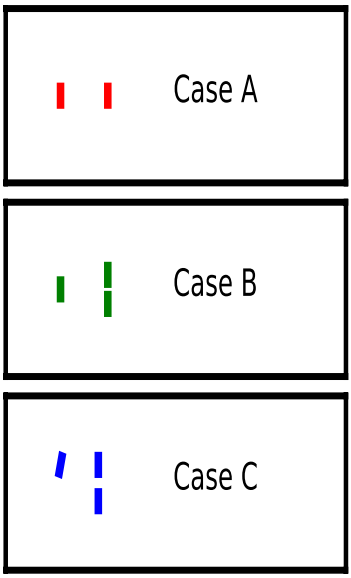

Fig. 1 Case constellation, turbine diameter is to scale

Table 1 Case setup paramters

Turbine	
Diameter	$D = 0.15\,\mathrm{m}$
Hub height	$h_{hub} = 0.125\,\mathrm{m}$
Rotation speed	$\Omega = 1190\,\mathrm{rpm}$
Inflow boundary	layer
Velocity	$U(h_{hub}) = 2.2\,\mathrm{m/s}$
Turbulence intensity	$\sigma_U(h_{hub}) = 1.0\%$
Mesh	
Domain size	$5.4 \times 1.8 \times 0.46\,\mathrm{m}^3$
Resolution	$360 \times 120 \times 64$

scales (Nicoud and Ducros 1999; Sanz Rodrigo et al. 2017). Table 2 presents the specific closure coefficients used. Actuator-disc models of the turbines are used in both LES and RANS, this avoids additional differences resulting from e.g. using actuator-line models in the LES. Details of the numerical setup such as boundary conditions and mesh resolution, as well as a validation of the CFD models with respect to wind tunnel measurements, can all be found in Steiner et al. (2022).

Throughout this work, turbulence models will be trained using the data from Case A only. LES data for Cases B and C is used exclusively for validation of the resulting models.

2.2 Optimal RANS Corrections Using the Frozen Approach

Given a reference LES field, finding a corrective field for the RANS equations is not trivial. In particular simply injecting the LES (or DNS) RST into the equations does not necessarily improve the correctness of the RANS mean-field (Thompson et al. 2016).

In this section we address this problem with the "frozen approach": we take an LES time-averaged flow-field of a statistically stationary flow, including mean velocity $U^\star$, turbulent kinetic energy $k^\star$, and Reynolds stresses $\tau_{ij}^\star$ (where an LES quantity is denoted

Table 2 Turbulence model parameters

WALE model	
C_e	0.93
C_k	0.0673
C_w	0.325
$k-\varepsilon$ model	
C_μ	0.03
$C_{\varepsilon 1}$	1.42
$C_{\varepsilon 2}$	1.92
σ_ε	1.3
σ_k	1.3

by a $\star$). If we inject these quantities into the $k-\varepsilon$ equations with appropriate BCs, the only remaining unknown is the turbulence dissipation rate ε. We can solve the ε -equation (from $k-\varepsilon$) to obtain an approximation to the dissipation rate, but the other equations (for k and U) will not be satisfied. This is due to the modelling assumptions required to obtain these equations—notably the Boussinesq assumption, and the modelling of the various terms in the k equation. As a direct consequence: solving the $k-\varepsilon$ equations can not lead us to the LES mean-solution.

If we wish to have a solution that corresponds to the LES mean flow, it is therefore necessary to modify the equations such that when LES data is injected they *are* satisfied. The way we do this is by adding spatially varying *corrective fields*, and since both the momentum equation and the k-equation are not satisfied, we need to add corrections to both these equations. We choose an additive correction $\tilde{R}(\mathbf{x})$ in the k-equation:

$$\frac{Dk^\star}{Dt} = \mathcal{P}_k^\star + \tilde{R}(\mathbf{x}) - \varepsilon + \frac{\partial}{\partial x_j}\left[\left(\nu + \nu_t/\sigma_k\right)\frac{\partial k^\star}{\partial x_j}\right], \tag{1}$$

and interpreting this as a modification to the production leads to a corresponding change in the ε equation:

$$\frac{D\varepsilon}{Dt} = \left[C_{\varepsilon 1}\left(\mathcal{P}_k^\star + \tilde{R}(\mathbf{x})\right) - C_{\varepsilon 2}\varepsilon\right]\cdot\frac{\varepsilon}{k^\star} + \frac{\partial}{\partial x_j}\left[\left(\nu + \nu_t/\sigma_\varepsilon\right)\frac{\partial \varepsilon}{\partial x_j}\right], \tag{2}$$

where the production term is known exactly from the LES data:

$$\mathcal{P}_k^\star := 2k^\star b_{ij}^\star \frac{\partial U_i^\star}{\partial x_j}. \tag{3}$$

The above coupled equations consist of a PDE for ε and an algebraic expression for $\tilde{R}$ (since all $\star$ quantities are known), and may be solved iteratively yielding the unknown fields ε and $\tilde{R}$.

We must also address the momentum equation, which we do by a correction to the Boussinesq approximation:

$$b_{ij}^\star := \frac{\tau_{ij}^\star}{2k^\star} - \frac{1}{3}\delta_{ij} = -\frac{\nu_t}{k^\star}S_{ij}^\star + \tilde{b}_{ij}^\Delta(\mathbf{x}). \tag{4}$$

so that the tensor-field $\tilde{b}_{ij}^\Delta(\mathbf{x})$ can be computed algebraically from (4) using the definition of the eddy-viscosity $\nu_t := C_\mu k^{\star 2}/\varepsilon$. The resulting fields ε, $\tilde{R}$ and $\tilde{b}_{ij}^\Delta$, satisfy the modified $k-\varepsilon$ equations with the LES data as an exact solution.

2.3 Specification of the Classification Target

The corrective fields $\tilde{b}_{ij}^\Delta(\mathbf{x})$ and $\tilde{R}(\mathbf{x})$ defined above are non-zero everywhere, but negligible in large regions of the flow. We define a single *classification target* $\tilde{\sigma}(\mathbf{x}) : \Omega \to \{0, 1\}$, a function of the spatial coordinate $\mathbf{x}$ taking the value 0 when no model correction is required, and 1 where correction is required—this assessment is based on the values of the corrective fields. Specifically we define

$$\tilde{\sigma}(\mathbf{x}) := \begin{cases} 1 \text{ if } \left(\frac{|\tilde{\mathcal{P}}_k^{\Delta}|}{|\mathcal{P}_k^{\star}|+\epsilon} > 0.2 \right) \cup \left(\frac{|\tilde{R}|}{|\mathcal{P}_k^{\star}|+\epsilon} > 0.2 \right) \\ 0 \text{ otherwise,} \end{cases} \tag{5}$$

where $\mathcal{P}_k^{\star}$ is the LES T.K.E. production, and

$$\tilde{\mathcal{P}}_k^{\Delta} := 2k^{\star} \tilde{b}_{ij}^{\Delta} \frac{\partial U_i}{\partial x_j}$$

is the extra T.K.E. production (beyond Boussinesq) due to $\tilde{b}_{ij}^{\Delta}$. This marker is thus active when the correction due to either $\tilde{R}$ or $\tilde{b}_{ij}^{\Delta}$ exceeds 20% of the reference T.K.E. production. Note that $\tilde{P}_k^{\Delta}$ may be zero, even when other measures of the tensor-valued correction b_{ij}^{Δ} are significant. This choice of classification target was made based on the observation that b_{ij}^{Δ} tends to have the biggest effect on the mean-flow via the production term.

A small threshold $\epsilon := 0.01$ is added to avoid division by zero. Finally multi-dimensional Gaussian smoothing with a filter width of two cells has been applied to all the fields in the marker expression before using them to calculate the marker. This mitigates noise in the input data.

The *implied* frozen corrections with classification then become:

$$\tilde{R}^{\sigma}(\mathbf{x}) := \tilde{\sigma}(\mathbf{x}) \cdot \tilde{R}(\mathbf{x}), \qquad \tilde{b}_{ij}^{\Delta,\sigma}(\mathbf{x}) := \tilde{\sigma}(\mathbf{x}) \cdot \tilde{b}_{ij}^{\Delta}(\mathbf{x}), \tag{6}$$

i.e. the same classification is applied to both corrections, and to all components of $\tilde{b}_{ij}^{\Delta}$. The effect is simply to switch corrections off when they fall below the threshold given in (5).

Note that the evolution of the incoming ABL does not match exactly between RANS and LES, and hence an additional correction is required (a function of wall-distance only), the details of which are identical to those of Steiner et al. (2022).

2.4 Modelling the Correction Terms and the Classifier

The objective of this section is to take the corrective fields $\tilde{b}_{ij}^{\Delta}(\mathbf{x})$ and $\tilde{R}(\mathbf{x})$, and the classification target $\tilde{\sigma}(\mathbf{x})$ (which are all functions of space), and make generalizable models for them in terms of local flow quantities available to RANS. This is the point at which the methods of supervised machine learning are valuable.

The input features we use are as comprehensive as we can achieve – later sparse regression will eliminate features that are not informative. We closely follow Wu et al. (2018), and use an integrity basis based on the set $\{\mathbf{S}, \mathbf{\Omega}, \mathbf{A}_p, \mathbf{A}_k\}$ where:

$$\mathbf{S} := \frac{1}{2}\frac{k}{\varepsilon}(\nabla U + \nabla U^T),$$

$$\mathbf{\Omega} := \frac{1}{2}\frac{k}{\varepsilon}(\nabla U - \nabla U^T),$$

$$\mathbf{A}_p := -\frac{\sqrt{k}}{\varepsilon} I \times \nabla\left(\frac{p}{\rho}\right),$$

$$\mathbf{A}_k := -\frac{\sqrt{k}}{\varepsilon} I \times \nabla k,$$

all of which are non-dimensional. We obtain a generalization of the Pope basis (Pope Nov 1975) augmented with pressure- and k-gradients, resulting in 47 scalar invariants $\mathbf{I} := [I_1, \dots, I_{47}]$. In addition we supplement the feature set with 11 non-dimensionalized physical features such as actuator forcing, $\mathbf{q} := [q_1, \dots, q_{11}]$. We use all these features when approximating both the corrective fields and the classification target. The full list of features can be found in Steiner et al. (2022).

Where approximating $\tilde{b}_{ij}^{\Delta}$, we employ the first four Pope basis tensors $T_{ij}^{(n)}$, and by construction guarantee Galilean and rotational invariance:

$$\hat{b}_{ij}^{\Delta}(\mathbf{I}, \mathbf{q}) := \sum_{n=1}^{4} T_{ij}^{(n)} \alpha_n(\mathbf{I}, \mathbf{q}). \tag{7}$$

An analogous modelling approach is taken for approximating $\tilde{R}$:

$$\hat{R}(\mathbf{I}, \mathbf{q}) := 2k \frac{\partial u_i}{\partial x_j} \left[\sum_{n=1}^{4} T_{ij}^{(n)} \beta_n(\mathbf{I}, \mathbf{q}) \right] + \varepsilon \cdot \gamma(\mathbf{I}, \mathbf{q}), \tag{8}$$

In the above $\alpha_n(\cdot)$, $\beta_n(\cdot)$ and $\gamma(\cdot)$ are arbitrary scalar-valued functions of the features.

In (8) we allow two types of terms: those that mirror a correction to the turbulence production, and those that represents a correction to the dissipation rate. Since the production term should already be correct (due to knowledge of the RST from LES), we expect most corrections to take the latter form. Nonetheless the former form allows us to capture other model-form errors - in our case notably the omission of the effect of the rotor forcing on the turbulence.

For the modelling the classification target $\tilde{\sigma}(\mathbf{x})$ the sigmoid function $s(x) := 1/(1 + \exp(-x))$ is used in conjunction with a scalar-valued function of the features, $\delta(\cdot)$:

$$\sigma(\mathbf{I}, \mathbf{q}) := s(\delta(\mathbf{I}, \mathbf{q})) \tag{9}$$

where—by construction—$\sigma \in [0, 1]$ for arbitrary $\delta(\cdot)$.

2.4.1 Sparse Regression

Any function representation can be used to parameterize the scalar-valued functions α_n, β_n, γ, and δ. We use a library approach with sparse regression (Schmelzer et al. 2019).

The $47 + 11 = 58$ input features are used to build a large library of $L \in \mathbb{N}$ candidate (basis) functions $(\ell_1, \dots \ell_L)$. This is done by recombining features with each other (up to a maximum of three features), and applying exponentiation by $\frac{1}{2}$ and 2. This results in a library cubically larger than the feature set. Each scalar function is then represented as:

$$\alpha(\mathbf{I}, \mathbf{q}) := \sum_{k=1}^{L} \theta_k \ell_k(\mathbf{I}, \mathbf{q}), \tag{10}$$

i.e. a linear representation with coefficients $\boldsymbol{\theta} \in \mathbb{R}^L$. An elastic net is then used to identify an optimal regressor with sparsity (most of the coefficients are zero) (Zou and Hastie Apr 2005).

Logistic regression is appropriate for the problem of discovering a classifier σ matching the classification target $\tilde{\sigma}$. In this case $\delta(\cdot)$ takes the form (10), and we solve:

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^L}\left\{\sum_{k=1}^{N}\log\left[\tilde{\sigma}(\mathbf{x}_k)-\sigma(\mathbf{I}(\mathbf{x}_k),\mathbf{q}(\mathbf{x}_k))\right]+\lambda\rho\|\boldsymbol{\theta}\|_1+0.5\lambda(1-\rho)\|\boldsymbol{\theta}\|^2\right\}. \tag{11}$$

where $1 \le k \le N$ indexes the mesh-points of the training data, and $\rho \in [0,1]$ and $\lambda \in \mathbb{R}^+$ control the level of sparseness, and coefficient magnitude respectively.

Given a classifier $\sigma(\mathbf{I},\mathbf{q})$ for the anisotropy correction we solve

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{4\times L}}\left\{\sum_{k=1}^{N}\sigma(\mathbf{I}_k,\mathbf{q}_k)\|\tilde{b}_{ij}^{\Delta}(\mathbf{x}_k)-\hat{b}_{ij}^{\Delta}(\mathbf{I}_k,\mathbf{q}_k)\|_F^2+\lambda\rho\|\boldsymbol{\theta}\|_1+\lambda(1-\rho)\|\boldsymbol{\theta}\|^2\right\},$$

where $\|\cdot\|_F$ is the Frobenius norm, and we have $4\times L$ coefficients in total (due to the four functions $\alpha_1,\ldots,\alpha_4$ in (7)). The presence of the multiplicative $\sigma(\cdot)$ term in the sum, limits training to locations where the (previously trained) classifier is active. The model for R is trained similarly.

Note that in the above $\boldsymbol{\theta}$ is a placeholder for the regression coefficients, and takes different values for R, b_{ij}^{Δ} and σ. Similarly the values of λ and ρ may be different for each model. The final correction models are then:

$$R(\mathbf{I},\mathbf{q}) := \sigma(\mathbf{I},\mathbf{q})\cdot\hat{R}(\mathbf{I},\mathbf{q}), \qquad b_{ij}^{\Delta}(\mathbf{I},\mathbf{q}) := \sigma(\mathbf{I},\mathbf{q})\cdot\hat{b}_{ij}^{\Delta}(\mathbf{I},\mathbf{q}),$$

analogously to (6).

The outline of the full procedure is:

1. Preprocessing: Use a mutual-information criterion to remove features that have no relationship to the target. Then build the library, and reduce it by *cliqueing* (identifying and removing clusters of multi-colinear functions).
2. Train a classifier: Match the classification target with a model expression $\sigma(\cdot)$.
3. Data reduction: Use the classifier σ from the previous step as a condition for inclusion of a point in the training dataset.
4. Model discovery: Use the elastic net to identify model forms. By varying regularization parameters λ and ρ, obtain an array of models with a variety of non-zero terms.
5. Model re-calibration: For each model from the previous step, recalibrate the non-zero model terms using Ridge regression (i.e. L_2 regularization only). Select a regularization parameter λ to encourage small coefficients.

The preprocessing step makes use of two probabilistic procedures: Mutual information (MI) (Moon et al. 1995; Ver Steeg and Galstyan 2013) and cliqueing (Alba 1973). MI can identify nonlinear relations between input features and correction terms and can hence help reduce the input feature set. Cliqueing checks if there is multi-collinearity in the input library and is thus useful for discarding co-linear input functions. Both of these procedures are vital for bringing for making the learning procedure manageable for our dataset.

3 Results and Discussion

For purposes of comparison in the following discussion we will use the LES data, as well as the predictions of the baseline $k-\varepsilon$ model, and the predictions of model we derived in Steiner et al. (2022). That latter model was obtained using the same elastic-net methodology

as used in this paper, but without the use of a classifier, and is given in Appendix A. It consists of 25 terms for b_{ij}^{Δ} and 11 terms for R. Our objective is to obtain similar accuracy and generalizability with classified-models, with significantly fewer terms.

3.1 Injection of Frozen Correction Terms

We first inject the frozen corrections of (6) into a RANS simulation. The fields $\tilde{R}^{\sigma}(\mathbf{x})$ and $\tilde{b}_{ij}^{\Delta,\sigma}(\mathbf{x})$ are calculated using the procedure of Sect. 2.2 for Case A, and are thereafter injected into a RANS simulation of the same case. This is therefore not a predictive test, but is useful for assessing the best-case scenario that can be obtained using our methodology. In the next subsection, our models for the correction terms will introduce additional errors.

Figure 2 shows the wall-normal profiles of the flow velocity and the turbulent kinetic energy (TKE) (non-dimensionalised by their values at the turbine hub height) as a function of non-dimensional height, at different stream-wise locations in the domain: from a distance of $-1D$ upstream of the first turbine T1 to a distance of $10D$ downstream of the second turbine T2, where D is the turbine diameter. The locations of the turbines and the locations of the profiles are indicated in the figure. Wake profiles are shown for the LES, the baseline RANS ($k-\varepsilon$), and the frozen RANS simulations.

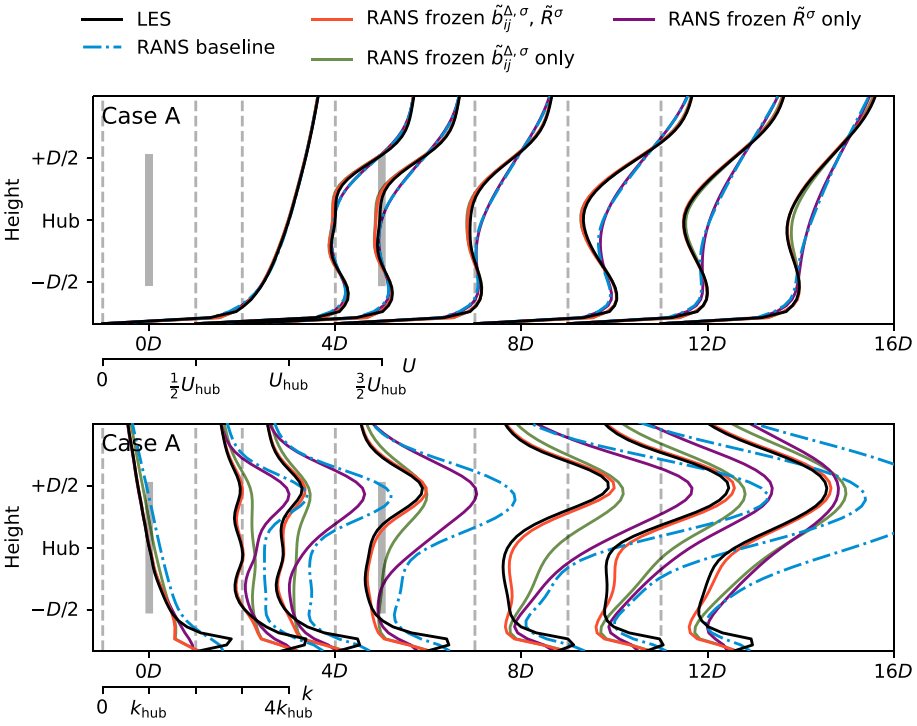

Fig. 2 Vertical velocity and T.K.E. profiles for Case A; positions of the two turbines, and the profile locations are shown in gray. Comparison of LES; RANS baseline; RANS with injected frozen corrections; and RANS with injection of $\tilde{R}^{\sigma}$ and $\tilde{b}_{ij}^{\Delta,\sigma}$ separately

The results in the figure show that—in terms of these quantities—indeed the frozen correction terms lead to an almost perfect match between the LES mean and frozen RANS velocity, and only localized mismatch of k near the wall. In particular we see that restricting the corrective fields based on the threshold of (5) has not significantly harmed the quality of the fit. On this basis we progress to discovering models for the corrections.

As an aside: we can assess the relative importance of the two corrective fields by injecting only one or the other—see Fig. 2. The anisotropy correction term $\tilde{b}_{ij}^{\Delta,\sigma}$ is significantly more important than the k-equation correction $\tilde{R}^{\sigma}$. In fact, if only a correct prediction of the velocity field is necessary, then R could be neglected completely. However, it does yield a significant improvement in the prediction of the turbulent kinetic energy, suggesting that in a predictive context it might become important.

3.2 Training of the Classifier

Based on the mutual information analysis, the feature set for the classifier was reduced to only four variables, listed in the first four rows of Table 3. Other features did not express significant correlation with the classification target. Varying the regularization parameters of the elastic net resulted in the identification of a large number of classifiers, of which five were selected for further testing based on complexity and achieved fit. These are denoted σ_1 to σ_5 in the following. The complexity of the chosen classifiers ranges from one to nine terms; and notably more much complex models do not show a significant increase in accuracy. Figure 3 is a visualization of the terms used. There is significant overlap between the terms used by the classifiers, notably turbulence intensity, velocity shear and eddy viscosity ratio are dominant in all. Note that not all features we used are Galilean invariant, notably turbulence intensity q_{TI} and actuator forcing q_F rely on a reference-frame fixed with the

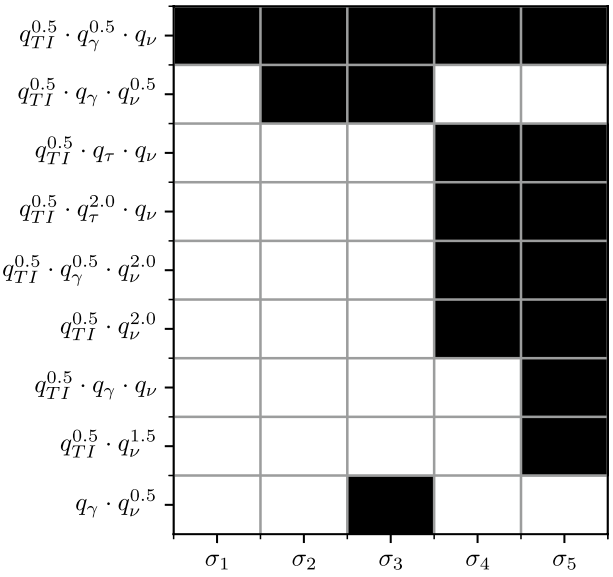

Fig. 3 Visualization of non-zero terms in the five discovered classifier models, σ_1 to σ_5. Model coefficients are not shown

Table 3 Non-dimensional features used in the discovered models of the classifiers and corrective fields. Other features not listed, while included in the model discovery process, were ultimately not part of any model

ID	Description	Expression	Normalization
q_γ	Shear parameter	$\left\|\frac{\partial U_i}{\partial x_j}\right\|$	$\frac{\epsilon}{k}$
q_τ	Ratio of total to normal Reynolds stresses	$\|\|\overline{u_i' u_j'}_{\text{Boussinesq}}\|\|$	k
q_ν	Viscosity ratio	ν_t	100ν
q_{TI}	Turbulence intensity	k	$\frac{1}{2}U_i U_i$
q_F	Actuator forcing	$\|F_{cell}\|$	$\frac{1}{2}\rho_0 A_{cell}\|U\|^2$
$q_\perp$	Nonorthogonality of U and ∇U	$\|U_i U_j \frac{\partial U_i}{\partial x_j}\|$	$\sqrt{U_l U_l U_i \frac{\partial U_i}{\partial x_j} U_k \frac{\partial U_k}{\partial x_j}}$
I_1	–	$\text{tr}\,\mathbf{S}^2$	–
I_2	–	$\text{tr}\,\mathbf{\Omega}^2$	–
I_{19}	–	$\text{tr}\,\mathbf{\Omega A}_k\mathbf{S}^2$	–
I_{25}	–	$\text{tr}\,\mathbf{A}_k^2\mathbf{S\Omega S}^2$	–
I_{35}	–	$\text{tr}\,\mathbf{A}_p\mathbf{A}_k\mathbf{S}^2$	–

ground. As such the developed models cannot be considered general-purpose, but specific to wind-farms.

In order to investigate the effect of the classifier models in isolation from the correction models, all five were implemented in the RANS solver in combination with the frozen correction terms. I.e. the corrections implemented were:

$$R(\mathbf{I},\mathbf{q}) := \sigma_l(\mathbf{I},\mathbf{q})\cdot\tilde{R}, \qquad b_{ij}^{\Delta}(\mathbf{I},\mathbf{q}) := \sigma_l(\mathbf{I},\mathbf{q})\cdot\tilde{b}_{ij}^{\Delta}, \qquad l\in\{1,\ldots,5\}.$$

and the classifier was updated at every iteration of the flow solver until the system converged.

Figure 4 shows T.K.E. profiles and the classifier fields (velocity profiles are almost identical for all classifiers and are not shown in the figure). There is minimal variation in the T.K.E. profiles, except close to the wall. The classifier values themselves show significant spread in the bottom part of the wake and towards the wall which does not seem to affect the mean-fields significantly. This could be an consequence of the corrections being generally small in the lower part of the wake. On this basis classifier σ_1 is considered the most promising candidate as it is the simplest, with a single term. We also retain σ_3, of intermediate complexity and slightly higher accuracy.

Note that the classifiers themselves might be considered not a particularly good fit to the classification target. There are two main reasons for this:

1. The classification target contains global information such as transport within the flow, whereas the the classifiers are function of local features only. As such there are regions in the lower part of the wake that cannot be effectively distinguished from regions in the upstream boundary-layer. Thus the classification problem is harder than it appears.
2. Our objective is to obtain good models that are as simple as possible—a better fit (of the classifier or the corrections) must be traded-off with model complexity. Increasing the

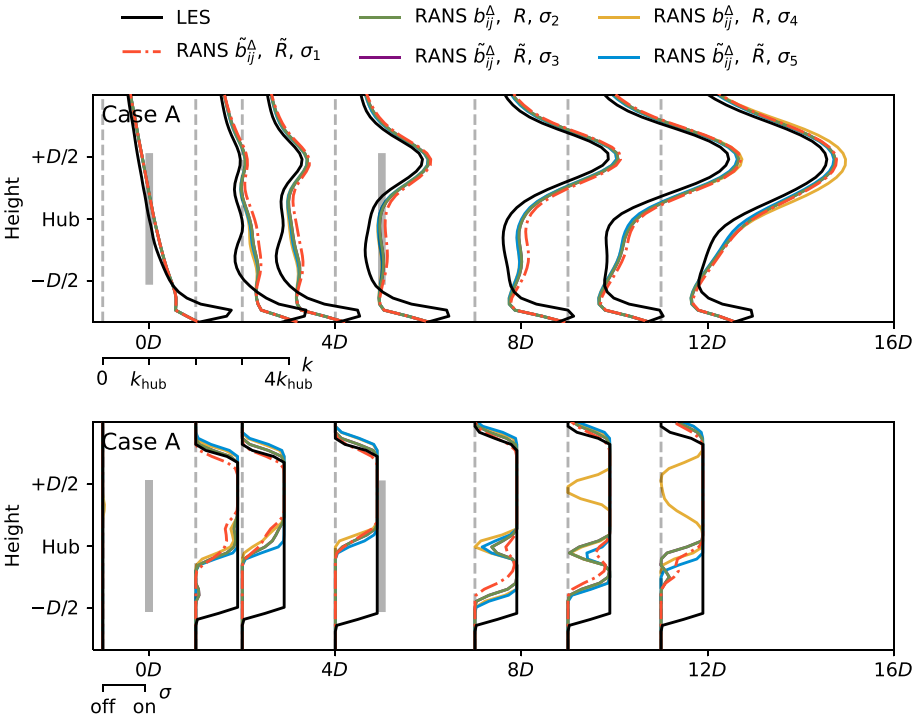

Fig. 4 LES; RANS baseline; and corrected RANS simulations of Case A. Correction terms are frozen and classifiers are coupled with the solver. The figures depict vertical slices the TKE (top) and classifier (bottom)

number of terms involved in our symbolic classifier dramatically did not significantly improve the fit, and was therefore not considered. It is perhaps the case that by using a random-forest, ANN or other very highly parameterized model, a better fit may be possible—but at the cost of significant model complexity.

3.3 Training of Correction Models

We now train correction models for the regions of the flow for which the classifier is active. The classifier is always trained first (see previous section), and used to discard points from the training data-set. As such the correction model only has to reproduce the frozen correction where the classifier is active, potentially allowing for a better fit with simpler models. Experiments with σ_1 and σ_3 showed no benefits of using the more complex σ_3, as such all models here are trained with data selected using σ_1.

3.3.1 Anisotropy Correction

All features were used in the model discovery phase resulting in a large number of candidate models. We used the procedures described in Sect. 2.4 to select a small set of promising models, and in addition we preferentially selected models without

$T^{(2)}, \ldots, T^{(4)}$ terms—i.e. linear eddy-viscosity models. Table 3 lists the features that were ultimately part of the resulting models.

Figure 5 visualizes the nonzero terms of the selected model formulations for the anisotropy correction term. We denote the models b_1^Δ to b_6^Δ in order of increasing model complexity, which ranges from 6 to 15 terms. Models b_1^Δ to b_4^Δ use only $T^{(1)}$ —making them *linear* EVMs—whereas b_5^Δ and b_6^Δ also use higher-degree base tensors, so we refer to them as non-linear EVMs. Four terms are used by all models (the first four rows of Fig. 5): these use $T^{(1)}$ combined with q_{TI} and q_F, and/or the invariant I_1. In the remaining terms, the physical features q_ν and $q_\perp$ are most often represented, followed by the invariant I_2. The most frequently used feature overall is I_1. For the non-linear eddy viscosity models a large overlap between the terms with nonlinear tensors is seen. The reference SpaRTA model without classifier is shown in the figure as "ref", and can be seem to be significantly more complex.

We assess the models for b_{ij}^Δ independently of the classifier and R models, by using the frozen $\tilde{R}$ and $\tilde{\sigma}$, and coupling only the model for b_{ij}^Δ. Figure 6 shows the effect of these partially coupled models for Case A. Again, the spread between the models is significantly larger for the k profiles than for the velocity profiles. Further, there is no significant spread between the models for the first turbine's wake, while the models differ for the second turbine—though all represent a significant improvement over the baseline model. Remarkable is is the simplest linear model b_1^Δ and the simplest nonlinear model b_6^Δ which yield the most consistent improvement over the baseline model, with the latter having a slight edge in the wake of T2.

3.3.2 T.K.E. Equation Correction

The nonzero terms of the correction models discovered for R are shown in Fig. 7, and in this case the models are denoted $R_1, \ldots, R_7$. The first observation is that terms describing a correction of the dissipation rate ϵ dominate these models, compared to terms describing a modification of production. Two terms are used in all models (the first two rows of Fig. 7): (i) a modified linear production and (ii) a dissipation associated with actuator forcing. When compared to the anisotropy correction models, the TKE production correction models use a wider range of features, most likely because this correction term is more complex and not as strongly related to velocity shear.

Once more, in Fig. 8 we visualize the effect of coupling these models with the flow solver using frozen $\tilde{b}_{ij}^\Delta$ and $\tilde{\sigma}$. The models are not distinguished in the plot in since their predictions are largely consistent with each other, with very little spread. Indeed, since this correction term mainly affects the turbulent kinetic energy, there is no visible difference in the velocity profiles. There is some spread in the turbulent kinetic energy profiles which is largest in the near wake of the second turbine. The most complex model yields the most consistent improvement over the baseline model, but the differences between models is not significant, so again we select the simplest model for further investigation.

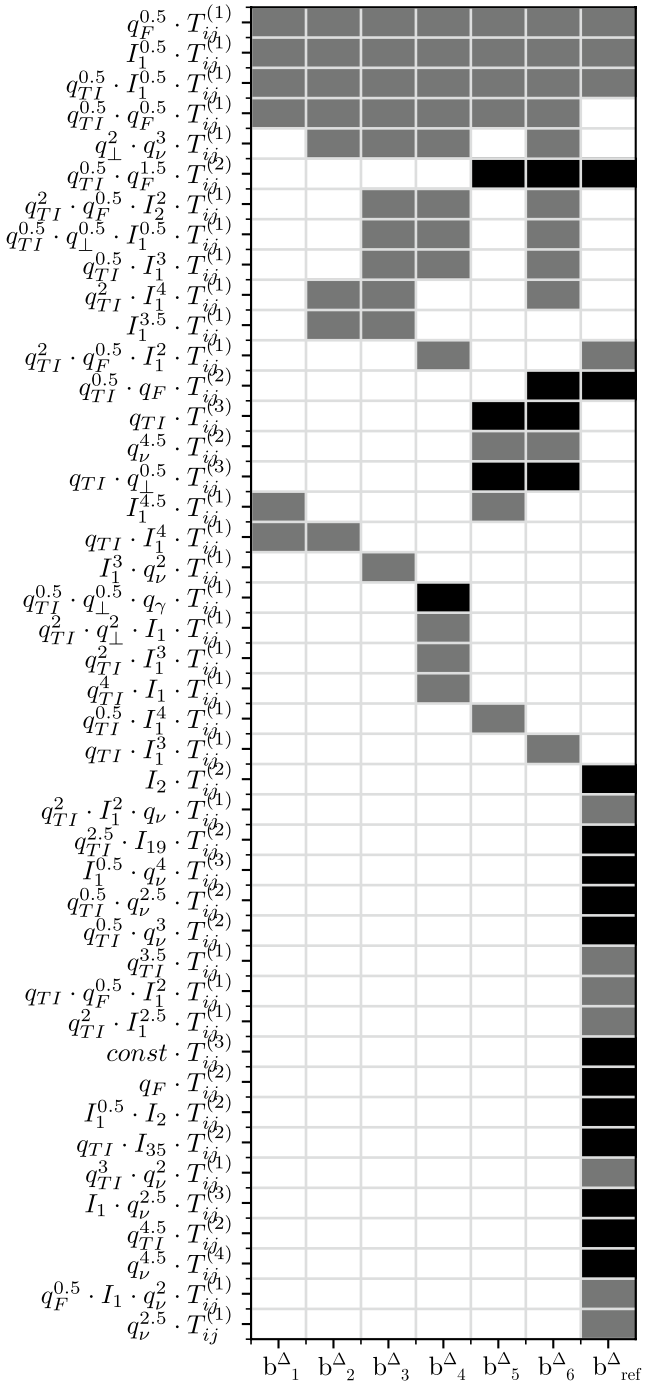

Fig. 5 Visualization of non-zero terms in the six discovered anisotropy-correction models, b_1^Δ to b_5^Δ, as well as b_{ref}^Δ from Steiner et al. (2022). Terms based on $T^{(1)}$ only are gray, while terms involving $T^{(2-4)}$ are black. The model coefficients are not shown

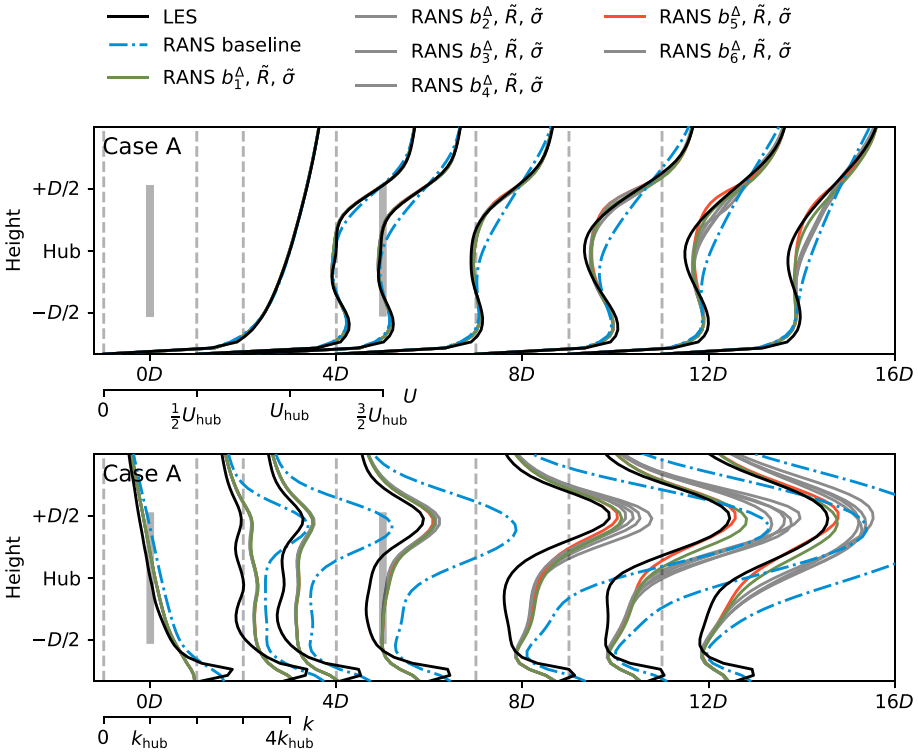

Fig. 6 LES, RANS baseline, and RANS coupled with different anisotropy correction models for Case A. The TKE production correction $\tilde{R}$ and the classifier $\tilde{\sigma}$ are frozen. Only two of the six b^Δ models are distinguished with color; the remaining are all gray (not distinguished) in order to visualize the spread of predictions

3.4 Robustness of Correction Terms

We observe the introduction of the classifier makes the data-driven SpaRTA models more prone to instabilities. NLEVMs are typically less stable than linear EVMs, and models produced by SpaRTA are no exception. However the classifier exacerbates this problem. The authors suspect this is because models derived in conjunction with classification are not required to be zero in non-wake regions. As such the models have become more sensitive to changes in the input features and tensors.

The instabilities manifest themselves as divergence of the anisotropy correction in the near wake of the turbines, and the R correction close to the rotor disk. Both manifestations are based on the same underlying effects:

1. The baseline $k-\epsilon$ model tends to over-predict k in the wake, and the actuator disc model does not remove turbulence energy from the flow. As a result in the near-wake our correction terms act mainly to remove energy.
2. The discovered models depend the shear strain invariant I_1, which increases in magnitude as shear increases.

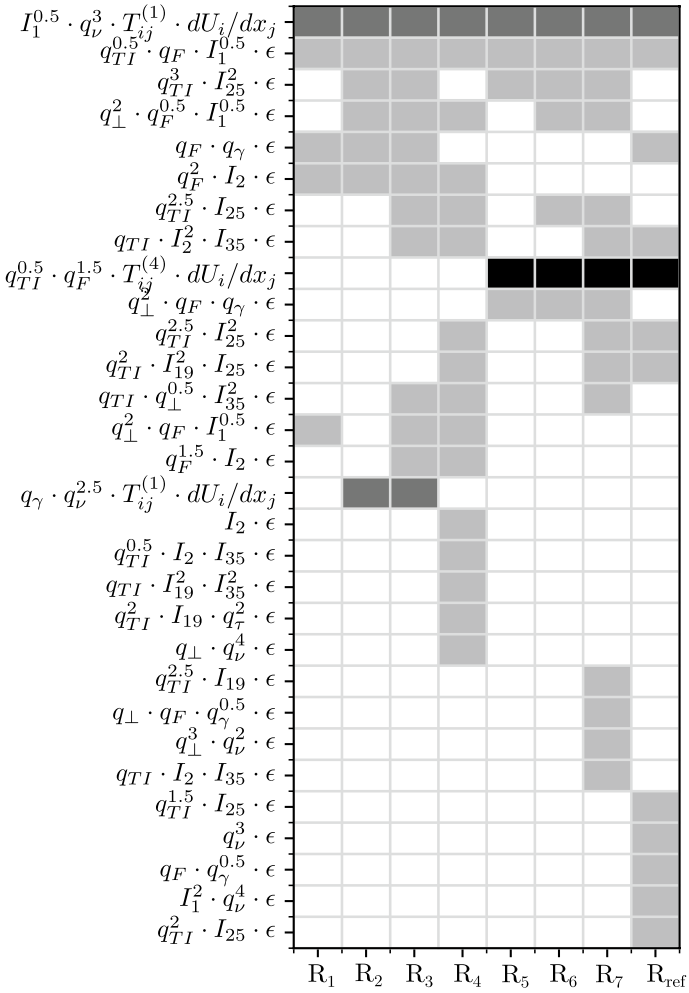

Fig. 7 Visualization of non-zero terms in the seven discovered R models, R_1 to R_7. Light-gray indicates a dissipation-rate correction; gray indicates the use of $T^{(1)}$; and black the use of $T^{(2-4)}$

As a consequence, a model removing too much energy in the near wake results in larger velocity gradients there (not seen during training), which leads to larger I_1 and increasing amounts of energy removed.

To break this positive feedback loop, in as flow-agnostic a way as possible, two limiters are proposed for the two corrective terms:

- Eddy viscosity limiter: Inspired by the k-ε-f_P model (van der Laan et al. 2013) the linear components of our anisotropy models were limited to a proportion of Boussinesq:

$$\alpha_1 = \min\left(0.8 \cdot \frac{\varepsilon}{k^2} \cdot \nu_t, \alpha_1\right). \tag{12}$$

- Form error limiter: This address tendency of the correction models to aggressively remove energy near the actuator discs. This limiter is only active in areas where

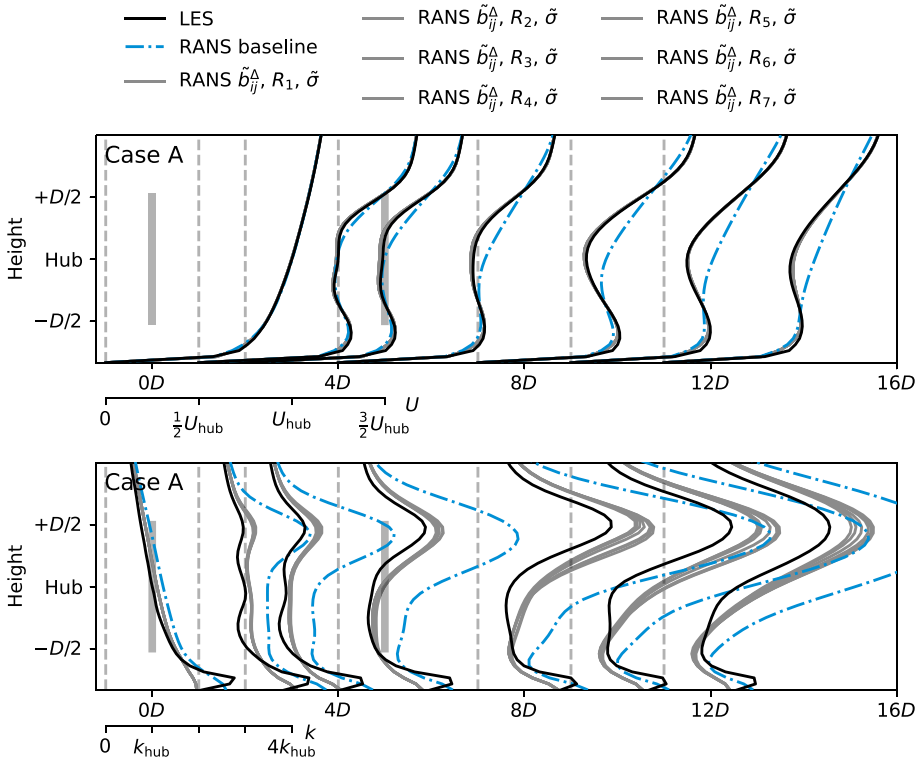

Fig. 8 LES, RANS baseline, and RANS coupled with seven different models for R for Case A. The TKE production correction $\tilde{R}$ and the classifier $\tilde{\sigma}$ are frozen. The seven R models are all gray (not distinguished) in order to visualize the spread of predictions

actuator forcing is applied, and is chosen based on the Boussinesq turbulent kinetic energy production as:

$$R = \text{sgn}(R) \cdot \min\left(0.5P_k^{\text{Boussinesq}}, |R|\right). \tag{13}$$

The thresholds were derived based on an analysis of the available data-set and are chosen sufficiently low that limiters are only active during the convergence of the solver.

3.5 Predictive Simulations

Any meaningful test of these models must be in a predictive setting on unseen cases, in this work Cases B and C. Based on the partially coupled results of Sects. 3.2 and 3.3, we select models composed of the following components:

- The simplest σ_1,
- The simplest linear b_1^{Δ} and nonlinear b_5^{Δ} models,
- The simplest R_1 and medium complexity R_4.

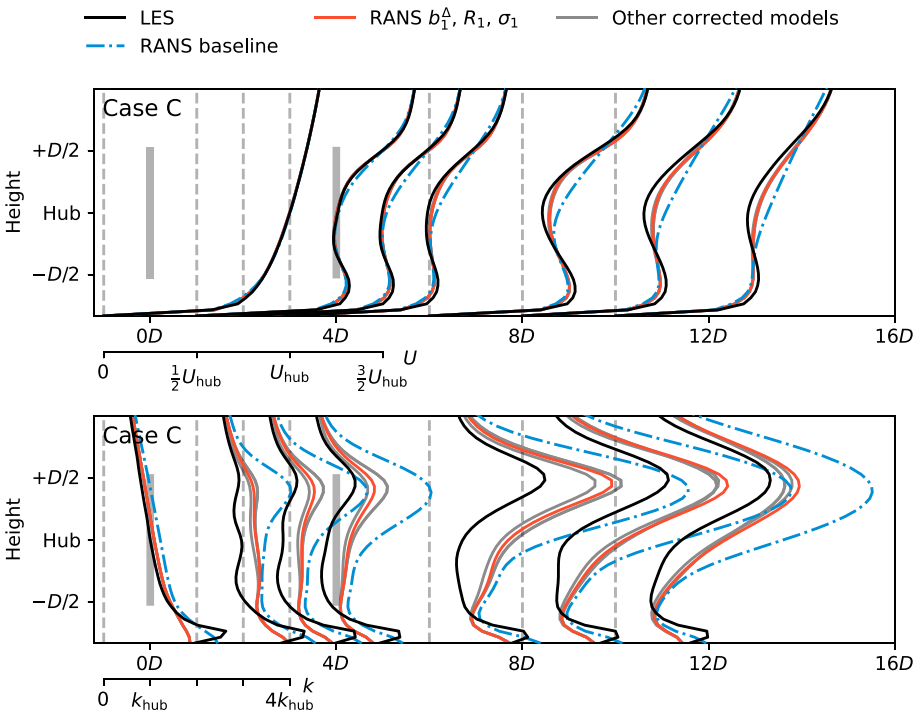

Fig. 9 LES, RANS baseline, and SpaRTA RANS models for Case C

All four model combinations were implemented in a fully coupled manner in the simulation code, and predictions compared with LES data for Cases B and C. Figure 9 shows the predictions for Case C. Additional figures for Case B (and Case A, not predictive) are presented in Appendix B and give broadly similar conclusions. All but the simplest model results are not distinguished in the figure, for the benefit of visualization.

Examining the predictions, firstly all the correction models yielded a significant improvement over the baseline $k-\varepsilon$ model. The variability in the predictions of the four models did not increase significantly beyond the variability observed in the partially coupled results. This suggests that there is no apparent strong interaction between the two correction terms which is reassuring. Given the similar performance of all four models, we again use our bias for parsimony to select a single "best" model consisting of b_1^Δ, R_1 and σ_1. Explicitly this model is:

$$\begin{aligned} b_1^\Delta := [&1.62 \cdot 10^{-1} \cdot q_{TI}^{1/2} \cdot q_F^{1/2} \\ &+4.84 \cdot 10^{-3} \cdot q_{TI}^{1/2} \cdot I_1^{1/2} \\ &-1.90 \cdot 10^{-11} \cdot q_{TI} \cdot I_1^4 \\ &+2.51 \cdot 10^{-2} \cdot q_F^{1/2} \\ &+2.00 \cdot 10^{-3} \cdot I_1^{1/2} \\ &+1.49 \cdot 10^{-15} \cdot I_1^{9/2}] \cdot T_{ij}^{(1)} \end{aligned} \tag{14a}$$

$$
\begin{aligned}
R_1 :=\ & 8.06 \cdot 10^{-5} \cdot I_1^{1/2} \cdot q_v^3 \cdot k \cdot T_{ij}^{(1)} \frac{\partial u_i}{\partial x_j} + \\
& [-2.91 \cdot 10^{1} \cdot q_{TI}^{1/2} \cdot q_F \cdot I_1^{1/2} \\
& +4.28 \cdot 10^{-1} \cdot q_\perp^2 \cdot q_F \cdot I_1^{1/2} \\
& -1.22 \cdot q_F \cdot q_\gamma \\
& +2.30 \cdot q_F^2 \cdot I_2] \cdot \epsilon
\end{aligned}
\tag{14b}
$$

$$
\sigma_1 := 1/\left(1 + \exp\left(-205.041112 \cdot q_{TI}^{1/2} \cdot q_\gamma^{1/2} \cdot q_v + 9.01862802\right)\right) \tag{14c}
$$

The magnitude of the coefficients of the terms can be misleading, because the range of magnitude of the features is quite large. For example, although the first term of R_1 has a small coefficient, it is one of the largest terms in the near wake. Similarly the tiny coefficient of the last term of b_1^Δ is counteracted by the high power of I_1, leading to this term having an effect in regions of high shear. Of course it is to be supposed that such high-powers and small coefficients may not generalize to other cases well—although they generalize to our Cases B and C here.

The ability of our procedure to produce many diverse models, all with similar performance, suggests on the one hand that procedure is reasonably robust, but on the other hand that perhaps the space of models is not being sufficiently explored. For example none of the discovered models come close to matching the frozen correction results (which themselves match LES well), even for the training Case A. It would be expected that with sufficiently many terms, an overfit model for Case A could be developed—but this was not achievable with less than 50 terms. There may be a case to be made for very highly parameterized models such as neural networks in this context.

In these results, the eddy viscosity limiter was active on average in about 1500 cells mainly in the upper part of the near wake, and the form error limiter was active on average in about 4000 cells mainly in the center of the rotor disk. Given that the total number of cells in the domain is around 3 million, the limiter is seldom used, which is deemed acceptable.

3.5.1 Comparison with Corrective Model Without Classifier

Figure 10 presents a comparison between the simplest model with classifier above, a the reference SpaRTA model discovered without classifier in Steiner et al. (2022), whose expression is given in Appendix A. In both cases the training data is Case A only, and prediction is for Case C.

There are some minor differences between predictions of the reference and the model with classifier, although overall the reference model performs slightly better. However, the reference model contains a total of 36 terms, 25 for b_{ij}^Δ and 11 for R, including terms involving $T^{(2)}, \ldots, T^{(4)}$. On the other hand our simple classifier model of (14) contains only 12 terms total—including the classifier, and does not exploit base tensors beyond $T^{(1)}$. If our premise is accepted, that—all else being equal—simple models should be preferred to complex ones, then the use of the classifier has been seen to have the potential to simplify the resulting models for similar predictive accuracy.

Admittedly, it is indeed the case that in our previous work (without classifiers), we were able to find models that worked without limiters, and that here (with classifiers) that was

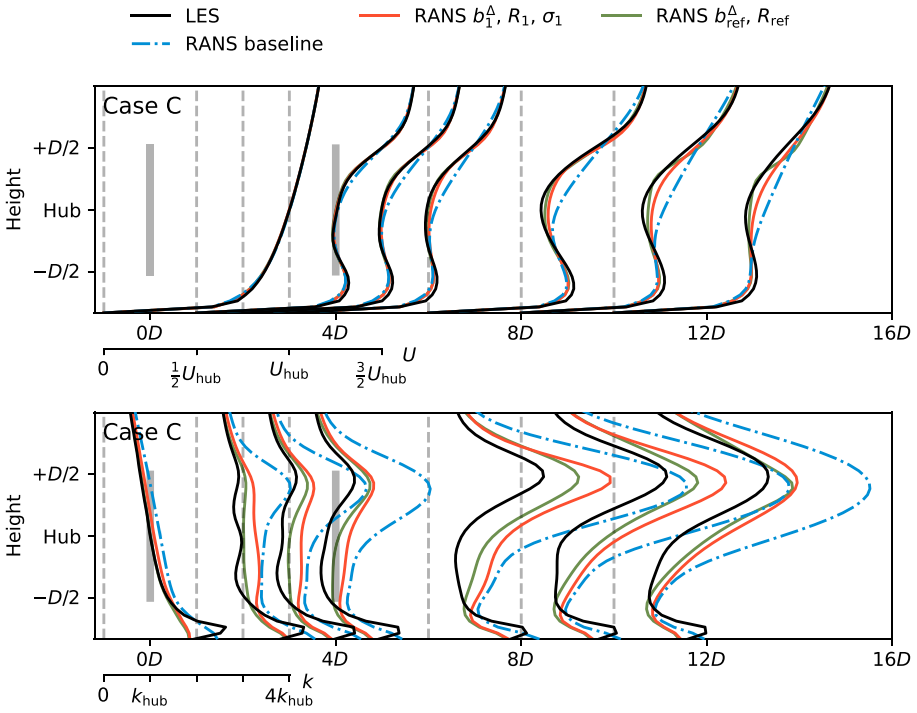

Fig. 10 LES, RANS baseline, SpaRTA model with classified, and a reference SpaRTA model without classifier for Case C

not possible. We believe that the reason for this is as follows: when training corrections without a classifier, the correction model is encouraged to predict zero, or very small, corrections in large regions of the flow—as a result it is often less aggressive everywhere. When training with a classifier, all regions of small correction have been filtered out by the classifier, and the corrector only has to match what remains. It can therefore be more aggressive, and matches the correction better with fewer terms. This aggressive fitting is what leads to the instabilities we observe. We expect to see a similar correspondence between goodness-of-fit and stability whenever the power of the fitting method increases—as we see in other work with random forests (Kaandorp and Dwight 2020), which also fit training corrections very well, and are often unstable in prediction.

To be clear: in general issues of instability are not restricted to our classifier-correction framework—in our experience they are a problem for data-driven RANS modelling in general, and our SpaRTA approach without classifier in particular. Although we were able to find stable SpaRTA models, by selecting them from the multiple models we generate, the more complex the cases studied, and in particular the more significant the corrections needed, the more difficult it becomes to find stable models. This is one of the main motivating factors behind so-called CFD-consistent modelling approaches (Holland et al. 2019; Waschkowski et al. 2022), in which the CFD code is inside the training loop, automatically making unstable models unfit. Our work attempts to achieve as much as possible *without* CFD-in-the-loop (Zhao et al. 2020) in order to build methods that are more scalable to expensive simulations and work without adjoints.

4 Conclusions

We introduced a data-driven RANS closure framework, that decomposes models for corrective terms into two stages: need for correction or not (a classifier), and magnitude of correction (regression model). This was motivated by the excellent performance of baseline RANS closures for attached flows, and their localized failure in—for example—wakes. We showed how to construct a need-for-correction measure by the frozen approach, of injecting reference data into the RANS equations. We went on to demonstrate that models resulting from this classifier-augmented approach could be made significantly simpler than similar models without classifier, while maintain similar predictive accuracy.

We also saw that the addition of the classifier has a significant drawback: it has a tendency to make the resulting models less numerically stable. We addressed this by limiting the model corrections to some proportion of the Boussinesq quantities—and made some effort to identify the source of the instabilities, which are likely the high-powers of the velocity-gradient in the discovered models. Future work must address this more thoroughly, perhaps by integration of the CFD solver into the training loop, so that unstable models are eliminated at the training stage.

The framework introduced does not rely on *prior* knowledge of where the turbulence model fails , and could be applied outside the wind-farm context investigated here . We demonstrate its power on the three-dimensional flow about multiple wind-turbines in an ABL, as a scenario where a customized turbulence closure may have engineering applications. We saw significant improvements compared to the baseline $k-\epsilon$ closure—so that RANS has the prospect of becoming a useful tool for this application in the future.

Appendix A Reference SpaRTA Model

The reference model with no classifier is taken from Steiner et al. (2022), and is:

$$\begin{aligned} R_{\text{ref}} = 2k\frac{\partial u_i}{\partial x_j}\,[\quad & 1.4771\cdot 10^{-4}\cdot I_1^{0.5}\cdot q_\nu^{3.0}\cdot \mathbf{T}_{ij}^{(1)} - 1.9183\cdot q_{TI}^{0.5}\cdot q_F^{1.5}\cdot \mathbf{T}_{ij}^{(4)}] \\ +\varepsilon\,[\quad & 1.0970\cdot 10^{1}\cdot q_{TI}^{0.5}\cdot q_F\cdot I_1^{0.5} + 6.1657\cdot 10^{-5}\cdot q_{TI}\cdot I_1^{2.0}\cdot I_{34} \\ & + 8.3864\cdot 10^{-3}\cdot q_{TI}^{1.5}\cdot I_{25} - 1.7888\cdot 10^{2}\cdot q_{TI}^{2.0}\cdot I_{25} \\ & - 1.3956\cdot 10^{1}\cdot q_F\cdot q_\gamma^{0.5} + 2.5231\cdot 10^{-7}\cdot q_{TI}^{2.5}\cdot I_{25}^{2.0} \\ & - 2.2330\cdot q_F\cdot q_\gamma - 5.2367\cdot 10^{-6}\cdot I_1^{2.0}\cdot q_\nu^{4.0} \\ & - 5.5597\cdot 10^{-2}\cdot q_\nu^{3.0},] \end{aligned} \tag{15}$$

and

$$
\begin{aligned}
b^{\Delta}_{\mathrm{ref}} = [\quad & 2.5095 \cdot 10^{-2} \cdot q_{TI}^{0.5} \cdot I_1^{0.5} + 1.090 \cdot 10^{-5} \cdot q_{TI} \cdot q_F^{0.5} \cdot I_1^{2.0} \\
& + 3.4089 \cdot 10^{-4} \cdot q_{TI}^{2.0} \cdot q_F^{0.5} \cdot I_1^{2.0} - 4.0175 \cdot 10^{-6} \cdot q_{TI}^{2.0} \cdot I_1^{2.0} \cdot q_\nu \\
& - 3.6356 \cdot 10^{-5} \cdot q_{TI}^{2.0} \cdot I_1^{2.5} + 9.6825 \cdot 10^{1} \cdot q_{TI}^{3.0} \cdot q_\nu^{2.0} \\
& - 2.8904 \cdot 10^{3} \cdot q_{TI}^{3.5} + 6.1482 \cdot 10^{-2} \cdot q_F^{0.5} \\
& - 9.4482 \cdot 10^{-5} \cdot q_F^{0.5} \cdot I_1 \cdot q_\nu^{2.0} - 2.1767 \cdot 10^{-3} \cdot q_\nu^{2.5} \\
& + 8.6126 \cdot 10^{-4} \cdot I_1^{0.5}] \cdot \mathbf{T}_{ij}^{(1)} \\
+ [& - 9.4932 \cdot 10^{-2} \cdot q_{TI}^{0.5} \cdot q_F + 1.0716 \cdot 10^{-2} \cdot q_{TI}^{0.5} \cdot q_F^{1.5} \\
& + 6.3229 \cdot 10^{-4} \cdot q_{TI}^{0.5} \cdot q_\nu^{2.5} + 6.3233 \cdot 10^{-5} \cdot q_{TI}^{0.5} \cdot q_\nu^{3.0} \\
& + 3.7871 \cdot 10^{-4} \cdot q_{TI} \cdot I_{34} + 7.5746 \cdot 10^{-4} \cdot q_{TI}^{2.5} \cdot I_{18} \\
& - 1.7673 \cdot 10^{3} \cdot q_{TI}^{4.5} + 4.8578 \cdot 10^{-3} \cdot q_F \\
& - 4.1741 \cdot 10^{-8} \cdot I_1^{0.5} \cdot I_2 + 1.3261 \cdot 10^{-6} \cdot I_1] \cdot \mathbf{T}_{ij}^{(2)} \\
+ [& - 1.3262 \cdot 10^{-3} - 2.7248 \cdot 10^{-6} \cdot I_1^{0.5} \cdot q_\nu^{4.0} \\
& + 6.5684 \cdot 10^{-7} \cdot I_1 \cdot q_\nu^{2.5}] \cdot \mathbf{T}_{ij}^{(3)} \\
& - 3.5887 \cdot 10^{-5} \cdot q_\nu^{4.5} \cdot \mathbf{T}_{ij}^{(4)} .
\end{aligned}
\tag{16}
$$

Appendix B Additional Figures: Cases A and B

Figures 11 and 12 show the results for the training data-set A. Figures 11 and 12 in the appendix show the results for the test data-set B.

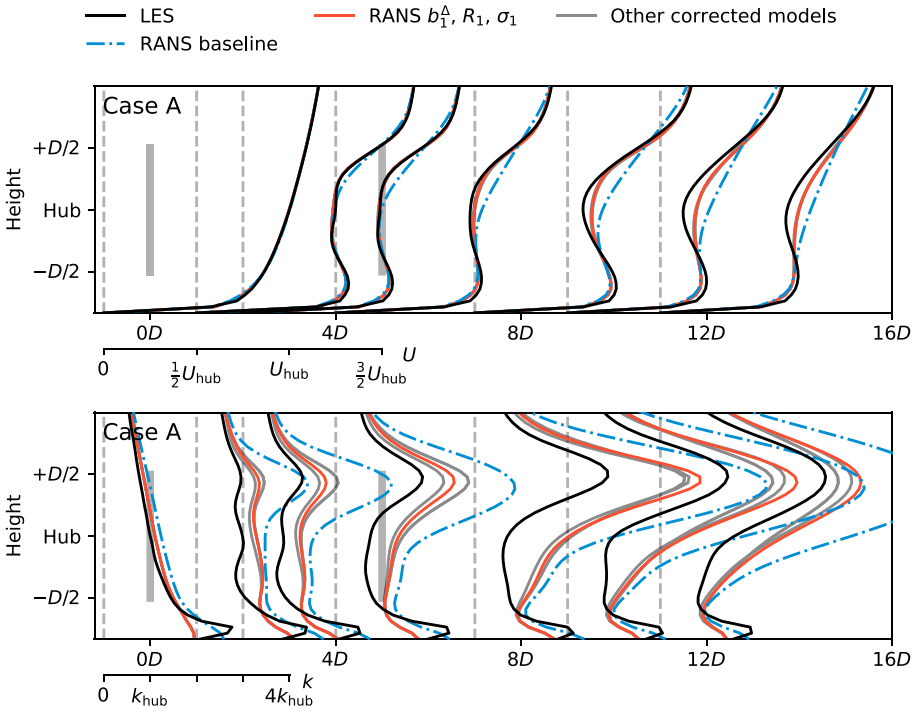

Fig. 11 LES, RANS baseline, and corrected RANS model predictions for Case A

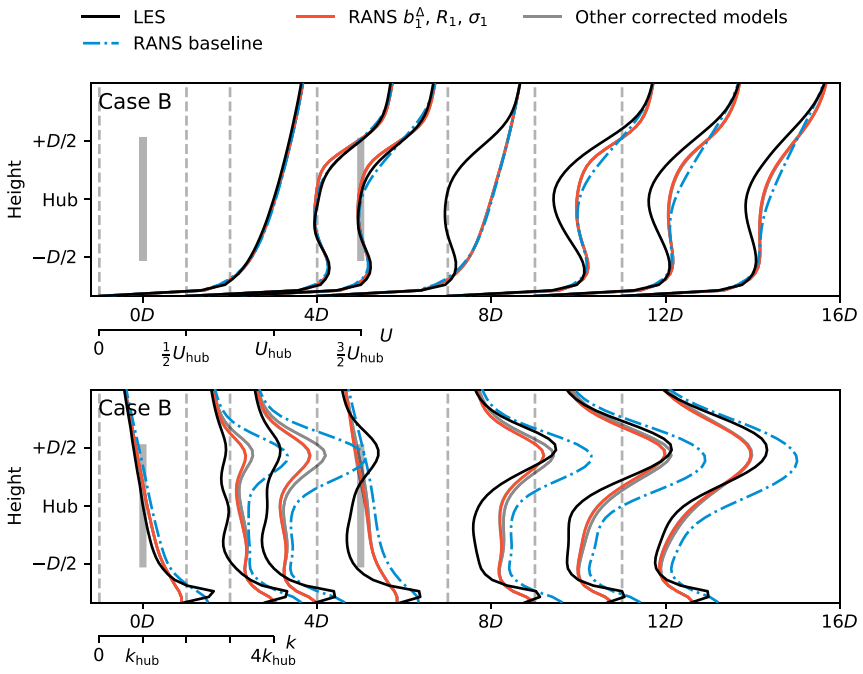

Fig. 12 LES, RANS baseline, and corrected RANS model predictions for Case B

Funding This study was funded by Rijksdienst voor Ondernemend Nederland (grant number TEHE116332).

Declarations

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

References

Alba, R.D.: A graph-theoretic definition of a sociometric clique. J. Math. Sociol. **3**, 113–126 (1973)

Breton, S.-P., Sumner, J., Sørensen, J.N., Hansen, K.S., Sarmast, S., Ivanell, S.: A survey of modelling methods for high-fidelity wind farm simulations using large eddy simulation. Philos. Trans. R. Soc. A **375**, 20160097 (2017)

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. USA **113**, 3932–3937 (2016)

Cabezón, D., Migoya, E., Crespo, A.: Comparison of turbulence models for the computational fluid dynamics simulation of wind turbine wakes in the atmospheric boundary layer. Wind Energy **14**, 909–921 (2011)

Chamorro, L.P., Porté-Agel, F.: Effects of thermal stability and incoming boundary-layer flow characteristics on wind-turbine wakes: a wind-tunnel study. Bound.-Layer Meteorol. **136**, 515–533 (2010)

Churchfield, M., Lee, S.: Simulator for wind farm aerodynamics (SOWFA). https://www.nrel.gov/wind/nwtc/sowfa.html, (2022). Accessed 5 Aug 2022

Duraisamy, K., Iaccarino, G., Xiao, H.: Turbulence modeling in the age of data. Ann. Rev. Fluid Mech. **51**, 357–377 (2019)

El Kasmi, A., Masson, C.: An extended model for turbulent flow through horizontal-axis wind turbines. J. Wind Eng. Ind. Aerodyn. **96**, 103–122 (2008)

Ghaisas, N.S., Archer, C.L., Xie, S., Wu, S., Maguire, E.: Evaluation of layout and atmospheric stability effects in wind farms using large-eddy simulation. Wind Energy **20**, 1227–1240 (2017)

Gómez-Elvira, R., Crespo, A., Migoya, E., Manuel, F., Hernández, J.: Anisotropy of turbulence in wind turbine wakes. J. Wind Eng. Ind. Aerodyn. **93**, 797–814 (2005)

Gorlé, C., Larsson, J., Emory, M., Iaccarino, G.: The deviation from parallel shear flow as an indicator of linear eddy-viscosity model inaccuracy. Phys. Fluids **26**, 051702 (2014)

Holland, J. R., Baeder, J. D., Duraisamy, K.: Field inversion and machine learning with embedded neural networks: physics-consistent neural network training, AIAA Aviation 2019 Forum, (2019)

Kaandorp, M.L., Dwight, R.P.: Data-driven modelling of the reynolds stress tensor using random forests with invariance. Comput. Fluids **202**, 104497 (2020)

Ling, J., Templeton, J.: Evaluation of machine learning algorithms for prediction of regions of high reynolds averaged navier stokes uncertainty. Phys. Fluids **27**, 085103 (2015)

Longo, R., Ferrarotti, M., Sánchez, C.G., Derudi, M., Parente, A.: Advanced turbulence models and boundary conditions for flows around different configurations of ground-mounted buildings. J. Wind Eng. Ind. Aerodyn. **167**, 160–182 (2017)

Luan, Y., Dwight, R.P.: Influence of turbulence anisotropy on RANS predictions of wind-turbine wakes. J. Phys. **1618**(6), 062059 (2020)

Menter, F. R., Matyushenko, A.: Development of a generalized $k-\omega$ two-equation turbulence model. In: Lechner, R., Dillmann, A., Heller, G., Krämer, E., Wagner, C., Tropea, C., Jakirlić, S. (Eds.), New results in numerical and experimental fluid mechanics XII, Springer International Publishing, Cham, pp. 101–109, (2019). https://doi.org/10.1007/978-3-030-25253-3_10

Moon, Y.-I., Rajagopalan, B., Lall, U.: Estimation of mutual information using kernel density estimators. Phys. Rev. E **52**, 2318–2321 (1995)

Nicoud, F., Ducros, F.: Subgrid-scale stress modelling based on the square of the velocity gradient tensor. Flow Turbul. Combust. **62**(3), 183–200 (1999)
Pope, S.B.: A more general effective-viscosity hypothesis. J. Fluid Mech. **72**, 331 (1975)
Prospathopoulos, J.M., Politis, E.S., Rados, K.G., Chaviaropoulos, P.K.: Evaluation of the effects of turbulence model enhancements on wind turbine wake predictions. Wind Energy **14**, 285–300 (2010)
Rethore, P.-E.: Wind turbine wake in atmospheric turbulence. PhD thesis, Roskilde: Riso National Laboratory for Sustainable Energy, Denmark (2009). Riso-PhD, No.53(EN)
Rudy, S.H., Brunton, S.L., Proctor, J.L., Kutz, J.N.: Data-driven discovery of partial differential equations. Sci. Adv. **3**, e1602614 (2017)
Sanderse, B., Pijl, S., Koren, B.: Review of computational fluid dynamics for wind turbine wake aerodynamics. Wind Energy **14**, 799–819 (2011)
Sanz Rodrigo, J., Churchfield, M., Kosovic, B.: A methodology for the design and testing of atmospheric boundary layer models for wind energy applications. Wind Energy Sci. **2**, 35–54 (2017)
Schmelzer, M., Dwight, R.P., Cinnella, P.: Discovery of algebraic reynolds-stress models using sparse symbolic regression. Flow Turbul. Combust. **104**, 579–603 (2019)
Steiner, J., Dwight, R., Viré, A.: Data-driven turbulence modeling for wind turbine wakes under neutral conditions. J. Phys. **1618**, 062051 (2020)
Steiner, J., Dwight, R.P., Viré, A.: Data-driven RANS closures for wind turbine wakes under neutral conditions. Comput. Fluids **233**, 105213 (2022)
Stevens, R.J., Meneveau, C.: Flow structure and turbulence in wind farms. Ann. Rev. Fluid Mech. **49**, 311–339 (2017)
Thompson, R.L., Sampaio, L.E.B., de Bragança Alves, F.A., Thais, L., Mompean, G.: A methodology to evaluate statistical errors in dns data of plane channel flows. Comput. Fluids **130**, 1–7 (2016)
van der Laan, M. P., Sørensen, N., Réthoré, P.-E., Mann, J., Kelly, M., Schepers, J.: Nonlinear eddy viscosity models applied to wind turbine wakes, in Proceedings for the ICOWES2013, Copenhagen, Denmark, 12, pp. 514–525 (2013)
van der Laan, M.P., Andersen, S.J.: The turbulence scales of a wind turbine wake: A revisit of extended k-epsilon models. J. Phys. **1037**, 072001 (2018)
van der Laan, M.P., Sørensen, N.N., Réthoré, P.-E., Mann, J., Kelly, M.C., Troldborg, N., Schepers, J.G., Machefaux, E.: An improved $k-\varepsilon$ model applied to a wind turbine wake in atmospheric turbulence. Wind Energy **18**, 889–907 (2014)
Ver Steeg, G., Galstyan, A.: Information-theoretic measures of influence based on content dynamics. Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13, (2013)
Wang, J.-X., Wu, J.-L., Xiao, H.: Physics-informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on dns data. Phys. Rev. Fluids **2**, 034603 (2017)
Waschkowski, F., Zhao, Y., Sandberg, R., Klewicki, J.: Multi-objective cfd-driven development of coupled turbulence closure models. J. Comput. Phys. **452**, 110922 (2022)
Weatheritt, J., Sandberg, R.: A novel evolutionary algorithm applied to algebraic modifications of the rans stress-strain relationship. J. Comput. Phys. **325**, 22–37 (2016)
Weatheritt, J., Sandberg, R.: The development of algebraic stress models using a novel evolutionary algorithm. Int. J. Heat Fluid Flow **68**, 298–318 (2017)
Wu, J.-L., Xiao, H., Paterson, E.: Physics-informed machine learning approach for augmenting turbulence models: a comprehensive framework. Phys. Rev. Fluids **3**, 074602 (2018)
Zhao, Y., Akolekar, H.D., Weatheritt, J., Michelassi, V., Sandberg, R.D.: Rans turbulence model development using cfd-driven machine learning. J. Comput. Phys. **411**, 109413 (2020)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. **67**, 301–320 (2005)