Theoretical and Applied Mechanics Letters 13 (2023) 100414

Contents lists available at ScienceDirect

Theoretical and Applied Mechanics Letters

journal homepage: www.elsevier.com/locate/taml

Letter

Linear logistic regression with weight thresholding for flow regime classification of a stratified wake

Xinyi L.D. Huang*, Robert F. Kunz, Xiang I.A. Yang

Department of Mechanical Engineering, Penn State University, University Park, 16802, PA, USA

ARTICLE INFO

Article history:
Received 25 July 2022
Revised 27 September 2022
Accepted 6 November 2022
Available online 21 November 2022

Keywords:
Stratified wake
Classification
Supervised learning
Full Reynolds stress modelling

ABSTRACT

A stratified wake has multiple flow regimes, and exhibits different behaviors in these regimes due to the competing physical effects of momentum and buoyancy. This work aims at automated classification of the weakly and the strongly stratified turbulence regimes based on information available in a full Reynolds stress model. First, we generate a direct numerical simulation database with Reynolds numbers from 10,000 to 50,000 and Froude numbers from 2 to 50. Order (100) independent realizations of temporally evolving wakes are computed to get converged statistics. Second, we train a linear logistic regression classifier with weight thresholding for automated flow regime classification. The classifier is designed to identify the physics critical to classification. Trained against data at one flow condition, the classifier is found to generalize well to other Reynolds and Froude numbers. The results show that the physics governing wake evolution is universal, and that the classifier captures that physics.

Thermal stratification is common in geophysical and oceanic environments and has received much attention in the fluid dynamics literature [1–6]. Figure 1 shows a sketch of a wake in a stratified environment. Such stratified wakes are found behind underwater vehicles. The wake is controlled by the bulk Reynolds number $Re_B = U_B D/\nu$ and the bulk Froude number $Fr_B = U_B/ND$, where D measures the size of the wake-generating object, U_B is the velocity of the wake-generating object (relative to the ambient fluid), ν is the kinematic viscosity, and N is the Brunt-Väisälä frequency. The wake is characterized by the competing effects of buoyancy, inertia, and viscosity, and the wake can be put into the "weakly stratified turbulence" (WST) regime, the "intermediately stratified turbulence" (IST) regime, and the "strongly stratified turbulence" (SST) regime [7–10]. In the WST regime, the turbulence is roughly isotropic. In the SST regime, the turbulence is highly anisotropic, and large-scale pancake-like structures emerge, see Fig. 1d. The wake may also be divided into the near-wake (NW) regime, the non-equilibrium (NEQ) regime, and the quasi-two-dimensional (Q2D) regime based on the behavior of the centerline velocity deficit [11]. We will concern ourselves with RANS modeling, i.e., the modeling of turbulence. Hence, the WST, IST, and SST regime classification is more relevant than the NW, NEQ, and Q2D regime classification, the latter of which concerns the behavior of the large-scale motions.

Consider the wake behind an underwater vehicle, for which $D = 10$ m, $U_b = 10$ m/s, and $\nu = 10^{-6}$ m^2/s. The bulk Reynolds number is 10^8. Direct numerical simulations (DNSs) and large-eddy simulations (LESs) at these Reynolds numbers are prohibitively costly [12–14]. Consequently, a naval engineer has to resort to cost-effective Reynolds-averaged Navier Stokes (RANS) simulations, where turbulence is modeled. The flows in the WST and the SST regimes require different RANS modeling strategies [15]. To apply different modeling strategies in different flow regimes, one must first identify and classify these flow regimes, which is the topic of this work.

Data-enabled methods have received much attention in the recent fluid dynamics literature [16–20]. Here, we attempt a data-based autonomous classification of the WST and the SST regimes. Autonomous flow regime classification was attempted in Refs. [21,22]. There, the developed classifiers were able to classify the flow regimes based on a dynamic mode decomposition of the instantaneous flow. However, because these classifiers require instantaneous flow information, they are not suitable for RANS applications.

We propose linear logistic regression with weight thresholding (LLR-WT). The intended application of the classifier is full Reynolds stress model (FRSM) of stratified wake flows [23]. Here, logistic regression is preferred to support vector machine, k-nearest neighbor

* Corresponding author.
E-mail address: xuh128@psu.edu (X.L.D. Huang).

https://doi.org/10.1016/j.taml.2022.100414

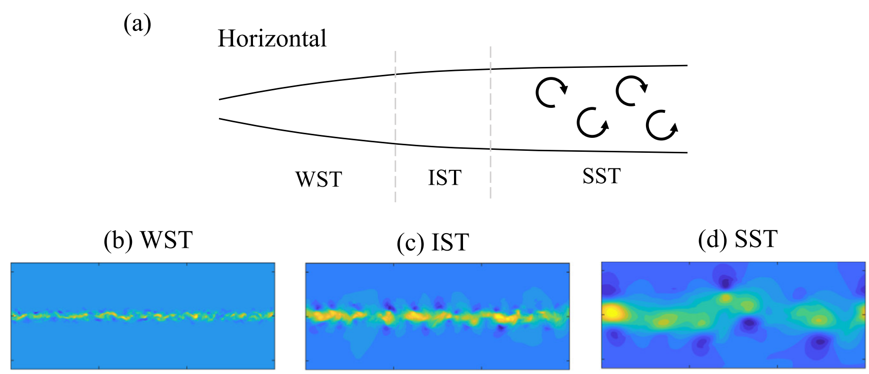


Fig. 1. (**a**) A sketch of a stratified wake in the horizontal plane. Large-scale pancake-like structures emerge in the SST regime [11]. (**b, c, d**) Contours of the streamwise velocity deficit at the center horizontal plane. (**b**) WST regime, (**c**) IST regime, and (**d**) SST regime. Here, yellow corresponds to high velocity deficit and blue corresponds to low velocity deficit. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

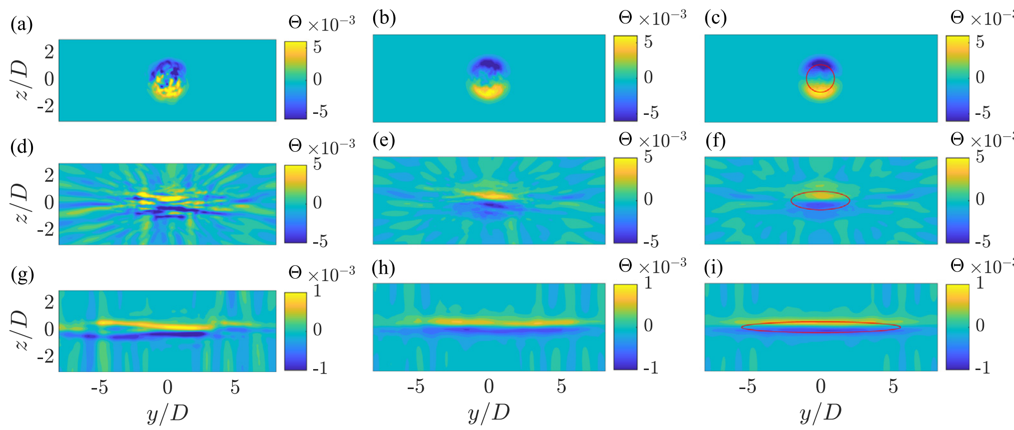


Fig. 2. The contours of the temperature field for the case R20F02. (**a,d,g**) 1 ensemble, (**b,e,h**) 10 ensemble, and (**c,f,i**) 100 ensemble. (a,b,c) $Nt = 4$, (d,e,f) $Nt = 46$, (g,h,i) $Nt = 400$, which fall in the WST regime, the IST regime and the SST regime, respectively. The red circles enclose the core of the wake. The horizontal and vertical sizes of the wake core L_{Hk} and L_{Vk} measure the distance between the center of the wake to the vertical and horizontal location where the turbulent kinetic energy decays to half of the maximum value [10]. Here, we show only part of the computational domain for presentation purposes.

method, neural network, and other machine learning tools because of its simplicity. We note that a RANS model is evaluated on every grid point, and simplicity is critical to the success of the model. FRSM is one type of RANS modeling. It solves the transport equations of the Reynolds stresses and handles flow anisotropy more competently than eddy viscosity models [24]. Considering the intended application, information not available in an FRSM, e.g., instantaneous flow field, is not invoked. Instead, the classifier relies on the terms in the Reynolds-averaged momentum equations and the Reynolds stress transport equations for classification. This gives rise to 129 terms, each corresponding to a specific physical process. By applying weight thresholding, the classifier identifies a few terms that are critical to the classification task. We will show that the trained classifier is able to extrapolate to conditions outside the training dataset.

First, we will generate a DNS database for training and testing. DNS resolves all the scales in a turbulent flow and therefore is very accurate [25]. It has been extensively used in our research [26–30]. We follow Refs. [31–33] and compute a temporally evolving stratified wake in a rectangular box. The box is periodic in the two horizontal directions. The wake at time t corresponds to a spatially evolving wake at a downstream distance $x = U_B t$. A temporal simulation is one realization of the wake at all simulated time instants. The previous DNS studies computed one realization [31,34–36]. However, one realization does not lead to converged statistics. Figure 2 shows the contours of the temperature at $Nt = 4$, 46, and 400 for $Re_B = 20000$ and $Fr_B = 2$. We compare statistics averaged among 1, 10, and 100 realizations. It should be clear that ensemble average is needed to get converged statistics.

Table 1 lists the details of the DNSs. Five combinations of Re_B and Fr_B are considered. The bulk Reynolds number ranges from $Re_b = 10000$ to 50000. The bulk Froude number ranges from 2 to 50. For each combination of Re_B and Fr_B, 80 to 100 realizations are computed. The nomenclature is R[Re_B/1000]F[Fr_B]. The domain size is such that the boundaries in the transverse and the vertical directions do not affect the wake evolution [33,35]. The mesh is such that the grid spacing is nowhere larger than 2 to 4 Kolmogorov length scales [31]. As the wake decays, the Kolmogorov length scale increases. We coarsen the grid resolution and expand the computational domain in the transverse and the vertical directions at time t_f to accommodate the increased Kolmogorov length scale and the expanded wake. As a result, our DNSs have multiple stages, as indicated by the superscripts. Our setup is similar to that in Ref. [31]., and further details of the DNSs including the initial condition, the time stepping, the boundary condition and the grid can be found there. The code we use is the second-order finite-difference incompressible Navier-Stokes solver AFiD. Details of the code can be found in Refs. [37–39].

To facilitate the discussion in the next parts, we summarize the flow phenomenology. Figure 3 shows the width L_{Hk} and the height L_{Vk} of the wake as a function of time, where L_{Hk} and L_{Vk} measure the distances between the center of the wake to the vertical and horizontal locations where the turbulent kinetic energy decays to half of its maximum value. These two length scales have been ex-

Table 1
Details of the DNSs. The nomenclature of the cases is R[Re_B]F[Fr_B]. L_x, L_y, and L_z are the domain sizes in the three Cartesian directions. N_x, N_y, and N_z are the number of grid points in the three Cartesian directions. Five cases are regridded at time t_f, giving rise to multiple stages, as indicated by the superscripts. The temporal simulations are repeated 80 to 100 times, as indicated by "Ensemble" in the Table.

Case	Re_B	Fr_B	L_x/D	L_y/D	L_z/D	N_x	N_y	N_z	t_fU_B/D	Ensemble
R10F02	10 000	2	60	38	14	1280	768	256	1380.18	100
R20F02^1	20 000	2	60	24	12	1280	768	384	109.29	100
R20F02^2	20 000	2	60	36	12	1280	576	192	800.62	100
R20F10^1	20 000	10	60	24	12	1280	768	384	109.29	80
R20F10^2	20 000	10	60	36	12	1280	576	192	1220.58	80
R20F50^1	20 000	50	60	24	12	1280	768	384	109.29	80
R20F50^2	20 000	50	60	36	12	1280	576	192	568.91	80
R20F50^3	20 000	50	60	48	24	640	384	192	5232.38	80
R50F50^1	50 000	50	60	8.25	8.25	2560	550	550	67.46	80
R50F50^2	50 000	50	60	23.04	11.52	1280	768	384	265.72	80
R50F50^3	50 000	50	60	46.08	23.04	640	512	256	5800.23	80

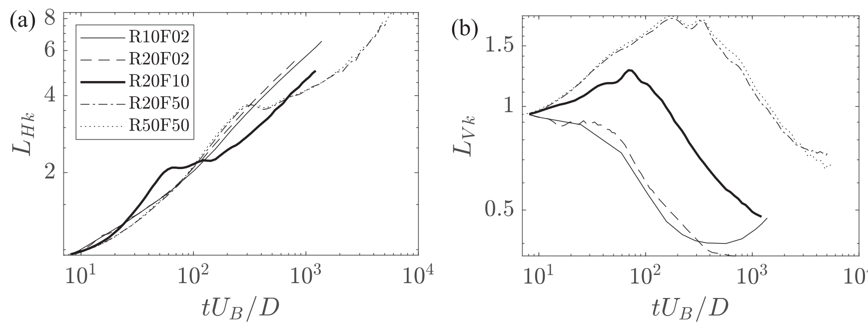


Fig. 3. Time history of the wakes' (**a**) horizontal sizes L_{Hk} and (**b**) vertical sizes L_{Vk}.

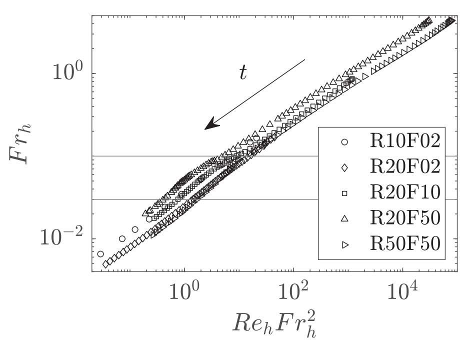


Fig. 4. Trajectories of the DNSs in the $Re_hFr_h^2 - Fr_h$ space. The arrow indicates the direction of time t. The two horizontal lines are at $Fr_h = 0.03$, and 0.1. We use different symbols for different cases.

tensively used to measure the size of a stratified wake [10,40,41]. We see that while L_{Hk} keeps increasing, L_{Vk} stops increasing in the late wake. Figure 4 shows the trajectories of the DNSs in the $Re_hFr_h^2 - Fr_h$ space, where $Re_h = u'_hL_{Hk}/\nu$ and $Fr_h = u'_h/NL_{Hk}$ are the local Reynolds number and the local Froude number, and u'_h is the center horizontal velocity fluctuation. $Re_hFr_h^2$ measures the ratio of the convective term to the viscous term. Fr_h measures the ratio of the convective term to the buoyancy term. Large Fr_h ($Fr_h \gtrsim 0.1$) corresponds to the WST flow regime, and small Fr_h ($Fr_h \lesssim 0.03$) corresponds to the SST flow regime. We see that all simulations continue well into the SST regime.

Having collected the training data, we now train a linear logistic regression classifier that predicts if a sample belongs to the SST regime or the WST regime according to:

$$\sigma(\eta) = \frac{1}{1+\exp(-\eta)}, \quad \eta = \sum_{j=1}^{N} \omega_j X_j + b, \tag{1}$$

where $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_N)$ is the weight vector, b is the bias, $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ is the input vector, N is the number of input features and σ is the sigmoid function. A sample that gives $\sigma \to 1$ is in the SST regime, and a sample that gives $\sigma \to 0$ is in the WST regime. Note that it makes no difference if we assign 1 to the WST regime and 0 to the SST regime. The weights are obtained by minimizing the following cross-entropy loss function

$$L_{\mathrm{CE}}(\hat{\mathbf{Y}}, \mathbf{Y}) \equiv -\sum_{i=1}^{n}\left[Y^{(i)}\log\hat{Y}^{(i)} + (1-Y^{(i)})\log(1-\hat{Y}^{(i)})\right] + \frac{C_0}{2}\sum_{j=1}^{N}\omega_j^2, \tag{2}$$

where the subscript CE is short for "cross entropy", Y is the labeled data, $\hat{Y} = \sigma(\eta)$ is the classifier output, the superscript (i) indexes the training sample, n is the number of training samples, the last term is the regularization term, and the coefficient $C_0 = 0.01$ is obtained through cross-validation (see the appendix for details).

The decision boundary of the above linear logistic regression classifier is given by

$$\sum_{j=1}^{N} \omega_j X_j + b = 0. \tag{3}$$

The classifier predicts $Y = 1$ if $\sum_{j=1}^{N} \omega_j X_j + b \geq 0$ and $Y = 0$ if $\sum_{j=1}^{N} \omega_j X_j + b < 0$ [42]. Equation (3) corresponds to a hyperplane in the input variable space. Note that this hyperplane is unique as long as the inputs are not linearly dependent because of the convexity of the cost function [43]. The ith weight ω_i measures the importance of the ith input X_i in the classification task. Figure 5 shows three decision boundaries in a 2D input space. We see that the first and the second inputs in Fig. 5a and 5b are not critical to the classification task. In Fig. 5a, the first input may take small or large values irrespective of the samples' classes; in Fig. 5b, the second input is small irrespective of the samples' classes.

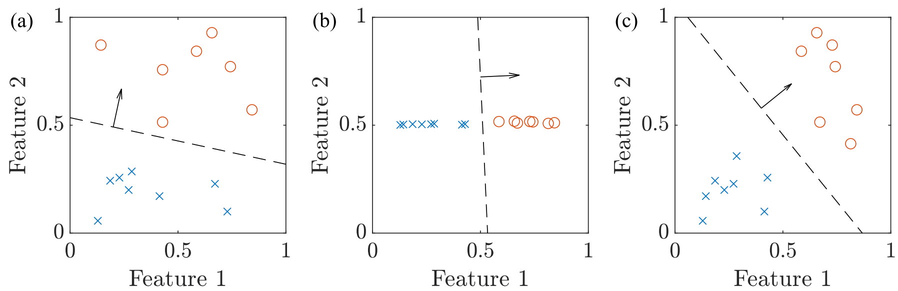


Fig. 5. Examples of binary classification in a 2D input space. The normal vectors are (0.2108 0.9775), (0.9990 0.0458), and (0.7775 0.6289) in (**a, b, c**). Class 0 is indicated by the blue cross while class 1 is indicated by the orange circle.

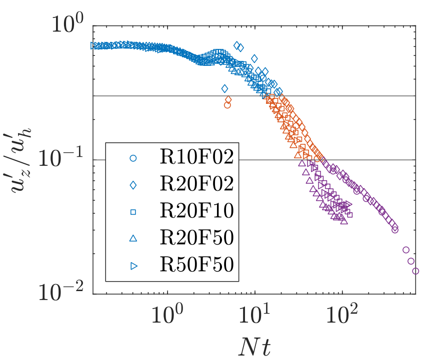


Fig. 6. u'_z/u'_h as a function of the non-dimensionalized time. The solid lines are at $u'_z/u'_h = 0.3$ and $u'_z/u'_h = 0.1$. Different cases are represented by different symbols, and the WST regime, the IST regime, and the SST regime are colored blue, orange, and purple, respectively.

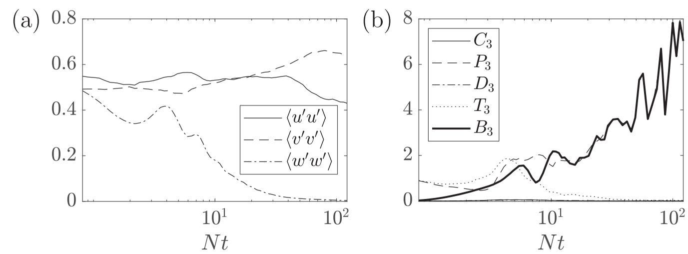


Fig. 7. (**a**) Input features $\langle u'u' \rangle$, $\langle v'v' \rangle$, and $\langle w'w' \rangle$ as a function of non-dimensional time in the case R20F10. (**b**) Input feature C_3, P_3, D_3, T_3, and B_3 as a function of non-dimensional time. These are the terms in the (Reynolds averaged) transport equation of the vertical momentum.

In these two examples, the normal vectors are (0.2108 0.9775) and (0.9990 0.0458), and the first and the second inputs have comparably small weights. In Fig. 5c, both inputs are critical to the classification task: the first input is more likely to take large values when the sample is in class 1, and the second input is more likely to take small values when the sample is in class 0. The normal vector is (0.7775 0.6289), and both inputs have large weights. In our training, we will apply "weight thresholding" and discard inputs that have small weights—these features are non-essential to the classification task. Discarding these non-essential features prevents overfitting and meanwhile improves computational efficiency. Specifically, we discard a feature if its weight is less than $0.1\max_j |\omega_j|$. The remaining features, which are terms in the momentum and Reynolds stress transport equations, are essential to the classification. We use these features as inputs and train the classifier. The above process gives rise to "linear logistic regression with weight thresholding" (LLR-WT). Here, principle component analysis is not pursued. That is, we do not try to reduce the dimension of the input feature space by combining two or more features linearly. We chose not to do principle component analysis because the input features here bear clear physical meanings, which would be lost when the features are combined and mixed.

Table 2
All input features.

Norm	Maximum of absolute value	Maximum - minimum
Computed	U, $\langle u'u' \rangle$, $\langle v'v' \rangle$, $\langle w'w' \rangle$, $\langle \theta'\theta' \rangle$, k, C_1, P_1, D_1, T_1, $\mathcal{C}_{11}, \mathcal{C}_{22}, \mathcal{C}_{33}$, $\mathcal{P}_{11}, \mathcal{P}_{22}, \mathcal{P}_{33}$, $\mathcal{D}_{v,11}, \mathcal{D}_{v,22}, \mathcal{D}_{v,33}$, $\mathcal{B}_{33}$, $\mathcal{C}_{\theta\theta}, \mathcal{P}_{\theta\theta}, \mathcal{D}_{v,\theta\theta}$	V, W, Θ, P, $\langle u'v' \rangle$, $\langle u'w' \rangle$, $\langle v'w' \rangle$, $\langle u'\theta' \rangle$, $\langle v'\theta' \rangle$, $\langle w'\theta' \rangle$, $C_2, C_3, P_2, P_3, D_2, D_3, T_2, T_3$, $B_3, C_\theta, D_\theta, T_\theta$, $\mathcal{C}_{12}, \mathcal{C}_{13}, \mathcal{C}_{23}, \mathcal{P}_{12}, \mathcal{P}_{13}, \mathcal{P}_{23}$, $\mathcal{D}_{v,12}, \mathcal{D}_{v,13}, \mathcal{D}_{v,23}, \mathcal{B}_{13}, \mathcal{B}_{23}$, $\mathcal{C}_{1\theta}, \mathcal{C}_{2\theta}, \mathcal{C}_{3\theta}, \mathcal{P}_{1\theta}, \mathcal{P}_{2\theta}, \mathcal{P}_{3\theta}$, $\mathcal{B}_{1\theta}, \mathcal{B}_{2\theta}, \mathcal{B}_{3\theta}$
Modeled	$\langle p'p' \rangle$, ε, $\varepsilon_{11}, \varepsilon_{22}, \varepsilon_{33}$, $\Phi_{11}, \Phi_{22}, \Phi_{33}$, $\mathcal{D}_{p,11}, \mathcal{D}_{p,22}, \mathcal{D}_{p,33}$, $\mathcal{D}_{T,11}, \mathcal{D}_{T,22}, \mathcal{D}_{T,33}$, $\varepsilon_{\theta\theta}, \mathcal{D}_{T,\theta\theta}$	$\langle u'p' \rangle$, $\langle v'p' \rangle$, $\langle w'p' \rangle$, $\langle \theta'p' \rangle$, $\langle u'u'u' \rangle$, $\langle u'u'v' \rangle$, $\langle u'u'w' \rangle$, $\langle u'v'v' \rangle$, $\langle u'v'w' \rangle$, $\langle u'w'w' \rangle$, $\langle v'v'v' \rangle$, $\langle v'v'w' \rangle$, $\langle v'w'w' \rangle$, $\langle w'w'w' \rangle$, $\langle u'u'\theta' \rangle$, $\langle u'v'\theta' \rangle$, $\langle u'w'\theta' \rangle$, $\langle v'v'\theta' \rangle$, $\langle v'w'\theta' \rangle$, $\langle w'w'\theta' \rangle$, $\langle u'\theta'\theta' \rangle$, $\langle v'\theta'\theta' \rangle$, $\langle w'\theta'\theta' \rangle$, $\varepsilon_{12}, \varepsilon_{13}, \varepsilon_{23}, \Phi_{12}, \Phi_{13}, \Phi_{23}$, $\mathcal{D}_{p,12}, \mathcal{D}_{p,13}, \mathcal{D}_{p,23}, \mathcal{D}_{T,12}, \mathcal{D}_{T,13}, \mathcal{D}_{T,23}$, $\varepsilon_{1\theta}, \varepsilon_{2\theta}, \varepsilon_{3\theta}, \Phi_{1\theta}, \Phi_{2\theta}, \Phi_{3\theta}$, $\mathcal{D}_{p,1\theta}, \mathcal{D}_{p,2\theta}, \mathcal{D}_{p,3\theta}, \mathcal{D}_{v,1\theta}, \mathcal{D}_{v,2\theta}, \mathcal{D}_{v,3\theta}$, $\mathcal{D}_{T,1\theta}, \mathcal{D}_{T,2\theta}, \mathcal{D}_{T,3\theta}$

To train the classifier, we need to first create labeled data. We label the SST regime, i.e., $Y = 1$, if $u'_z/u'_h < 0.1$ (turbulence is highly anisotropic) and the WST regime, i.e., $Y = 0$, if $u'_z/u'_h > 0.3$ (turbulence is roughly isotropic) [10]. Here, $u'_z = \sqrt{\langle w'w' \rangle}$ is the vertical velocity fluctuation and $u'_h = \sqrt{\langle u'u' \rangle + \langle v'v' \rangle}$ is the horizontal velocity fluctuation at the wake center. Recall that the purpose of this work is to train a classifier and deploy the trained classifier in RANS of stratified wake flows without requiring the RANS practitioners to do any re-training. For this purpose, generalization is a question we must concern ourselves with. Corrupted training data, on the other hand, is not an issue since we deal with numerical simulations rather than experimental measurements.

Figure 6 shows u'_z/u'_h as a function of the non-dimensional time Nt. The statistics u'_z/u'_h is a decreasing function of time t. The lines do not collapse, but the transition from the WST regime to the IST regime and the transition from the IST regime to the SST regime are approximately at $Nt \approx 20$ and $Nt \approx 60$. The IST regime is not part of the classification task, considering the intended application of the classifier: if there were models that handle the WST and the far-wake SST regimes, the two models could be blended to handle the IST regime, and there would be no need to identify the IST regime.

Now, we train the classifier. Table 2 lists all the inputs. They include the first-order velocity and temperature, the second-order Reynolds stresses and heat fluxes, and the terms in the momentum and stress transport equations. There are 129 terms in total.

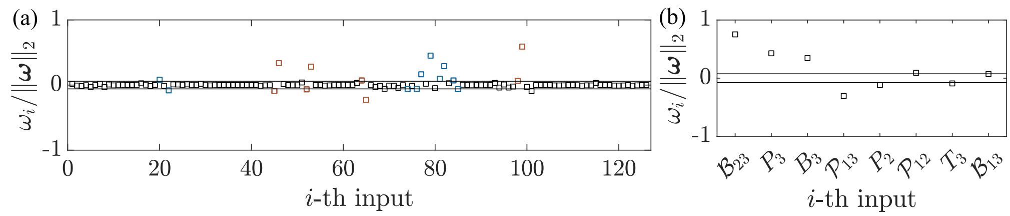


Fig. 8. (**a**) Input weights. The thin black lines are at $\pm 0.1 \max_i |\omega_i|/\|\boldsymbol{\omega}\|$. The inputs are color coded. The weights of the black ones are between $\pm 0.1 \max_i |\omega_i|/\|\boldsymbol{\omega}\|$. The weights of the orange and blue ones are either greater than $0.1 \max_i |\omega_i|/\|\boldsymbol{\omega}\|$ or smaller than $-0.1 \max_i |\omega_i|/\|\boldsymbol{\omega}\|$. The blue ones are modeled in an FRSM. The orange ones are computed in an FRSM. (**b**) Input weights. The 8 inputs are ranked from the most important to the least important.

Here, U, V, W are the mean velocity in the streamwise, transverse, and vertical directions; Θ, P are the mean temperature and the mean pressure; $'$ denotes temporal fluctuation, and $\langle\cdot\rangle$ denotes ensemble average; C_i, P_i, D_i, T_i, and B_i are the convective term, the pressure term, the viscous diffusion term, the turbulent diffusion term, and the buoyancy term in the mean momentum and temperature equations, and the subscript i is either 1, 2, 3 or θ, denoting the mean momentum equations in the three Cartesian directions and the temperature equation; $\mathcal{C}_{ij}$, $\mathcal{P}_{ij}$, Φ_{ij}, $D_{p,ij}$, $D_{\nu,ij}$, $D_{T,ij}$, ε_{ij}, and $\mathcal{B}_{ij}$ are the convection term, the production term, the pressure-redistribution term, the pressure diffusion term, the viscous diffusion term, the turbulent diffusion term, the dissipation term, and the buoyancy term in the momentum and thermal stress transport equations, and the subscripts i and j are either 1, 2, 3 or θ. Further details of the FRSM and the definition of the terms can be found in Ref. [44] and are not repeated here for brevity. In Table 2, a distinction is made between the computed (closed) terms and the modeled (unclosed) terms. This distinction is not critical in this study, since we will compute all the terms from DNS. However, when the classifier is deployed in FRSM, we prefer closed terms over unclosed terms as the latter incurs larger uncertainties in RANS calculations [45–48].

At a given downstream location, all the terms in Table 2 are functions of the transverse coordinate y and the vertical coordinate z (2D fields). Classification based on multiple 2D fields is possible [49,50], nonetheless we will boil down the 2D fields to scalars for computational efficiency in RANS. Specifically, for the terms that are always positive or the terms in the transport equations of $\langle u'u'\rangle$, $\langle v'v'\rangle$, and $\langle w'w'\rangle$, we take the maximum of their absolute values, whereas for the terms that are both positive and negative, e.g., $\langle u'v'\rangle$, we take the difference between the maximum and the minimum values. Here, we only consider values in the core of the wake, see Fig. 2c, 2f, and 2i. Non-trivial behaviors of a term outside the core of the wake are due to gravitational waves, which are not necessarily modeled (or computed) in a RANS. Last but not least, we follow Ref. [10] and normalize all terms locally (at a given streamwise location) with u'_h, L_{Hk}, and the background temperature gradient $\partial\Theta_0/\partial z$. We note that further feature scaling is not needed. Small terms are kept small. It is our intention to discard small terms, which suffer from a higher percentage of uncertainty in a RANS calculation and does not contribute significantly to the transport equations. Again, the cost function of logistic regression is convex, which guarantees a unique decision boundary.

Figure 7 shows the time history of the following input features: $\langle u'u'\rangle$, $\langle v'v'\rangle$, and $\langle w'w'\rangle$ in (a) and C_3, P_3, D_3, T_3, and B_3 in (b) as examples. From Fig. 7a, we see that while the velocity fluctuations in the two horizontal directions are comparable throughout, the vertical velocity fluctuation is suppressed in the late wake. From Fig. 7b, we see that the turbulent diffusion term and the pressure gradient term dominate in the early wake, whereas the pressure gradient term and the buoyancy term dominate in the late wake. These are consistent with the results in Fig. 3 and in the literature [10,33]. The question is whether our classifier is able to learn

Table 3
Classification results for R20F10. For a given column, the table tells how many WST (or SST) are classified by the classifier (ML in the table) as WST and SST. For a given row, the table tells how many of the predicted WST (or SST) actually belong to WST and SST.

ML \ DNS	WST	SST
WST	37	0
SST	0	14

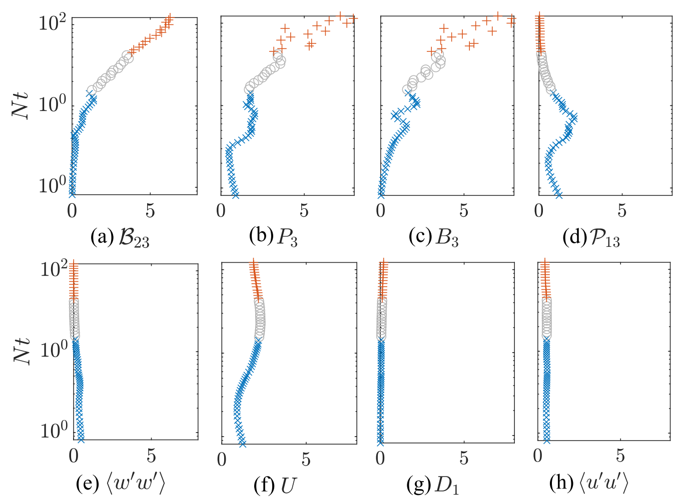


Fig. 9. Time histories of a few input features in R20F10 (**a**) $\mathcal{B}_{23}$, (**b**) P_3, (**c**) B_3, (**d**) $\mathcal{P}_{13}$, (**e**) $\langle w'w'\rangle$, (**f**) U, (**g**) D_1, (**h**) $\langle u'u'\rangle$. We use different symbols for different regimes, where the blue crosses indicate the WST regime, the grey circles indicate the IST regime and the orange pluses indicate the SST regime. Again, all inputs are locally normalized.

these physics from data and tell the WST regime apart from the SST regime or not.

The classifier is trained against the R20F10 data only. First, we pre-train the classifier using all inputs in Table 2. Figure 8a shows the weights of the 129 inputs. 18 inputs among the 129 inputs remain after applying weight thresholding. Among these 18 inputs, 8 are computed in an FRSM and 10 are modeled. The computed 8 terms are $\mathcal{B}_{23}, P_3, B_3, \mathcal{P}_{13}, P_2, \mathcal{P}_{12}, T_3, \mathcal{B}_{13}$, and they will be the inputs to our key classifier. Re-training the classifier using the 8 inputs only, we show the updated input weights in Fig. 8b. We observe the following. All but two (P_2 and $\mathcal{P}_{12}$) of the 8 inputs involve the vertical velocity component: P_3, T_3, and B_3 are the terms in the vertical momentum transport equation, and $\mathcal{P}_{13}$, $\mathcal{B}_{13}$, and $\mathcal{B}_{23}$ are the terms in the $\langle u'w'\rangle$ and $\langle v'w'\rangle$ transport equations. We know that the vertical velocity component behaves quite differently in the WST and the SST regimes. We see that the trained classifier is able to pick up that physics. Furthermore, the classifier is able to learn what terms are more critical and what terms are less critical: among the 8 inputs, B_3, P_3, $\mathcal{B}_{23}$, and $\mathcal{P}_{13}$ have comparably

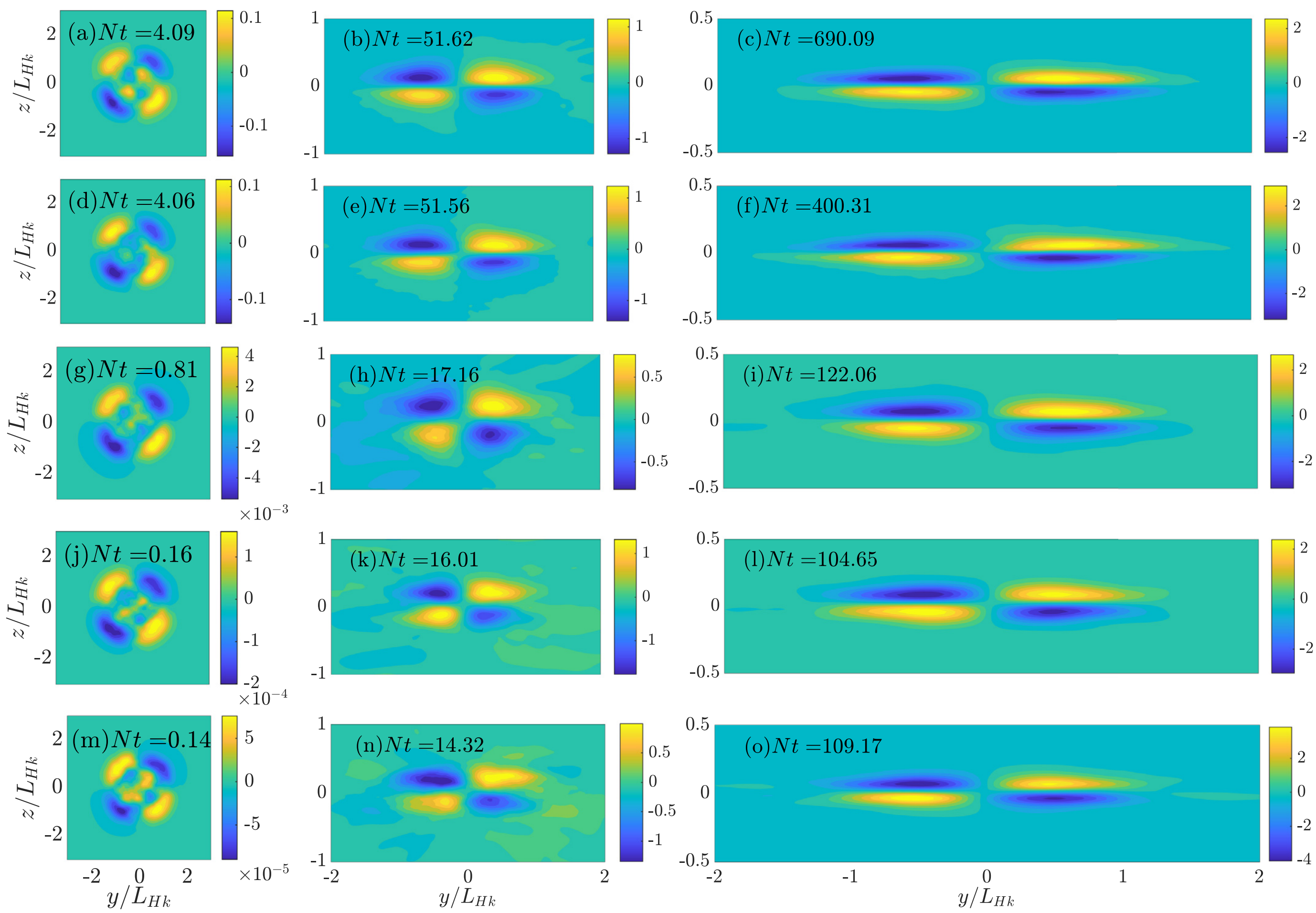


Fig. 10. Contours of the locally normalized buoyancy term $\mathcal{B}_{23}$ in all 5 cases. From top to bottom: R10F02, R20F02, R20F10, R20F50, and R50F50. The left column is at the beginning of the WST regime, the middle column is in the IST regime and $Nt \sim \mathcal{O}(10)$, and the right column is in the SST regime and $Nt \sim \mathcal{O}(100)$.

larger weights and are the more critical inputs to the classification task; P_2, T_3, $\mathcal{P}_{12}$, and $\mathcal{B}_{13}$ have comparably small weights and are the less critical inputs to the classification task. Besides, we notice that the u_h' and u_z' are filtered out. This is intriguing because the data is labeled per these two statistics. In the following, we attempt to explain this interesting observation from a flow physics perspective. Figure 9 shows the time history of B_3, P_3, $\mathcal{B}_{23}$, and $\mathcal{P}_{13}$, i.e., the 4 inputs that survive weight thresholding, and u_h', u_z', D_1 and U, which do not survive the weight thresholding. We see that the 4 terms that survive the weight thresholding are large whereas the 4 terms that do not survive weight thresholding are small: the maximum values of the locally normalized $\mathcal{B}_{23}$, P_3, and B_3 are about 8, but $\langle w'w'\rangle$, D_1, $\langle u'u'\rangle$ are nowhere greater than 0.5. This is advantageous because the larger terms are less susceptible to uncertainties in a RANS calculation. Moreover, the 4 terms that survive weight thresholding exhibit noticeably different behaviors in the WST and the SST regimes but the 4 terms that do not survive weight thresholding behave similarly in the WST and the SST regimes: $\mathcal{B}_{23}$, P_3, B_3 tend to take small values in the WST regime and large values in the SST regime, and $\mathcal{P}_{13}$ tends to take large values in the WST regime and small values in the SST regime, the (locally normalized) U takes small values in both the WST regime and the SST regime, and $\langle w'w'\rangle$, D_1, $\langle u'u'\rangle$ are small in both the WST and the SST regimes.

Before we proceed to testing the classifier, we show the classification results for the training data in Table 3. We see that the classifier is able to accurately classify the WST and the SST regimes for the training data.

Now, we apply the classifier to R10F02, R20F02, R20F50 and R50F50. From a flow physics perspective, the classifier is able to extrapolate only if the physics that governs the training data also governs the testing data. We test if it extrapolates so. Figure 10 shows the contour of $\mathcal{B}_{23}$, an input that is considered to be critical to the classification task, at 3 time instants in all 5 cases. We see that the term is small in the early wake and large in the late wake. Furthermore, we see that the term has similar behavior at the 5 flow conditions. Hence, we can conclude that the physics that governs the term $\mathcal{B}_{23}$ in case R20F10 also governs in other cases. Figure 11 shows the time histories of P_3, $\mathcal{P}_{13}$, $\mathcal{B}_{23}$, and B_3 in all 5 cases. We see that the terms behave similarly at all 5 flow conditions: $\mathcal{B}_{23}$, P_3, B_3 tend to take small values in the WST regime and large values in the SST regime, $\mathcal{P}_{13}$ tends to take large values in the WST regime and small values in the SST regime. Hence, the physics that governs R20F10 also governs the other cases.

Finally, the classification results are shown in Table 4. We see that the classifier generalizes very well and almost perfectly separates the WST regime and the SST regime.

Before we conclude, we train the classifier using all DNS data. Figure 12a shows the input weights. The results in Figs. 12a and 8b are similar: $\mathcal{B}_{23}, \mathcal{P}_{13}, P_3, \mathcal{P}_{12}$ are the more critical inputs, and B_3, P_2, T_3, $\mathcal{B}_{13}$ are comparably less critical. Figure 12b shows the classification results, and the classifier is able to perfectly separate the SST regime and the WST regime. In all, training using data from 1 case and from all 5 cases lead to similar classifiers.

To conclude, we conduct DNSs of temporally evolving wakes in stratified environments at Reynolds numbers from 10000 to 50000

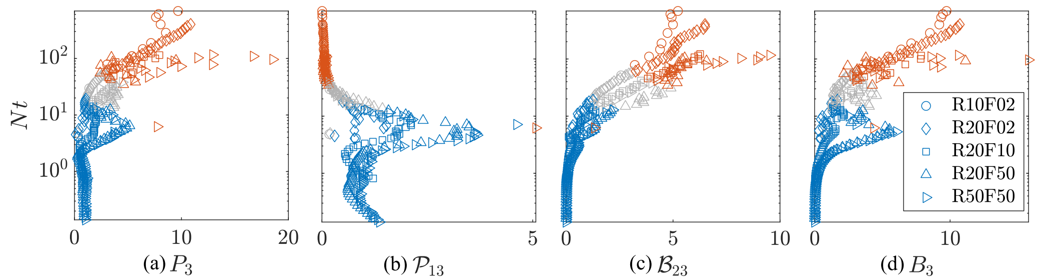
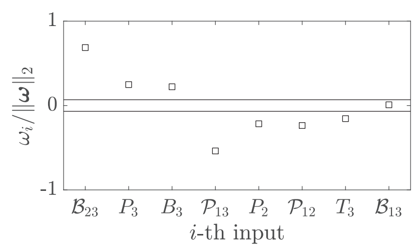


Fig. 11. The development of feature values in all the cases, including the training data set and the testing data sets. (**a**) P_3, (**b**) $\mathcal{P}_{13}$, (**c**) $\mathcal{B}_{23}$, (**d**) B_3. All the features are normalized using the local velocity/length scale. We use different colors for different regimes, where the blue symbols indicate the WST regime, the grey symbols indicate the IST regime and the orange symbols indicate the SST regime.

Table 4
Testing results. The meaning of the table headers is the same as Table 3.

R10F02			**R20F02**		
ML \ DNS	WST	SST	ML \ DNS	WST	SST
WST	3	0	WST	16	0
SST	0	9	SST	0	24
R20F50			**R50F50**		
ML \ DNS	WST	SST	ML \ DNS	WST	SST
WST	55	0	WST	54	1
SST	0	15	SST	0	14

(a)



(b)

ML \ DNS	WST	SST
WST	166	0
SST	0	76

Fig. 12. (a) Input weights for the classifier that is trained against all DNS data. (b) Classification results.

and Froude numbers from 2 to 50. Rather than computing one realization, 80 to 100 realizations are computed to get converged statistics from the ensemble average. The DNS data are subsequently labeled for training and testing purposes.

The intended application of the classifier is full Reynolds stress model (FRSM), and we limit ourselves to information that is available in an FRSM. A linear logistic regression classifier with weight thresholding is proposed. Weight thresholding helps to remove features that are less critical to the classification task and to identify the critical physics. Here, the production and the buoyancy terms in the transport equation of $\langle u'w' \rangle$ and the production and the buoyancy term in the transport equation of $\langle W \rangle$ are found to be critical to the classification of the wake into the weakly stratified turbulence and the strongly stratified turbulence regimes.

The classifier is trained against the flow at only one flow condition, i.e., $Re = 20000$, $Fr = 10$, and applied to other flow conditions with Reynolds numbers and Froude numbers both lower and higher than the training condition. The classifier is found to generalize very well. Two conclusions can be drawn from this result. First, the physics that governs the flow in the weakly stratified turbulence and the strongly stratified turbulence regimes are universal (with respect to the Reynolds number and the Froude number). Second, the classifier is able to identify and "learn" about the physics.

The intended application of this classifier is FRSM, more generally, RANS modeling. A RANS model is evaluated at every grid point at every iteration. It is undesirable to have a model that is costly to evaluate. This makes logistic regression a better choice than more advanced tools—if it does generalize, and we show that the classifier generalizes. That said, the discussion in this paper is limited to the classification of the wake flow regimes. Although the classifier generalizes well, we are yet to test it in the context of FRSM.

Declaration of Competing Interest

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

Acknowledgment

The authors would like to thank Jiaqi Li and Naman Jain for their assistance in the DNS simulations. This work is supported by ONR, contract N000142012315.

Appendix

In the classifier, the cross-entropy loss function L_{CE} in Eq. (2) has a regularization term. The coefficient C_0 is determined through a cross validation process. The data from R20F10 is used as the training data and the data from other cases is used for cross validation. The linear logistic classifier with weight thresholding is trained for multiple C_0 candidates, from 10^{-4} to 1. The results in Fig. A1 show that $C_0 = 0.01$ gives the smallest L_{CE}. Hence, we have $C_0 = 0.01$.

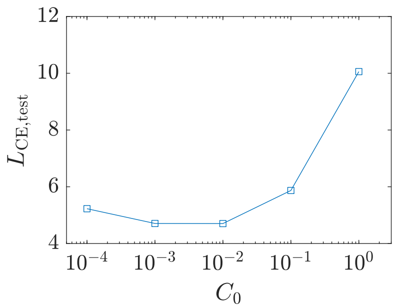


Fig. A1. The cross validation of the coefficient C_0.

References

[1] J.-T. Lin, Y.-H. Pao, Wakes in stratified fluids, Ann. Rev. Fluid Mech. 11 (1979) 317–338.
[2] L. Mahrt, Stratified atmospheric boundary layers, Boundary-Layer Meteorol. 90 (1999) 375–396.
[3] E.J. Hopfinger, Turbulence in stratified fluids: a review, Journal of Geophysical Research: Oceans 92 (1987) 5287–5303.
[4] J.J. Riley, M.-P. Lelong, Fluid motions in the presence of strong stable stratification, Ann. Rev. Fluid Mech. 32 (2000) 613–657.
[5] X.I.A. Yang, S. Pirozzoli, M. Abkar, Scaling of velocity fluctuations in statistically unstable boundary-layer flows, J. Fluid Mech. 886 (2020) A3.
[6] X.I.A. Yang, P.E.S. Chen, R. Hu, et al., Logarithmic-linear law of the streamwise velocity variance in stably stratified boundary layers, Boundary-Layer Meteorol. 183 (2022) 199–213.
[7] E. Lindborg, The energy cascade in a strongly stratified fluid, J. Fluid Mech. 550 (2006) 207–242.
[8] G. Brethouwer, P. Billant, E. Lindborg, et al., Scaling analysis and simulation of strongly stratified turbulent flows, J. Fluid Mech. 585 (2007) 343–368.
[9] S.M. de Bruyn Kops, J.J. Riley, The effects of stable stratification on the decay of initially isotropic homogeneous turbulence, J. Fluid Mech. 860 (2019) 787–821.
[10] K. Chongsiripinyo, S. Sarkar, Decay of turbulent wakes behind a disk in homogeneous and stratified fluids, J. Fluid Mech. 885 (2020) A31.
[11] G.R. Spedding, The evolution of initially turbulent bluff-body wakes at high internal froude number, J. Fluid Mech. 337 (1997) 283–301.
[12] X.I.A. Yang, K.P. Griffin, Grid-point and time-step requirements for direct numerical simulation and large-eddy simulation, Phys. Fluids 33 (2021) 015108.
[13] X.I.A. Yang, J. Hong, M. Lee, et al., Grid resolution requirement for resolving rare and high intensity wall-shear stress events in direct numerical simulations, Phys. Rev. Fluids 6 (2021) 054603.
[14] J. Li, X.I.A. Yang, R.F. Kunz, Grid-point and time-step requirements for large-eddy simulation and Reynolds-averaged Navier-Stokes of stratified wakes, Phys. Fluids 34 (2022) 115125.
[15] X.L.D. Huang, N. Jain, M. Abkar, et al., Determining a priori a RANS model's applicable range via global epistemic uncertainty quantification, Computers & Fluids 230 (2021) 105113.
[16] K. Duraisamy, G. Iaccarino, H. Xiao, Turbulence modeling in the age of data, Ann. Rev. Fluid Mech. 51 (2019) 357–377.
[17] S.L. Brunton, B.R. Noack, P. Koumoutsakos, Machine learning for fluid mechanics, Ann. Rev. Fluid Mech. 52 (2020) 477–508.
[18] T.-R. Xiang, X.I.A. Yang, Y.-P. Shi, Neuroevolution-enabled adaptation of the Jacobi method for Poisson's equation with density discontinuities, Theor. Appl. Mech. Lett. 11 (2021) 100252.
[19] X.L.D. Huang, X.I.A. Yang, R.F. Kunz, Wall-modeled large-eddy simulations of spanwise rotating turbulent channels-comparing a physics-based approach and a data-based approach, Phys. Fluids 31 (2019) 125105.
[20] X.L.D. Huang, X.I.A. Yang, A Bayesian approach to the mean flow in a channel with small but arbitrarily directional system rotation, Phys. Fluids 33 (2021) 015103.
[21] C.-Y. Ohh, G.R. Spedding, Wake identification of stratified flows using dynamic mode decomposition, Phys. Rev. Fluids 7 (2022) 024801.
[22] V.K. Chinta, C.-Y. Ohh, G. Spedding, et al., Regime identification for stratified wakes from limited measurements: a library-based sparse regression formulation, Phys. Rev. Fluids 7 (2022) 033803.
[23] M.M. Gibson, B.E. Launder, Ground effects on pressure fluctuations in the atmospheric boundary layer, J. Fluid Mech. 86 (1978) 491–511.
[24] J.P. Panda, H.V. Warrior, S. Maity, et al., An improved model including length scale anisotropy for the pressure strain correlation of turbulence, J. Fluid Eng. 139 (2017).
[25] P. Moin, K. Mahesh, Direct numerical simulation: a tool in turbulence research, Ann. Rev. Fluid Mech. 30 (1998) 539–578.
[26] H.H.A. Xu, S.J. Altland, X.I.A. Yang, et al., Flow over closely packed cubical roughness, J. Fluid Mech. 920 (2021) A37.
[27] H.H.A. Xu, S. Lynch, X.I.A. Yang, Direct numerical simulation of slot film cooling downstream of misaligned plates, Flow 2 (2022) E7.
[28] W. Zhang, X. Zhu, X.I.A. Yang, et al., Evidence for Raupach et al.'s mixing-layer analogy in deep homogeneous urban-canopy flows, J. Fluid Mech. 944 (2022) A46.
[29] X.I.A. Yang, M. Ge, Revisiting Raupach's flow-sheltering paradigm, Boundary Layer Meteorol 179 (2021) 313–323.
[30] Y. Bin, X.I.A. Yang, Y. Yang, et al., Evolution of two counter-rotating vortices in a stratified turbulent environment, J. Fluid Mech. 951 (2022) A47, doi:10.1017/jfm.2022.905.
[31] K.A. Brucker, S. Sarkar, A comparative study of self-propelled and towed wakes in a stratified fluid, J. Fluid Mech. 652 (2010) 373–404.
[32] D.G. Dommermuth, J.W. Rottman, G.E. Innis, et al., Numerical simulation of the wake of a towed sphere in a weakly stratified fluid, J. Fluid Mech. 473 (2002) 83–101.
[33] P.J. Diamessis, G.R. Spedding, J.A. Domaradzki, Similarity scaling and vorticity structure in high-Reynolds-number stably stratified turbulent wakes, J. Fluid Mech. 671 (2011) 52–95.
[34] M.J. Gourlay, S.C. Arendt, D.C. Fritts, et al., Numerical modeling of initially turbulent wakes with net momentum, Phys. Rev. Fluids 13 (2001) 3783–3802.
[35] M.B. de Stadler, S. Sarkar, K.A. Brucker, Effect of the Prandtl number on a stratified turbulent wake, Phys. Fluids 22 (2010) 095102.
[36] J.A. Redford, T.S. Lund, G.N. Coleman, A numerical study of a weakly stratified turbulent wake, J. Fluid Mech. 776 (2015) 568–609.
[37] Y. Yang, E.P. Van Der Poel, R. Ostilla-Mónico, et al., Salinity transfer in bounded double diffusive convection, J. Fluid Mech. 768 (2015) 476–491.
[38] Y. Yang, W. Chen, R. Verzicco, et al., Multiple states and transport properties of double-diffusive convection turbulence, Proceedings of the National Academy of Sciences 117 (2020) 14676–14681.
[39] J. Li, Y. Yang, Thermohaline interleaving induced by horizontal temperature and salinity gradients from above, J. Fluid Mech. 927 (2021) A12.
[40] S. Nidhan, O.T. Schmidt, S. Sarkar, Analysis of coherence in turbulent stratified wakes using spectral proper orthogonal decomposition, J. Fluid Mech. 934 (2022) A12.
[41] J.L. Ortiz-Tarin, S. Nidhan, S. Sarkar, High-Reynolds-number wake of a slender body, J. Fluid Mech. 918 (2021) A30.
[42] D.G. Kleinbaum, K. Dietz, M. Gail, et al., Logistic regression, Springer, 2002.
[43] M.P. LaValley, Logistic regression, Circulation 117 (2008) 2395–2399.
[44] N. Jain, H.T. Pham, X. Huang, et al., Second moment closure modeling and direct numerical simulation of stratified shear layers, J Fluids Eng 144 (2022) 041102.
[45] A. Eidi, R. Ghiassi, X.I.A. Yang, et al., Model-form uncertainty quantification in RANS simulations of wakes and power losses in wind farms, Renew Energy 179 (2021) 2212–2223.
[46] M. Emory, J. Larsson, G. Iaccarino, Modeling of structural uncertainties in Reynolds-averaged Navier-Stokes closures, Phys. Fluids 25 (2013) 110822.
[47] S.D. Hornshøj-Møller, P.D. Nielsen, P. Forooghi, et al., Quantifying structural uncertainties in Reynolds-averaged Navier–Stokes simulations of wind turbine wakes, Renew Energy 164 (2021) 1550–1558.
[48] H. Xiao, P. Cinnella, Quantification of model uncertainty in RANS simulations: a review, Prog. Aerosp. Sci. 108 (2019) 1–31.
[49] Y. Liu, X. Chen, H. Peng, et al., Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207.
[50] Y. Sun, L. Zhu, G. Wang, et al., Multi-input convolutional neural network for flower grading, Journal of Electrical and Computer Engineering 2017 (2017).