Proceedings of Machine Learning Research vol 245:1–10, 2024 2024 International Conference on Machine Learning and Intelligent Computing

Research on Random Forest Regression Algorithms for Predicting Reynolds Stress Anisotropy in Separation Flow around Near-Wall Cylinder

Kewei Deng 970981704@QQ.COM and **Jianhong Sun*** JHSUN@NUAA.EDU.CN
College of Aerospace Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

Standard RANS model is widely used in turbulence modelling and numerical simulation, due to its linear assumption of eddy viscosity, it is not suitable to accurately predict the Reynolds stress, especially in separated flows with anisotropy of Reynolds stress. In this study, a machine learning model is proposed by applying the random forest regression algorithm to learn the deviation between eddy-viscosity and high-fidelity models for separation flows around near-wall cylinder. The output features which represent magnitude, shape and directions of Reynolds stress are extracted by decomposing Reynolds stress tensor, while 8 types of input features are extracted from raw local flow data sets to represent the main physical characteristics of the flow field. Both input and output features satisfy Galilean invariance, which contributes to improving prediction accuracy and generalization performance of the random forest regression model. The results show that the random forest regression model has a great potential to effectively predict the Reynolds stress anisotropy distribution of different Reynolds numbers.

Keywords: eddy-viscosity model; Reynolds stress anisotropy; random forest (RF) regression; Reynolds-Averaged Navier-Stokes (RANS) model.

1. Introduction

Turbulence refers to the phenomenon in fluid motion where the distribution of flow state in time and space is extremely chaotic, complex, irregular, and filled with unpredictable pulsating states (Jovanović, 2021). In turbulent motion, it is difficult to predict the direction and speed of fluid particles with a strong sense of randomness. Due to the unpredictability and irregularity of turbulence, systematic research on turbulence has always been a huge challenge.

With the rapid development of computer science and parallel computing technology, turbulence models such as direct numerical simulation (DNS) and large eddy simulation (LES) are proposed, which require intensive computations and have been widely applied (XU, 2009). However, DNS and LES methods require highly refined grid model, resulting in long computation time and low efficiency (Singh and Duraisamy, 2016). Therefore, the Reynolds-averaged Navier-Stokes (RANS) method is still very important in engineering applications. Reynolds stress closure of RANS model is usually based on the Boussinesq linear hypothesis, therefore, when it comes to complex flow fields generated by large curvature, separated flows, and shock wave formation, RANS simulation results are often inaccurate or even contra-dictory to actual situations.

DNS and LES methods can provide high fidelity and credibility in numerical simulation, but using both methods come with massive amounts of data, bringing difficulties to subsequent data processing. To handle these large-scale data sets, statistical analysis, data mining, and machine learning have become the main tools. In recent years, researchers have increasingly tended to find

the relationships among variables and describe flow field characteristics by mining the intrinsic connections of the data (Brunton et al., 2020). Combining machine learning with turbulence models and their numerical simulations can improve simulation accuracy, accelerate computation speed, excavate flow field characteristics, and optimize turbulence control. More and more researchers have also begun to explore the use of machine learning methods to study turbulent processes Vinuesa, making it significant to research on applying machine learning to turbulence models and their numerical simulations.

Machine learning has been used to identify and model the differences in Reynolds stress tensor between RANS models and high-fidelity simulations. Researchers have used support vector machines, decision trees, and random forest methods to classify and predict high-uncertainty areas in Reynolds stress tensor. Tang et al. (2023) proposed a robust data-driven Reynolds-averaged turbulence model with uncertainty quantification and non-linear correction with the Bayesian deep neural network. The results show the proposed model has a good generalization performance when simulating turbulent flows with large scale separation. There is also a growing trend to take account of necessary physical mechanisms and laws Ling et al. (2016), and apply them into existing turbulence models, improving model prediction accuracy and stability.

2. Overview of the Research Problem

The problem researched in this paper is to improve the calculation accuracy of the RANS turbulence model based on eddy viscosity assumption by using machine learning algorithms to learn from high-precision flow field data. It is known that models using eddy viscosity assumption often perform poorly in flows which are characterized by Reynolds stress anisotropy. Thus, it is expected to correct the Reynolds stress to compensate for its shortcomings. Machine learning algorithms are used as the research method for this problem, in order to reduce the dimensionality of the data and extract the most relevant features, and to construct a function relationship between the input and output features, the flow field data need to be pre-processed by feature extraction and regularization, meanwhile, machine learning models are considered to be embedded certain physical experience to improve their accuracy and interpretability, such as physical symmetries and invariance. In this paper, the design of input and output features for the model is presented based on Galilean invariance. As the random forest algorithm is suitable for high-dimensional feature space and has good robustness to non-important features and outlier data, we use random forest algorithm to construct the machine learning model in this paper. The flow studied in this paper is chosen as the flow around near-wall cylinder, which is characterized by separation and Reynolds stress anisotropy.

3. Construction of RF Regression Model

3.1. Output Features

The output features should be able to characterize the Reynolds stress state. The Reynolds stress is a second-order symmetric tensor and is not suitable as a direct output target. Therefore, it needs to be decomposed, and as a second-order symmetric tensor, it has six degrees of freedom, and only six target features need to be determined. Here we decompose the Reynolds stress and consider Galilean invariance to use matrix knowledge to select output targets based on properties such as invariants of vectors or tensors.

Reynolds stress tensor is a second-order symmetric tensor, it can be normalized to obtain the Reynolds stress anisotropy tensor b_{ij}:

$$b_{ij} = \frac{\tau_{ij}}{2k} - \frac{1}{3}\delta_{ij}. \tag{1}$$

where τ_{ij} is Reynolds stress tensor, k is turbulent kinetic energy, and δ_{ij} is Kronecker delta.

It can be seen that b_{ij} is also a symmetric tensor, and all its elements are real numbers. For a real symmetric matrix, it can be diagonalized. Therefore, dropping the subscript, Reynolds stress tensor can be written as

$$\tau = 2k(\frac{1}{3}I + V\Lambda V^T). \tag{2}$$

where I is unit matrix, V is orthonormal matrix composed of eigenvectors of b, and Λ is diagonal matrix composed of eige-nvalues of b (the eigenvalues are sorted in decreasing order along diagonal).

In order to provide a more convenient and intuitive repre-sentation of turbulent states, the Centroid Triangle (Tang et al., 2023) is used in this paper to analyze turbulence states. To calculate the coo-rdinates (η, ξ) in Centroid Triangle, assume three eigenvalues of b are λ_1, λ_2 and λ_3 ($\lambda_1 > \lambda_2 > \lambda_3$), introduce C_1, C_2 and C_3 as

$$C_1 = \lambda_1 - \lambda_2. \tag{3}$$

$$C_2 = 2(\lambda_2 - \lambda_3). \tag{4}$$

$$C_3 = 3\lambda_3 + 1. \tag{5}$$

Let coordinates of three vertices in the Centroid Triangle be (η_{1c}, ξ_{1c}), (η_{2c}, ξ_{2c}), (η_{3c}, ξ_{3c}) which representing three limiti-ng states. To plot any turbulence states corresponding to its C_1, C_2 and C_3 in Centroid Triangle, take a convex combination of the three limiting states:

$$\eta = C_1\eta_{1c} + C_2\eta_{2c} + C_3\eta_{3c}. \tag{6}$$

$$\xi = C_1\xi_{1c} + C_2\xi_{2c} + C_3\xi_{3c}. \tag{7}$$

Fig.1 shows an example of Reynolds stress anisotropy dist-ributions in Centroid Triangle.

Next, we analyze the unitary orthogonal matrix V which consists of three unit-length eigenvectors of b. Referring to the rotation of a rigid body, the matrix V can be obtained by sequ-entially rotating the unit matrix around coordinate axes: (1) rotating around z-axis by an angle of φ_1; (2) rotating around x-axis by an angle of φ_2; (3) rotating around z-axis by an angle of φ_3.

Finally, we obtained six output variables to describe Rey-nolds stress tensor which have certain physical interpretation, while satisfying Galilean invariance. They are magnitude (k), shape (η, ξ), and direction (φ_1, φ_2, φ_3). In this paper, we use de-viations of these variables between low-fidelity flow and targ-et flow to be the output features of RF model.

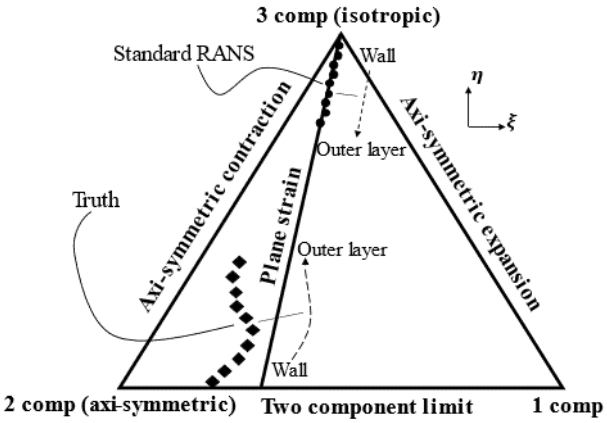

Figure 1: Centroid Triangle.

3.2. Input Features

The input features should sufficiently represent the flow field which needs to be corrected, without any contribution from the high-fidelity flow field dataset. Therefore, the input features can only be extracted from uncorrected flow. In addition, to improve the generalization performance of machine learning model, the input features should satisfy Galilean invariance and be dimensionless.

Therefore, to construct the input features, the following raw local variables are considered: the mean pressure P and its gradient vector, the mean velocity U and its gradient tensor, the turbulence kinetic energy k and its gradient vector, the turbulent dissipation rate ε, the eddy viscosity v_t, the molecular viscosity v, and the distance d to the nearest wall. These variables can be obtained from the RANS solver.

According to a series of turbulent features proposed by Ling et al. (2016), the following features are selected as input features for machine learning model, as shown in Table 1:

Table 1: Input Features

Feature q_β	Description	Original Formulation $\hat{q}_\beta$	Normalization Factor q^*_β
q_1	Q criterion	$\frac{1}{2}\left(\|\Omega\|^2 - \|S\|^2\right)$	$\| S\|^2$
q_2	Turbulence intensity	k	$\frac{1}{2}U_iU_i$
q_3	Turbulence Reynolds number	$\min\left(\frac{\sqrt{k}d}{50v},\ 2\right)$	
q_4	Pressure gradient along streamline	$U_k\frac{\partial P}{\partial x_k}$	$\sqrt{\frac{\partial P}{\partial x_j}\frac{\partial P}{\partial x_j}U_iU_i}$
q_5	Ratio of turbulent time scale to mean strain time scale	$\frac{k}{\varepsilon}$	$\frac{1}{\|S\|}$
q_6	Viscosity coefficient	v_t	$100v$
q_7	Frobenius norm of Reynold stress tensor	$\left\|\overline{u_i'u_j'}\right\|$	k
q_8	Non-orthogonality between velocity and its gradients S. Banerjee and Zenger (2007)	$\left\|U_iU_j\frac{\partial U_i}{\partial x_j}\right\|$	$\sqrt{U_lU_iU_i\frac{\partial U_i}{\partial x_j}U_k\frac{\partial U_k}{\partial x_j}}$

In Table 1, $\|.\|$ is its Frobenius norm, Ω is rotation rate tensor, and S is strain rate tensor. The normalization of the original features is performed according to the normalization factors given, using the method proposed by Ling et al. (2016) (except for q_3), which is as follows:

$$q_\beta = \frac{\hat{q}_\beta}{|\hat{q}_\beta| + \left|q_\beta^*\right|}, \beta = 1, 2, 4, 5, \cdots, 8. \tag{8}$$

The normalization method can ensure most of the inputs lie in the range of [-1,1], and make all features have similar meas-urement scales. For feature q_3, normalization is not required because it is already dimensionless and its values range from 0 to 2.

3.3. Hyperparameter Optimization

In machine learning (ML), hyperparameter optimization or tuning is the problem of choosing a set of optimal hyperpara-meters for a learning algorithm. Cross-validation is often used to estimate generalization performance, and therefore choose the set of values for hyperparameters that maximize it.

Common types of cross-validation methods are k-fold cross-validation (such as ten-fold cross-validation) and leave-one-out cross-validation. K-fold cross-validation divides the training set into K equal parts, sequentially using each part as a validation set and the remaining K-1 parts as a training set. This process is repeated by K times, and the results are avera-ged or weighted to evaluate the model parameters. The leave-one-out method takes one sample from the training set as the validation set in turn and trains with the remaining N-1 sam-ples (assuming that there are N samples in total).

However, for hyperparameter optimization of random for-est models, the aforementioned cross-validation methods are not used. Instead, the out-of-bag estimate method is used bec-ause bootstrap sampling used in random forest models prod-uces out-of-bag samples. Each decision tree has corresponding out-of-bag samples that can be used to test the error and perfo-rmance of each decision tree, so it is unnecessary to perform cross-validation or use an independent test set to obtain an un-biased estimate of the error.

In this section, the mean squared error (MSE) is used to evaluate the impact of model parameters on model performa-nce, and four important hyperparameters of the random forest model are optimized: the number of trees (n_trees), the maxi-mum number of splits per tree ($\max$ _splits), the minimum number of samples per leaf ($\min$ _sample_leaf), and the maxim-um number of input features per node split ($\max$ _features).

Experiments using the method of controlling variables are conducted. For each variable, set different values of it while keeping other variables constant to research its effects on the regression model. When controlling for other variables, use the default values: n_trees = 100 for the number of trees, $\min$ _sample_leaf = 5 for the minimum number of samples per leaf, and $\max$ _features = 3 for the maximum number of input features per node split. And change the parameter values of studied variable, calculate the corresponding MSE values, and record the elapsed time. Fig.2 shows the variations of MSE and elapsed time with each hyperparameter.

As shown in Fig.2, considering the influence of various hyperparameters on the model error, and in order to ensure that the model has good generalization performance while avoiding the machine learning model being too complex and large, and avoiding the training time being too long, the opti-mal hyperparameters are set as follows:

n_trees = 30; $\min$ _sample_leaves = 5; $\max$ _splits = 100; $\max$ _features = 4.

3.4. Case set-up

In this paper, the flow around a near-wall cylinder is cho-sen as the research object. To improve the accuracy of eddy-viscosity model, a random forest regression model is constru-cted from the train-ing flow (unfixed flow, calculated by $k-\varepsilon$ model) to the target flow (high fidelity flow, calculated by RSM model). If the predicted results of ML (Machine Learn-ing) model are consistent with the target flow, then the model can be applied to other flow case to correct the results of eddy-viscosity model. The following Table II provides the informat-ion about training flow and target flow:

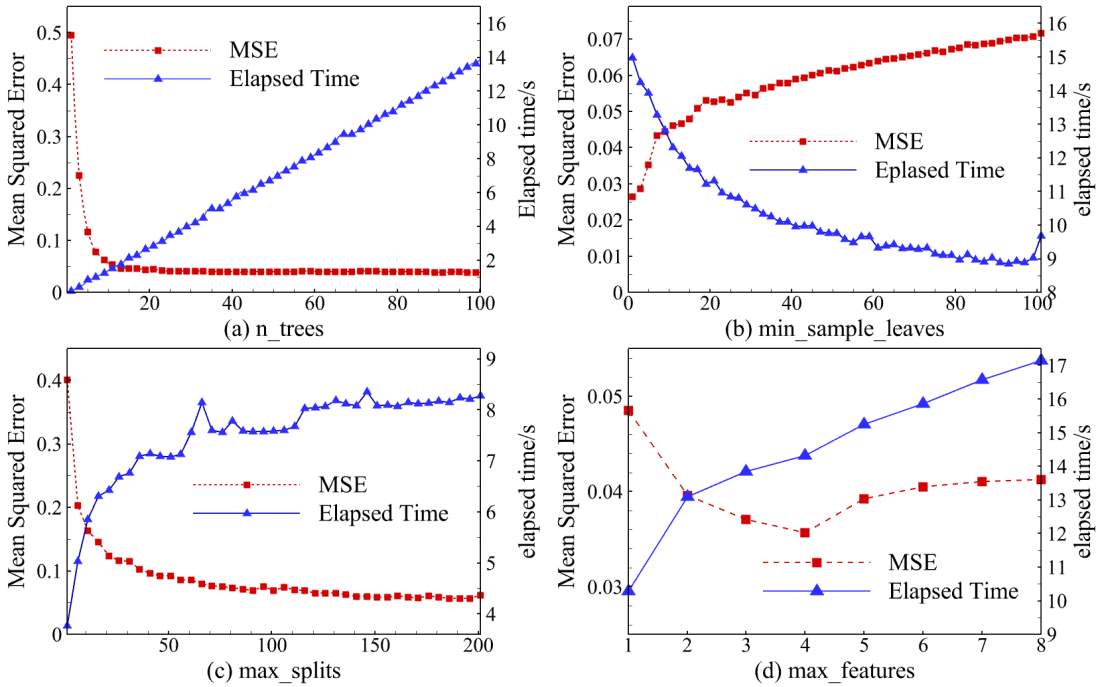

Figure 2: The impacts of hyperparameters on the regression model's out-of-bag error (MSE) and elapsed time.

The Reynolds numbers are calculated based on the charac-teristic length of the corresponding model (the diameter D of cylinder) and the incoming main flow velocity. A structured mesh is used, and to ensure computational accuracy, the first grid layer near the wall has a $y+$ (non-dimensional wall dista-nce) value less than 1, and wall function is not utilized. The geometry and boundary conditions are shown in Fig.3. The length L_x of the computational domain is 12D, and the height L_y of the computational domain is 5.333D. To simulate the int-eraction between the cylinder and the boundary layer of plane, a Blasius velocity profile is imposed at the inlet with a boun-dary layer thickness δ of 1.022D. The upper boundary of com-putational domain is set as a moving boundary, while the bott-om of computational domain is set as a no-slip wall boundary. Here, D represents the diameter of the cylinder, and G repre-sents the distance between the cylinder and the wall. Here gap ratio G/D=1. The changes of Reynolds number are achieved by modifying the fluid density or kinematic viscosity while keeping the incoming velocity and characteristic length cons-tant. The calculations are performed for 15 vortex shedding periods. A total of 33,634 sample points is taken.

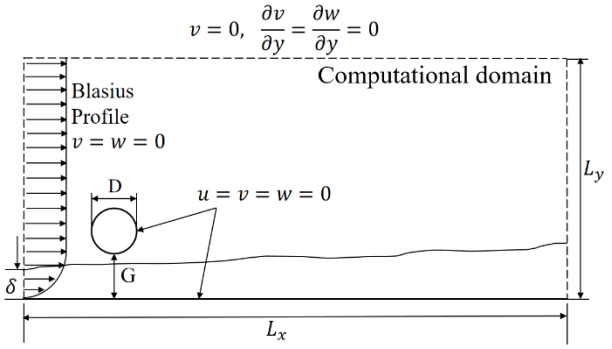

Figure 3: Geometry and boundary conditions.

For the supervised learning regression problem, the input features of the training set are taken from the $k-\varepsilon$ calculation of the flow field (low-fidelity flow), while the output targets (la-bels) are obtained from the RSM calculation of the flow field (high-fidelity flow). The deviations of output features Δk, $\Delta\eta$, $\Delta\xi$, $\Delta\varphi_1$, $\Delta\varphi_2$ and $\Delta\varphi_3$, corresponding to the magnitude (k), sha-pe (η, ξ), and direction (φ_1, φ_2, φ_3) of the Reynolds stress tens-or, are taken as the responses of the random forest regression model.

Let $\Delta\tau_\alpha$ $(\alpha = 1, 2, 3, \cdots, 6)$ represent the six deviations Δk, $\Delta\eta$, $\Delta\xi$, $\Delta\varphi_1$, $\Delta\varphi_2$ and $\Delta\varphi_3$ in sequence, let q denote the input featu-res of training flow (low-fidelity flow). Then the goal is to co-nstruct regression function f_α that maps input to output:

$$f_\alpha \colon q \rightarrow \Delta\tau_\alpha(\alpha = 1, 2, \cdots, 6). \tag{9}$$

According to the hypothesis proposed by Tracey and Dura-isamy et al. Ling and Templeton (2015), the differences among the six quantities of $\Delta\tau_\alpha$ are independent of each other. Therefore, a separate regression function is constructed for each quantity.

4. Results and Analysis

4.1. Verification results of ML model

It is necessary to verify the predictive performance of the random forest regression model on the training set to avoid prediction distortion. The deviations of Reynolds stress aniso-tropy, turbulence kinetic energy, and components of Reynolds stress are predicted, and the corrected results are compared with the target values on the training set to test whether the regression model can learn the deviations between uncorrected flow ($k-\varepsilon$) and target flow (RSM). Since $k-\varepsilon$ model uses Boussinesq hypothesis to deal with Reynolds stress term, it cannot fully reflect Reynolds stress anisotropy. In contrast, RSM model builds transport equations for components of Reynolds stress separately, which could effectively reflect Reynolds stress anisotropy. Therefore, accurately predicting Reynolds stress anisotropy is essential for correcting turbulent states.

Fig.4 shows the predicted results of Reynolds stress anisotropy for the training flow by random forest regression model under different locations in the wake region of the cylinder, where x/D equals 1, 2, 3, and 4, respectively. The x-coordinate of the cylinder center is 0. It can be seen that the predicted Reynolds stress distribution of $k-\varepsilon$ flow is a straight line starting from the vertex of

the triangle, showing complete isotropy. In contrast, the Reynolds stress calculated by RSM model shows a certain extent of anisotropy. The predicted results of the regression model follow the same distribution pattern as RSM model, indicating that the random forest regression model can learn the deviations of the crucial regions between uncorrected flow and target flow, and effectively predict Reynolds stress anisotropy without overfitting or underfitting.

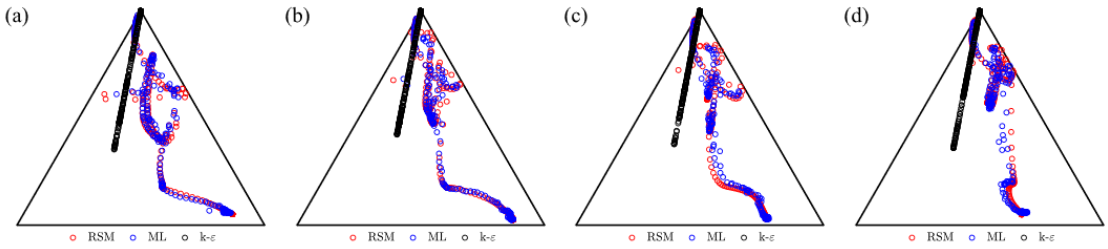

Figure 4: Verification results of Reynolds anisotropy distribution (Re=4000); (a)x/D=1; (b)x/D=2; (c)x/D=3; (d)x/D=4.

4.2. Prediction results at different Reynolds number

To verify the generalization performance of random forest regression model, the prediction is carried out considering increased Reynolds number and decreased Reynolds number.

Fig.5 and Fig.6 shows the prediction results at Re=2000 and Re=6000, respectively. The inputs are input features of original flow field ($k-\varepsilon$) corresponding to its Reynolds number, the outputs are obtained by adding the predicted deviations to original flow field.

As shown in Fig. 5 and Fig. 6, the Reynolds stress distri-bution obtained by the prediction model is in good agreement with the distribution of RSM, that is, the distance and dispersion of the point distribution near the boundary line of the free shear layer and the logarithmic layer, as well as the space region of outer layer are basically consistent with target flow. However, for larger Reynolds number, the prediction results of the wake region close to the fully developed region have a certain degree of deviation from the target flow, as shown in Fig.6(c) and Fig.6(d).

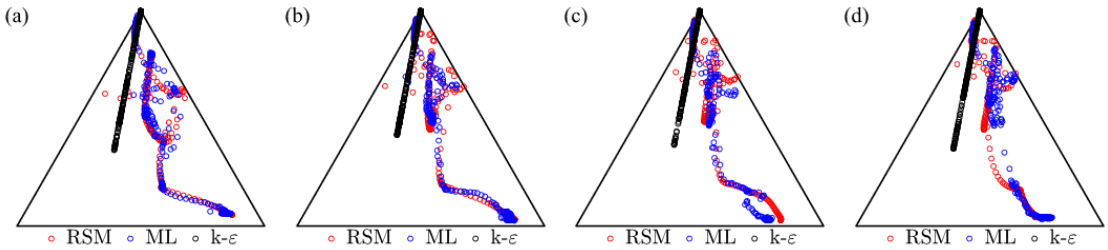

Figure 5: Prediction results of Reynolds anisotropy distribution (Re=2000); (a)x/D=1; (b)x/D=2; (c)x/D=3; (d)x/D=4.

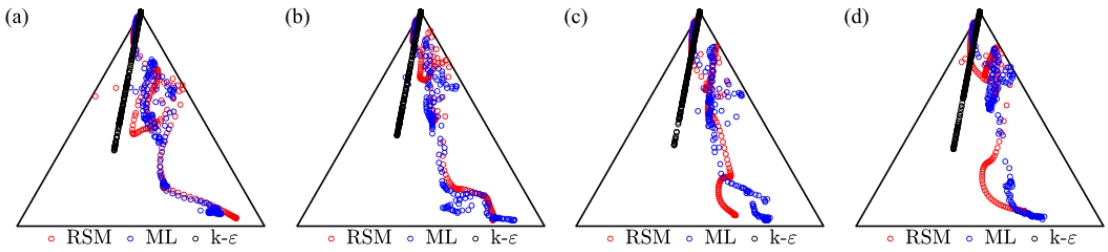

Figure 6: Prediction results of Reynolds anisotropy distribution (Re=6000); (a)x/D=1; (b)x/D=2; (c)x/D=3; (d)x/D=4.

5. Summary and Conclusion

In this paper, a prediction model is proposed by applying the RF algorithm to learn the deviations between $k-\varepsilon$ and RSM models for the flow around a near-wall cylinder. The output variables are six features of Reynolds stress tensor (magnitude (k), shape (η, ξ), and direction (φ_1, φ_2, φ_3), which can be obtained by decomposing Reynolds stress tensor. The input variables are the features extracted from the flow field which need to be corrected. Since the flow field data is usually massive, it is crucial to reduce the dimension of the input data. In this paper, eight types of features are selected to represent the main physical characteristics of the flow field. Both the input variables and output variables satisfy Galilean invariance, which can improve the prediction accuracy and generalization performance of the RF regression model.

By using the prediction model, the important information of turbulent disturbance state in boundary layer and the direction and magnitude of stress pulsation during the generation of maximum/minimum turbulent kinetic energy are obtained effectively, and it's feasible to be applied in the prediction of different Reynolds number or even different geometry model as long as there are enough training scenarios.

More importantly, most of previous studies have been based on steady-state classical flows, such as flow around a cylinder, back step flow, periodic hill flow, etc. The innovation of this paper lies in the use of transient simulated flow around a near-wall cylinder. Unlike ordinary flow around a cylinder, which changes with time, the flow around a near-wall cylinder takes into account the interaction between the near-wall boundary layer and the wake region of the flow around a cylinder. Therefore, this kind of flow has no regularity with time and is a classic turbulent flow which is characterized by randomness. In future work, machine learning methods can be used to predict each time step (that is, the flow field at each moment), providing new ideas for machine learning methods to predict Reynolds stress anisotropy of transient flows.

Acknowledgments

This work was supported in part by CSC (China Scholarship Councill) Scholarship.

References

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(Volume 52, 2020):477–508, 2020. ISSN 1545-4479. doi: https://doi.org/10.1146/annurev-fluid-010719-060214.

Mihailo R. Jovanović. From bypass transition to flow control and data-driven turbulence modeling: An input–output viewpoint. *Annual Review of Fluid Mechanics*, 53(Volume 53, 2021):311–345, 2021. ISSN 1545-4479. doi: https://doi.org/10.1146/annurev-fluid-010719-060244.

J. Ling and J. Templeton. Evaluation of machine learning algorithms for prediction of regions of high Reynolds averaged Navier Stokes uncertainty. *Physics of Fluids*, 27(8):085103, 08 2015. ISSN 1070-6631. doi: 10.1063/1.4927765.

Julia Ling, Andrew Kurzawski, and Jeremy Templeton. Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 807:155–166, 2016. doi: 10.1017/jfm.2016.615.

F. Durst S. Banerjee, R. Krahl and Ch. Zenger. Presentation of anisotropy properties of turbulence, invariants versus eigenvalue approaches. *Journal of Turbulence*, 8:N32, 2007. doi: 10.1080/14685240701506896.

Anand Pratap Singh and Karthik Duraisamy. Using field inversion to quantify functional errors in turbulence closures. *Physics of Fluids*, 28(4):045110, 04 2016. ISSN 1070-6631. doi: 10.1063/1.4947045.

Hongwei Tang, Yan Wang, Tongguang Wang, Linlin Tian, and Yaoru Qian. Data-driven Reynolds-averaged turbulence modeling with generalizable non-linear correction and uncertainty quantification using Bayesian deep learning. *Physics of Fluids*, 35(5):055119, 05 2023. ISSN 1070-6631. doi: 10.1063/5.0149547.

Steven L. McKeon Beverley J. Vinuesa, Ricardo Brunton. The transformative potential of machine learning for experiments in fluid mechanics. 5:536–545. ISSN 2522-5820. doi: 10.1038/s42254-023-00622-y.

HONGYI XU. Direct numerical simulation of turbulence in a square annular duct. *Journal of Fluid Mechanics*, 621:23–57, 2009. doi: 10.1017/S0022112008004813.