Evaluation of machine learning algorithms for predictive Reynolds stress transport modeling

J P PANDA, H V WARRIOR

Department of Ocean Engineering and Naval Architecture
Indian Institute of Technology Kharagpur, India

† *Corresponding Author Email*: *jppanda.iit@gmail.com*

(Received −; accepted −)

ABSTRACT

The application machine learning (ML) algorithms to turbulence modeling has shown promise over the last few years, but their application has been restricted to eddy viscosity based closure approaches. In this article we discuss rationale for the application of machine learning with high-fidelity turbulence data to develop models at the level of Reynolds stress transport modeling. Based on these rationale we compare different machine learning algorithms to determine their efficacy and robustness at modeling the different transport processes in the Reynolds Stress Transport Equations. Those data driven algorithms include Random forests, gradient boosted trees and neural networks. The direct numerical simulation (DNS) data for flow in channels is used both as training and testing of the ML models. The optimal hyper-parameters of the ML algorithms are determined using Bayesian optimization. The efficacy of above mentioned algorithms is assessed in the modeling and prediction of the terms in the Reynolds stress transport equations. It was observed that all the three algorithms predict the turbulence parameters with acceptable level of accuracy. These ML models are then applied for prediction of the pressure strain correlation of flow cases that are different from the flows used for training, to assess their robustness and generalizability. This explores the assertion that ML based data driven turbulence models can overcome the modeling limitations associated with the traditional turbulence models and ML models trained with large amounts data with different classes of flows can predict flow field with reasonable accuracy for unknown flows with similar flow physics. In addition to this verification we carry out validation for the final ML models by assessing the importance of different input features for prediction.

Keywords: Machine learning, Pressure strain correlation,Turbulence modeling, CFD

NOMENCLATURE

R_{ij}	Reynolds stress tensor
b_{ij}	Reynolds stress anisotropy
δ_{ij}	Kronecker delta
k	turbulent kinetic energy
ρ	density
ν	Kinetic viscosity
U_i	Mean velocity component
u_i	Fluctuating velocity component
P_{ij}	Production of turbulence
T_{ijk}	Diffusive transport
ε_{ij}	Dissipation rate tensor
ϕ_{ij}	Pressure strain correlation

1. INTRODUCTION

Computational fluid dynamics(CFD) utilizes numerical techniques to model and solve problems that involve flow of fluids. CFD based methods(Versteeg and Malalasekera 2007) have became popular with the advancement in computational facilities and formulation of models for turbulence(Pope 2000). The CFD based techniques are widely used in the early years of 21st century. The success of CFD techniques is largely dependent on the development of the techniques to model the turbulence. Mainly there are four

broad classes of CFD approaches in the turbulence modeling framework, those are eddy viscosity models(Pope 1975; Pope 2000; Panda, Sasmal, Maity, and Warrior 2020), Reynolds stress transport models(Mishra and Girimaji 2017; Panda and Warrior 2018), Large eddy simulation(Pope 2000) and Direct numerical simulation(Lee and Moser 2015; Lee and Moser 2018). In industrial applications eddy viscosity based are widely used, since those have least level of complexity and a few number of equations are solved for the modeling of the flow field. In recent years with booming in computational facilities researchers have started using LES for flow prediction in complex and larger domains. But LES can not be used for modeling and prediction complex flows of Engineering interest, since the cost of simulations of LES is very high. The LES and DNS data are mainly used for development and calibration of new turbulence models. In contrast to all the above mentioned approaches, the Reynolds stress transport models(Panda and Warrior 2018; Mishra and Girimaji 2017) are superior to eddy viscosity models, since in these models equations are solved for all the components of Reynolds stresses from which the Reynolds stress field is approximated. The building block of such models are the Reynolds stress transport equations(Panda 2019), in which, the two important terms that need to be modeled are the pressure strain correlation(Panda 2020) and the dissipation term. The models for all these terms are developed by calibrating with few cases of experiments such as grid turbulence, grid turbulence with mean strain and shear flows. So, when these models are applied to other realistic cases of flow predictions may provide inaccurate results.

To overcome these difficulties recent emphasis of turbulence modelers have been shifted towards development of data-driven turbulence models(Duraisamy, Iaccarino, and Xiao 2019). Machine learning is a process of function approximation, the function can be approximated by correlating the input parameters with the output. There are various machine learning algorithms. Few important of those are neural networks, Random forests(Wang, Wu, and Xiao 2017; Kaandorp and Dwight 2020; Luan and Dwight 2020), gradient boosted trees(Wu, Li, Qiu, and Liu 2020), Gene expression programming(Zhao, Akolekar, Weatheritt, Michelassi, and Sandberg 2020) and sparse symbolic regression(Schmelzer, Dwight, and Cinnella 2020). In turbulence modeling, the problem is posed as one of supervised learning, where the model attempts to minimize the prediction error over a training data-set. This data required for training, validation and testing the ML models can be obtained from experimental investigations, high fidelity DNS and LES data-sets of turbulent flows. In the context of supervised learning, data driven approaches of turbulence modelling can be classified into three main categories a) quantification of uncertainties in the RANS models (with this approach the uncertainties in Reynolds stress tensor can be quantified), b) finding the discrepancies in the model coefficients and magnitude of the terms in the governing equation and c) direct modeling of turbulence parameters such as Reynolds stress(EVM), pressure strain correlation (RSTM) and the subgrid scale stress(LES) in terms of the mean flow parameters. Among the three mentioned approaches of data-driven turbulence modeling, the models developed through the third approach can be trained with huge database of turbulence and can be applied for prediction of unseen flow cases of similar flow physics.

(Singh, Medida, and Duraisamy 2017) developed model augmentations for Spalart-Allmaras(SA) turbulence model using adjoint based full field inference using experimentally measured lift coefficient data. These models forms are reconstructed using neural networks, and applied in CFD solver, to predict flow in different operating conditions. (Tracey, Duraisamy, and Alonso 2015) used a shallow neural network (one hidden layer) to model the source terms of the SA turbulence model. (Parish and Duraisamy 2016) learned a turbulence production term using machine learning and applied that to the k-equation of the $k-\omega$ turbulence model. (Maulik, San, Rasheed, and Vedula 2018) used neural networks to model eddy viscosities for RANS simulations. (Ling, Kurzawski, and Templeton 2016) used a special kind of neural network called tensor basis neural network to model the Reynolds stress anisotropy and also compared the model prediction against a simple multi layer perception. The optimal hyperparameters of the model were recommended using Bayesian optimization. (Zhu, Zhang, Kou, and Liu 2019) used neural networks to construct a mapping function between the turbulent eddy viscosity and the mean flow variables and ML model completely replaces the partial differential equation model. (Wu, Xiao, and Paterson 2018) proposed a physics based implicit treatment to model Reynolds stresses using random forests. The optimal eddy viscosity and nonlinear part of the Reynolds stresses were both predicted. They used seven distinct physics based features to model the Reynolds stresses, those are strain rate tensor, rotation rate tensor, pressure gradient, turbulence kinetic energy gradient, wall distance based

Reynolds number, turbulence intensity and ratio of turbulent time-scale to mean strain time-scale. They used different approaches to achieve in-variance in machine learning. (Weatheritt and Sandberg 2016) used Gene expression programming to formulate non-linear consecutive stress-strain relationship. The mathematical model was created by using high fidelity and uncertainty measures. The learning method has the capability to produce a constraint free model. (Weatheritt and Sandberg 2017) used symbolic regression to model the algebraic form of the Reynolds stress anisotropy tensor. The equations were trained using hybrid RANS/LES data and the new model was employed in RANS closure to test the prediction of model in 3D geometries. (Schmelzer, Dwight, and Cinnella 2020) discovered algebraic Reynolds stress models using sparse regression and they have used high fidelity LES/DNS data for training and cross validation of the model. There case of separated flows were considered, those are periodic hills, converging-diverging channel and curved backward facing step. The prediction of the machine learnt model was better than the $k-omega$ SST model. (Huijing, Dwight, and Schmelzer 2021) used the model developed by (Schmelzer, Dwight, and Cinnella 2020) to predict the fully three dimensional high Reynolds number flows, e.g. wall mounted cubes and cuboids. (Fang, Sondak, Protopapas, and Succi 2020) used neural networks to model the Reynolds stress anisotropy using neural networks. They proposed different modification of the neural network structure to accommodate effect of Reynolds number, non-locality and wall effects into the modeling basis. With such distinct feature injection, significant improvement of model prediction was observed. (Beck, Flad, and Munz 2019) proposed a novel data-driven strategy for turbulence modeling for LES using artificial neural networks.

The methodology used in many of the studies discussed till now uses large high fidelity datasets obtained from DNS or LES simulations(Parashar, Srinivasan, and Sinha 2020). This data is used to train and validate machine learning based models, that can vary from deep learning based models or ensembled meta-models. The training of this model involves the inference of optimal coefficients for the closure of the turbulence model. In these studies, the turbulence model form corresponds to 2-equation eddy viscosity based models (EVM) or Algebraic Reynolds Stress Models (ARSM). While these studies have had considerable success, this outlined methodology may be hamstrung by the discordance between the high fidelity of datasets and the fidelity of the baseline model form. As an illustration DNS data is able to replicate the high degree of anisotropy in turbulent flows. However eddy viscosity based models are not capable of replicating a high degree of turbulence anisotropy due to the linear eddy viscosity hypothesis(Mishra, Duraisamy, and Iaccarino 2019). The eddy viscosity hypothesis assumes that turbulence anisotropy is a linear function of the instantaneous mean strain rate. Thus the anisotropy predicted by any eddy viscosity based model is forced to lie on the plane strain segment on the barycentric triangle (Edeling, Iaccarino, and Cinnella 2018). As a result the information regarding the turbulence anisotropy in the high fidelity data is ignored by the machine learning algorithm, due to the model form of the baseline eddy viscosity closure. Additionally the high fidelity data reflects the complicated dependence of turbulent statistics on streamline curvature and the mean rate of rotation. This is important information to model the mean and frame rotation effects on turbulence evolution. However eddy viscosity based closures model the Reynolds stresses as functions of the mean rate of strain only(Pope 2001). Thus the information regarding the effect of streamline curvature or frame rotation is ineffectual because of the form of the baseline model (Mishra and Girimaji 2019). Algebraic Reynolds Stress Models assume that turbulent diffusive and convective fluxes are relatively small (explicitly that the flow is source dominated (Gatski and Speziale 1993)). This is a restrictive assumption and is not valid for most turbulent flows. In light of these illustrations the adoption of a different baseline model formulation, flexible enough to take advantage of the high fidelity information in the data may be advisable. An alternative that can meet these requirements is the Reynolds Stress Modeling approach. Instead of assuming the form of relationship between the Reynolds stresses and the mean gradient, the Reynolds stress models use the Reynolds Stress Transport Equations as evolution equations for components of turbulent anisotropy. Explicit computation of the evolution of components of turbulent anisotropy results in improved representation of anisotropy. Reynolds stress models can account for directional effects of the Reynolds stresses. Reynolds stress models can represent the limiting states of turbulent flow, for example the return to isotropy of turbulence in decaying turbulent flows, and the Rapid Distortion Limit where turbulence behaves like an elastic medium (instead of a viscous medium). Because of the separate modeling of the turbulent transport processes, Reynolds stress models account for effects of flow stratification, buoyancy, streamline curvature, etc. Thus in a machine learning framework, utilizing Reynolds Stress Models as the base-

line model can enable the algorithm to exploit a substantially higher proportion of physics and information in the high fidelity data.

However there has been little research to extend the potential of Reynolds Stress Modeling approach by utilizing machine learning algorithms. This is a central novelty of this investigation.

In the Reynolds Stress Modeling approach separate models are formulated for the terms in the Reynolds Stress Transport Equation, where each such term represents a different turbulence transport process. These transport processes include turbulent diffusion, rotational effects, rate of dissipation and the pressure strain correlation. While high fidelity models for all these terms are important, accurate and robust modeling of the pressure strain correlation term is a long standing challenge in turbulence modeling. The pressure strain correlation term represents physics responsible for the transfer of energy between different components of the Reynolds stress tensor (Mishra and Girimaji 2014). It is responsible for the non-local interactions in turbulent flows, the initiation of instabilities in rotation dominated flows, the return to isotropy observed in decaying flows, etc (Mishra and Girimaji 2013). While classical models have been developed for the pressure strain correlation term, such physics driven models have many limitations in their ability to account for streamline curvature effects, realizability requirements, their performance in complex engineering flows (Mishra and Girimaji 2017). In this work we have modeled the pressure strain correlation of turbulence using three different machine learning approaches, those are artificial neural network, random forest and gradient boosted decision trees. The ML models developed with above mentioned algorithms are trained and tested for DNS data of turbulent channel flow at different Reynolds numbers. We have grouped the data-sets in 4 different combination to perform 4 distinct training and testing of the ML models. The Bayesian hyper-parameter optimization was also used to find the best hyper-parameters of the ML models that perform well for unknown prediction problems. One main advantage of hyper-parameter optimization is that it reduces the chance of overfitting in tree based algorithms and also enhance the generalizability of the model.

The remainder of the article is summarized as follows: Section 2 provides a background on the turbulence modeling framework. In Section 3 the limitations of the turbulence models in predicting complex flow fields is discussed. Section 4 describes the different machine learning algorithms used in modeling the turbulence parameters. In section 5 the detailed methods of optimization of the hyper-parameters of the ML models are discussed. In the subsequent sections the training-testing of the models and feature importance are discussed followed by conclusion and future scope.

2. REYNOLDS STRESS TRANSPORT MODELING FRAMEWORK

Different fidelities of turbulence simulation approaches can be differentiated based on the scales of motion that are directly simulated based on their governing equations and the scales of motion that are modeled. For example, in Direct Numerical Simulation, all scales of motion are computed. In Reynolds Averaged Navier Stokes (RANS) approaches, all turbulent scales of motion are modeled. RANS closures can carry out this modeling of turbulent scales of motion in different manner. One equation models may use the concept of mixing length to model turbulent scales of motion. Two-equation models like the $k-\varepsilon$ and the $k-\omega$ models use the eddy viscosity hypothesis to model turbulent motions. Reynolds Stress Models are based off the Reynolds Stress Transport Equations. Here, each turbulent transport process is present as a separate term in the evolution equations. In Reynolds stress transport modeling, such transport terms such as pressure strain correlation and dissipation tensors can be modeled using data driven approaches. In this section we discuss the traditional method of modeling of the pressure strain correlation starting from the basic governing equations.

The Reynolds stress transport model(Mishra and Girimaji 2010; Panda and Warrior 2018; Panda, Warrior, Maity, Mitra, and Sasmal 2017) does not rely upon any ad hoc definitions of turbulent stresses in terms of strain field, rather the stress field is directly computed from the modeled transport equation of Reynolds stress components. The Reynolds stress transport equation has four terms in its right side, those are production, transport, dissipation and the pressure strain correlation term. The pressure strain correlation term mainly accounts for the complex flow features resulting from the stream line curvature and flow separation. The Reynolds stress model much more reliable and accurate than its two equation counter parts. The Reynolds stress transport equation

has the form(Panda 2019):

$$\partial_t \overline{u_i u_j} + U_k \frac{\partial \overline{u_i u_j}}{\partial x_k} = P_{ij} - \frac{\partial T_{ijk}}{\partial x_k} - \varepsilon_{ij} + \phi_{ij},$$

where,

$$\begin{aligned} P_{ij} &= -\overline{u_k u_j}\frac{\partial U_i}{\partial x_k} - \overline{u_i u_k}\frac{\partial U_j}{\partial x_k}, \\ T_{kij} &= \overline{u_i u_j u_k} - \nu\frac{\partial \overline{u_i u_j}}{\partial x_k} + \delta_{jk}\overline{u_i \frac{p}{\rho}} + \delta_{ik}\overline{u_j \frac{p}{\rho}} \\ ,\varepsilon_{ij} &= -2\nu\overline{\frac{\partial u_i}{\partial x_k}\frac{\partial u_j}{\partial x_k}}, \\ \phi_{ij} &= \overline{\frac{p}{\rho}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)} \end{aligned} \tag{1}$$

P_{ij} denotes the production of turbulence, T_{ijk} is the diffusive transport, ε_{ij} is the dissipation rate tensor and ϕ_{ij} is the pressure strain correlation. The pressure fluctuations are governed by a Poisson equation:

$$\frac{1}{\rho}\nabla^2(p) = -2\frac{\partial U_j}{\partial x_i}\frac{\partial u_i}{\partial x_j} - \frac{\partial^2 u_i u_j}{\partial x_i \partial x_j} \tag{2}$$

The fluctuating pressure term is split into a slow and rapid pressure term $p = p^S + p^R$. Slow and rapid pressure fluctuations satisfy the following equations

$$\frac{1}{\rho}\nabla^2(p^S) = -\frac{\partial^2}{\partial x_i \partial x_j}(u_i u_j - \overline{u_i u_j}) \tag{3}$$

$$\frac{1}{\rho}\nabla^2(p^R) = -2\frac{\partial U_j}{\partial x_i}\frac{\partial u_i}{\partial x_j} \tag{4}$$

It can be seen that the slow pressure term accounts for the non-linear interactions in the fluctuating velocity field and the rapid pressure term accounts for the linear interactions(Mishra and Girimaji 2015). The pressure strain correlation is modeled using rational mechanics approach. The rapid term can be modeled as(Pope 2000)

$$\phi^R_{ij} = 4k\frac{\partial U_l}{\partial x_k}(M_{kjil} + M_{ikjl}) \tag{5}$$

where,

$$M_{ijpq} = \frac{-1}{8\pi k}\int \frac{1}{r}\frac{\partial^2 R_{ij}(r)}{\partial r_p \partial r_p} dr \tag{6}$$

where, $R_{ij}(r) = \langle u_i(x)u_j(x+r)\rangle$ For homogeneous turbulence the complete pressure strain correlation can be written as

$$\phi_{ij} = \varepsilon A_{ij}(b) + kM_{ijkl}(b)\frac{\partial \overline{v}_k}{\partial x_l} \tag{7}$$

The most general form of slow pressure strain correlation is given by

$$\phi^S_{ij} = \beta_1 b_{ij} + \beta_2(b_{ik}b_{kj} - \frac{1}{3}II_b\delta_{ij}) \tag{8}$$

Established slow pressure strain correlation models including the models of (Rotta 1951) and (Sarkar and Speziale 1990; Panda, Warrior, Maity, Mitra, and Sasmal 2017; Warrior, Mathews, Maity, and Sasmal 2014) use this general expression. Considering the rapid pressure strain correlation, the linear form of the model expression is

$$\begin{aligned} \frac{\phi^R_{ij}}{k} &= C_2 S_{ij} + C_3(b_{ik}S_{jk} + b_{jk}S_{ik} - \\ &\frac{2}{3}b_{mn}S_{mn}\delta_{ij}) + C_4(b_{ik}W_{jk} + b_{jk}W_{ik}) \end{aligned} \tag{9}$$

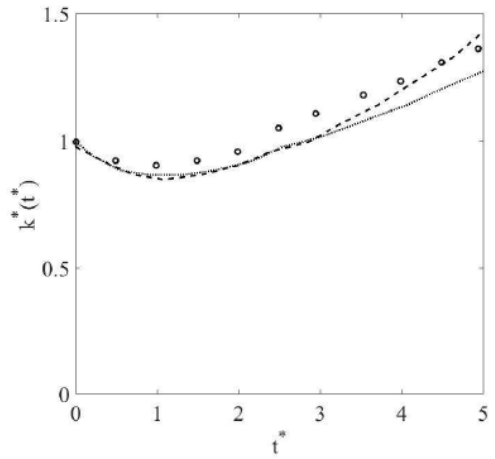
Here $b_{ij} = \frac{\overline{u_i u_j}}{2k} - \frac{\delta_{ij}}{3}$ is the Reynolds stress anisotropy tensor, S_{ij} is the mean rate of strain and W_{ij} is the mean rate of rotation. Rapid pressure strain correlation models like the models of (Mishra and Girimaji 2017; Speziale, Sarkar, and Gatski 1991; Panda and Warrior 2018) use this general expression. For this work the Reynolds stress model of (Speziale, Sarkar, and Gatski 1991) is used which has the form:

$$\begin{aligned} \phi^{(R)}_{ij} &= (C_1 - C^*_1 II^{0.5})KS_{ij} + \\ &C_2K(b_{ik}S_{jk} + b_{jk}S_{ik} - 2/3b_{mn}S_{mn}\delta_{ij}) \\ &+ C_3K(b_{ik}W_{jk} + b_{jk}W_{ik}) \end{aligned} \tag{10}$$

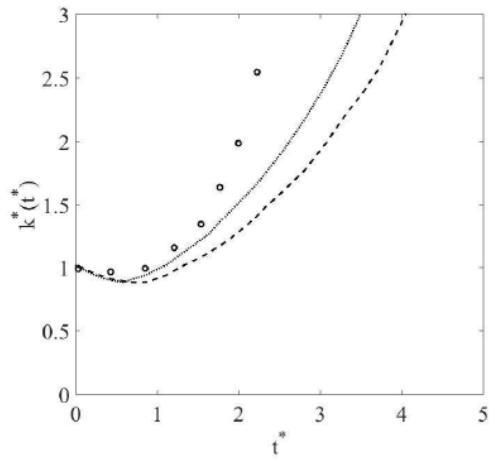
The closure coefficients are taken as $C_1 = 0.8$, $C^*_1 = 1.3$, $C_2 = 1.25$ and $C_3 = 0.4$.

3. Limitation of the existing pressure strain correlation model in the prediction of complex turbulent flows

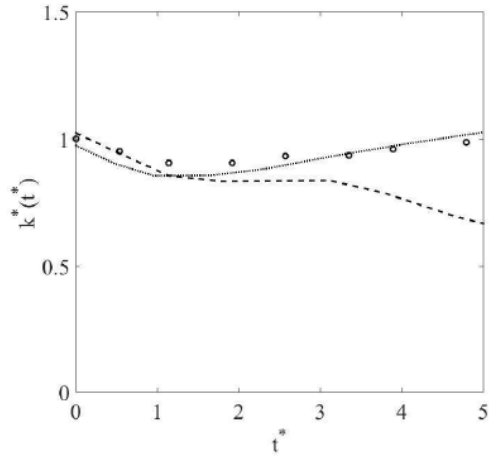
Most of the Reynolds stress transport models available in literature have some tuned coefficients, those are obtained from the calibration of set of equations against few standard experimental/DNS data-sets like flow in channels, grid turbulence and grid turbulence, so when those models are tested for complex flow configurations produce unrealistic results. The Reynolds stress transport models are also associated with another type of uncertainty, i.e. those can not accurately capture the non-local nature of flow(The flow at one point may be affected by flow physics of upstream points), since the model form does not have any terms for accommodation of the non-local nature



(a)



(b)



(c)

Fig. 1. Evolution of turbulence kinetic energy for rotating shear flows, a) $W/S = 0$, b) $W/S = 0.25$ and c) $W/S = 0.5$, Symbols LES data, dashed lines predictions of LRR model, dotted lines predictions of SSG model.

Case	Training Set	Testing Set
1	$Re_\lambda = 550, 1000, 2000$	$Re_\lambda = 5200$
2	$Re_\lambda = 550, 1000, 5200$	$Re_\lambda = 2000$
3	$Re_\lambda = 550, 2000, 5200$	$Re_\lambda = 1000$
4	$Re_\lambda = 1000, 2000, 5200$	$Re_\lambda = 550$

Table 1. Four training and test cases for the turbulent channel flow

of flow. In order to overcome these limitations of the Reynolds stress transport models data driven models of the pressure strain correlation model can be developed using machine learning approaches.

We have performed numerical simulations to check the predictive capability of different pressure strain correlation models in rotation dominated flow fields. It is noticed that, there is a strong disparity between the model predictions and the LES data particularly at higher values of rotation to strain ratio. The results of those comparisons are presented in fig.1. fig.1 a, b and c correspond to rotation to strain ratio of 0, 0.25 and 0.5 respectively. The symbols in the fig.1 represent the LES results of(Bardino, Ferziger, and Reynolds 1983).

4. Description of the Data Set

The dataset used in the modeling of the pressure strain correlation of turbulence was obtained from the Oden Institute Turbulence File Server(Lee and Moser 2015), in which DNS simulations were performed for flow in channels with friction Reynolds number ranging from 550 to 5200. The DNS simulations were performed with a B-spline collocation method in the wall-normal direction and for the stream-wise and span-wise directions Fourier-Galerkin method was used. The detailed information on the on the methods employed in the DNS of channel flow is available in (Lee and Moser 2015). We have used four distinct data-sets for training and testing of the ML models. The four training and testing data-sets are presented in table1.

5. Physics Based Input Features for ML Models

The turbulence parameter e.g. pressure strain correlation can be accurately modeled using suitable input features. The most important features for the modeling of the pressure strain correlation can be the Reynolds stress anisotropy, dissipation, velocity gradient and the turbulence kinetic energy. The input

features must be wisely chosen so that there is no over-fitting. Secondly this ensures that physics based constraints are met in the final model. For instance due to Galilean in-variance we should ensure that the features in the modeling basis also obey this requirement. The functional mapping for the ML models can be written as:

$$\phi_{uv} = f_1(b_{uv}, \varepsilon, \frac{du}{dy}, k) \tag{11}$$

We have used the formula: $\alpha^* = \frac{\alpha - \alpha_{min}}{\alpha_{max} - \alpha_{min}}$ for normalizing the inputs to the ML models, so that those will be in the range 0 and 1. This avoids clustering of training in one direction and enhance convergence in the training.

6. DATA DRIVEN TURBULENCE MODELING WITH MACHINE LEARNING

In turbulence modeling frameworks such as eddy viscosity and Reynolds stress transport models the Reynolds stress and pressure strain correlation terms plays major role and with accurate modeling of those two parameters, the model form uncertainties can be eradicated. In this work, we have modeled above mentioned terms using three different machine learning algorithms, those are deep neural networks, Random forests and gradient boosted trees. In turbulence modeling and fluid dynamics applications deep neural networks are widely used. Next, we have used random forests in the models, those have the capability to check the feature importance, i.e. which of the input features has dominant correlation with the output. We also have considered gradient boosted trees in the modeling of the turbulence parameters. The basic difference between RF and GB trees is that RF creates random training samples from full training set based on Bootstrap Aggregating(Bagging) and a weak learner is trained in parallel and in GB trees ensemble methods consist in fitting several weak learners sequentially as shown in figure. 2.

6.1 Artificial Neural Networks

Artificial Neural Networks(ANN)(fig.3) are the machine learning systems inspired from the biological neural networks, these are also known as multilayer perception(MLP). The biological neural networks(BNN) are the circuits that carry out a specific task when activated. These are population of neurons interconnected by synapses. Similar to BNNs, ANN/MLP have artificial neurons. A very basic unit of a MLP is a perceptron as shown in fig.3a. The input data from a (l-1)th layer are multiplied by a weight W, which are linearly combined and allowed to pass through a non-linear activation function η:

$$q_i^l = \eta(\sum_j W_{ij}^l q_j^{l-1}). \tag{12}$$

A MLP can be constructed by combining a number of perceptrons. A typical MLP is shown in fig.3b. The weights in a MLP are optimized to minimize a cost function E by back propagation. There are several optimization techniques by which the weights can be calculated. Those are gradient descent, Quasi- Newton, Stochastic Gradient Descent or the Adaptive moment Estimation etc. The activation functions, which can be used in MLP are sigmoid $\eta(\beta) = 1/(1+e^{-\beta})$, hyperbolic tangent(tanh) $\eta(\beta) = (e^{-\beta} - e^{-\beta}).(e^{-\beta} + e^{-\beta})$ and RELU $\eta(\beta) = max[0, \beta]$.

6.2 Random forests(RF)

The random forest algorithm was proposed by (Breiman 2001). The basic building block of a random forest is a decision tree. The structure of the decision tree is shown in fig.4a. The boxes in the decision tree represent group of features and data. The decision tree seeks if/then/else rules for obtaining the desired output. Random forest(RF)(fig.4b) regression combines performance of multiple decision trees for predicting a output variable. It is a assemble learning technique, works on the concept of bagging method. In RF regression trees are constructed using a subset of random sample drawn with replacement from training data. In RF to depict the growth of the tress random vectors are generated and the trees are not allowed to prun. To perform splitting of the dataset, a random combination of features is selected at each and every node.

6.3 Gradient boosted decision trees(GBDT)

Similar to random forest algorithm, the building block of GBDT are the decision trees. Random forests utilizes method of bagging to combine many decision trees to create an ensemble. Bagging simply means combining the decision trees in parallel. However, boosting means combining decision trees in series to achieve strong learner. The decision trees are the weak learners. The boosting algorithms learn slowly, since trees are added sequentially. Each tree in the boosting trees focus on errors from previous one, making the boosting algorithm an efficient and accurate model. Boosting process is slow, since the trees in the GBDT are added sequentially.

7. HYPER-PARAMETER OPTIMISATION

The hyper-parameter optimization is the process of finding the best hyper-parameters for a machine

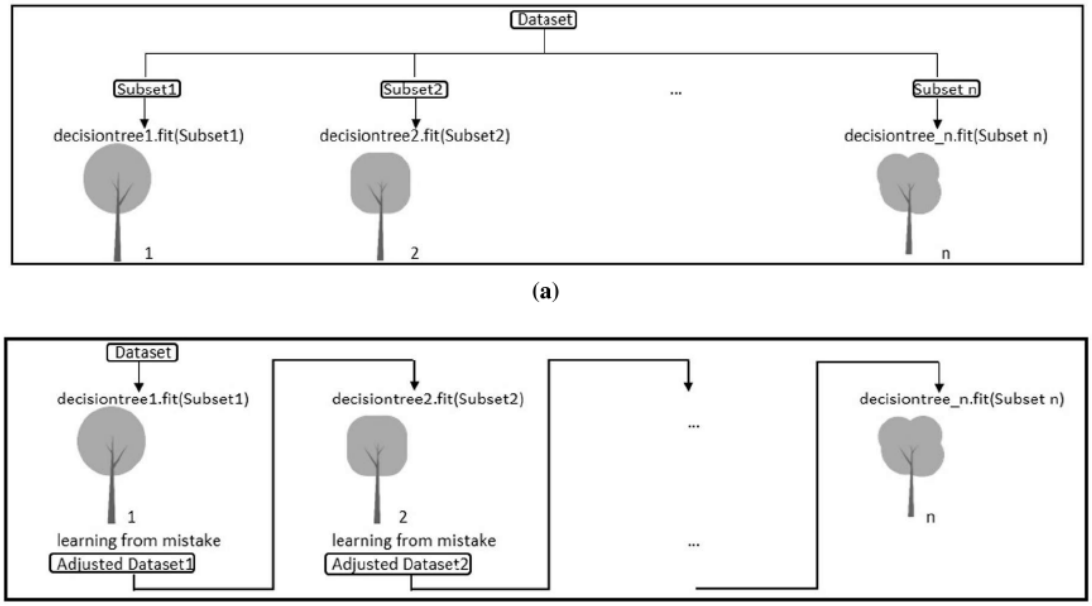


Fig. 2. a) Bagging vs b) Boosting

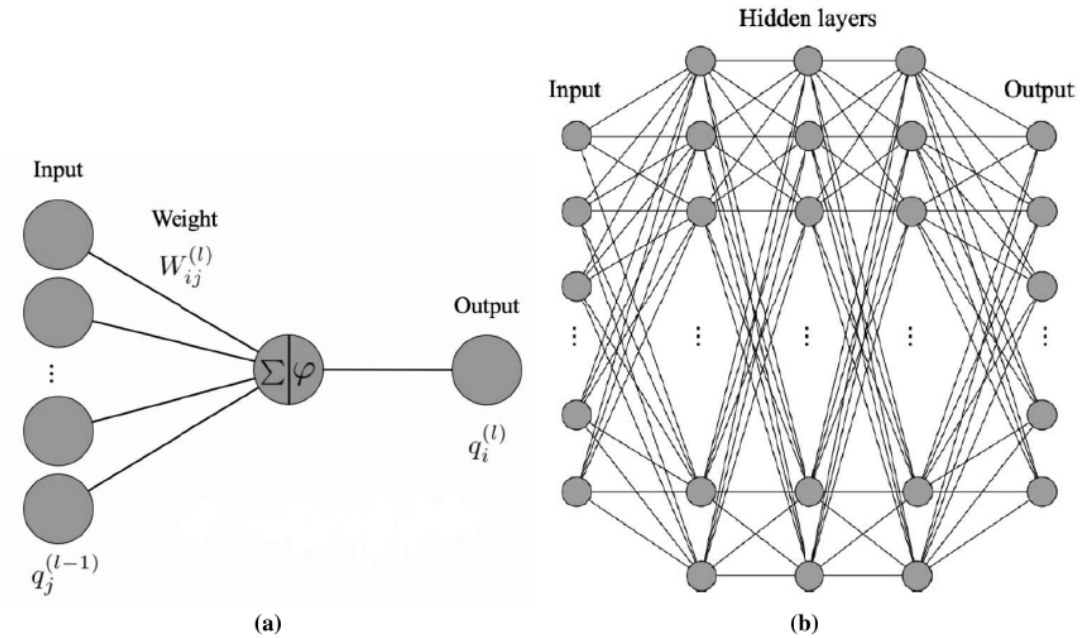


Fig. 3. a) A perceptron b) An artificial neural network(Multi layer perception)

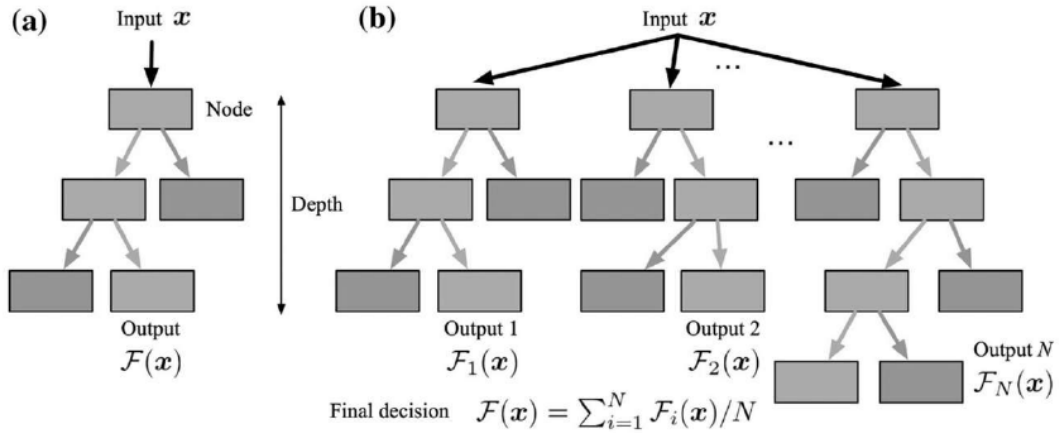


Fig. 4. Architecture of of a random forest, a)a single decision tree b)a random forest

learning algorithm that returns best performance while measured on a validation set. The hyper-parameters are turned by application engineers during model developments. There are mainly two methods of hyper-parameter tuning, those are either performed manually or done by some optimization algorithm. The automated algorithms are grid search, random search and Bayesian optimization. The grid search and random search not informed by past evaluations. The Bayesian optimization keep track of past evaluations, hence we have considered the Bayesian optimization for automated hyper-parameter tuning.

7.1 Manual search

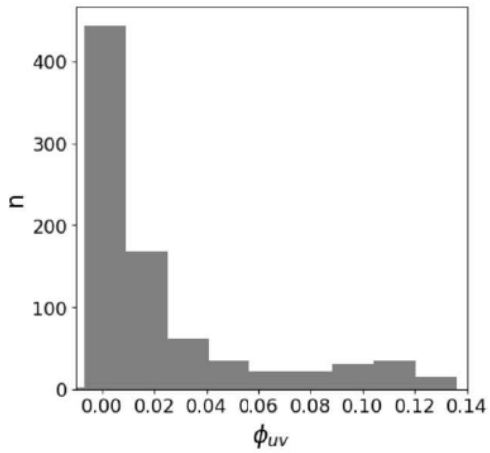
The hyper-parameters of different machine learning algorithms, which need to finely tuned for improving the predictive capability of any machine learning model. For neural networks the important hyper-parameters are number of layers and number of neurons in each layer. Similarly for random forest and gradient boosted decision trees the hyper-parameters are number of decision trees and maximum depth of the trees. Few other parameters are also important in tree algorithms, those are minimum sample leaf and learning rate for gradient boosting decision trees. However fine tuning of all such parameters using manual search method is tedious task and also there is also the chance of over fitting. In manual search, we have considered R^2 as the criteria for finding the hyper-parameters. By manual search the optimal hyper-parameters are tuned as follows: for neural network: 5 layers with 10 neurons in each layer and for both RF and GBDT the number of decision trees and max depth are taken as 5 and 10 respectively. By considering above mentioned hyper-parameters the R^2 and MSE values for test data are calculated as follows for case 4: ANN: 0.9897, $9.97 * 10^{-6}$, RF: 0.9841, $1.66 * 10^{-5}$, GBDT: 0.981, $1.99 * 10^{-5}$.

7.2 Bayesian optimization

In Bayesian optimization, the parameter space is sampled and a gradient boosted tree was constructed and trained. The performance of the model is evaluated using validation data. The optimal hyper-parameters are those, that yield the lowest error on the validation data. The model validation error was treated as a sample from Gaussian process at each hyper-parameter setting. In contrast to other optimization techniques, the Bayesian optimization requires relatively lesser number of model evaluations, since at each step the information from all previous states is utilized to inform the GP model of the validation error. This method is highly useful, when computational cost of evaluation of objective function is very high. The detailed steps in Bayesian optimization are as follows:

1. Build a surrogate probability model of the objective function
2. Finding the hyper-parameters that performs best on the surrogate model
3. Application of these hyper-parameters to the true objective function
4. Updating of the surrogate model incorporating the new results
5. The steps 2 and 4 must be repeated until max iterations or time is reached

The hyper-parameters mentioned in table 2 are only



(a)

Fig. 5. Histogram of the data variability for training data in case 1, There are very few training samples for high values of pressure strain correlation.

ML models	Parameter I	parameter II
MLP	N. of neurons	Layers
RF	N. of decision trees	Depth
GBDT	N. of decision trees	Depth

Table 2. The hyper-parameters of the machine learning models. (N. stands for number).

a limited set of important parameters, through which the model can be developed. The models with those limited tuned parameters may not perform well for unknown prediction data-sets on which the model is trained. To enhance the predictive capability of any machine learning model at least few other hypo-parameters must be considered in the model development stage. However, with increase in number of hyper-parameters, manual tuning may be a tedious process and advanced optimization techniques like Bayesian optimisation must be used for fine turning of the hyper-parameters. For brevity, we have only considered the hyper-parameter optimization of gradient boosted regression tree using Bayesian optimization. In addition to depth and number of decision trees, three other hypo-parameters are considered, those are max features, minimum simple leaf and minimum sample split. From Bayesian optimization, the optimal hyper-parameters(maximum depth, Number of estimators, minimum sample split, minimum sample leaf, maximum features) are (66, 63, 9, 5, sqrt). The maximum depth correspond to highest depth of the individual regression tree. The best value of this parameter depends on the interaction of the input variables. Number of estimators are the number of boosting stages used in the model development. The larger is the number of the estimators, better is the performance of the the model. The minimum sample leaf is the minimum number of minimum number of samples required to split an internal node. The other terms has their usual meaning. A detailed discussion of all these hyper-parameters of GBDT is available in (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay 2011). The R^2 values of the optimized GBDT was found to be 0.9915 and 0.9076 for the training and testing respectively for case 4. The predictions of the GBDT with optimized hyper-parameters for unknown flow cases are summarized in section 9..

8. TRAINING OF THE ML MODELS

All the three ML models are trained for the DNS data of turbulent channel flow for four different test cases. The four different cases are shown in table 1. Since the turbulence statistics are available for four different friction Reynolds numbers, in each training and testing phase, we have considered data for three friction Reynolds numbers as training and the other was considered for testing. We have considered two different approaches for finding the hyper-parameters of the ML models as discussed in section 6. Using the

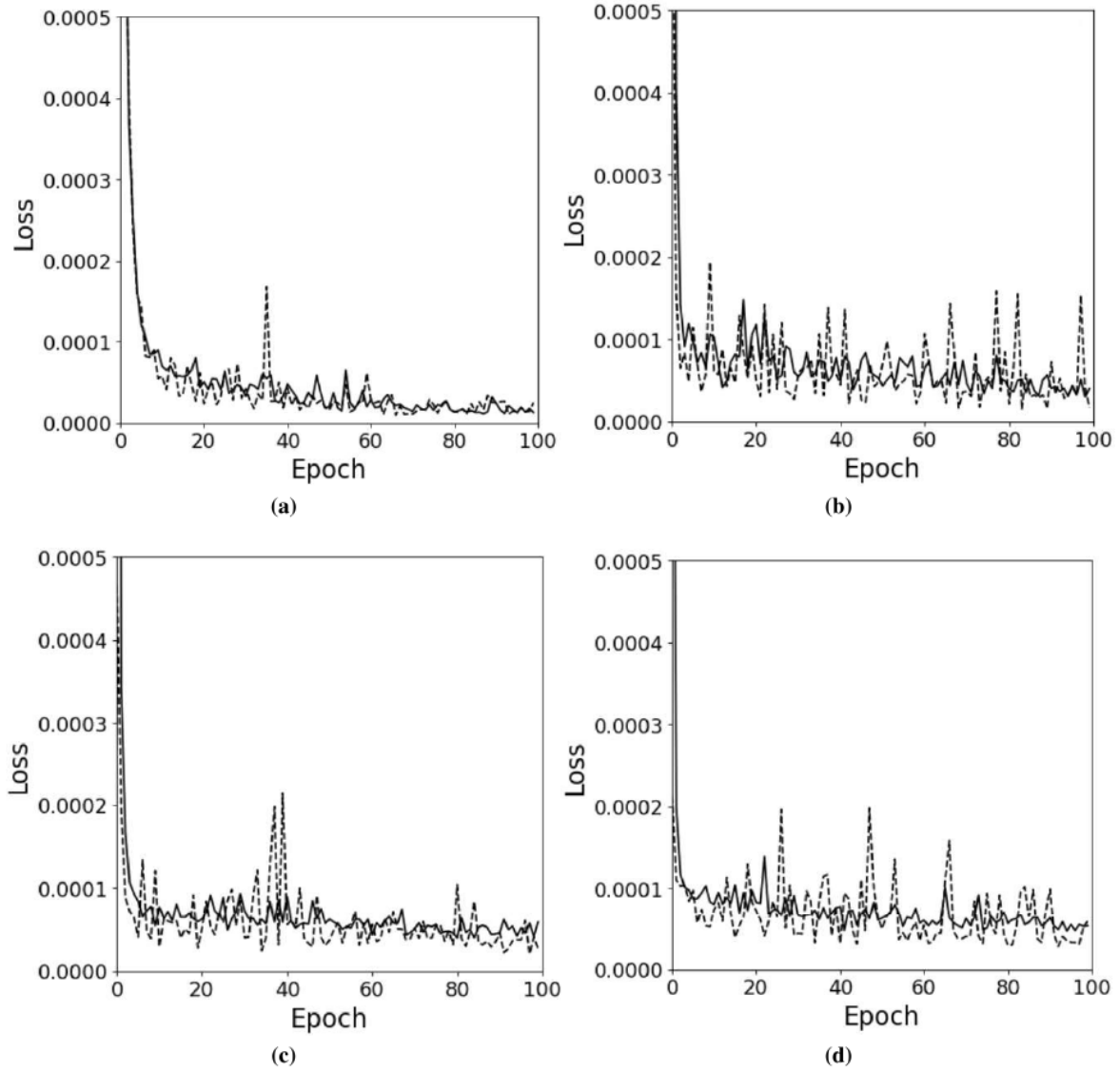


Fig. 6. Loss convergence in MLP training. a) case 4 b) case 3 c) case 2 d) case 1. Solid line and dashed line correspond to training and validation losses respectively.

	Case 1		Case 2		Case 3		Case 4	
Models	**Train R^2**	**Test R^2**	**Train R^2**	**Test R^2**	**Train R^2**	**Test R^2**	**Train R^2**	**Test R^2**
MLP	0.967	0.978	0.978	0.984	0.921	0.891	0.992	0.892
RF	0.971	0.978	0.949	0.966	0.962	0.988	0.984	0.885
GBDT	0.982	0.871	0.968	0.958	0.951	0.955	0.982	0.871

Table 3. R^2 of ϕ^*_{uv} predictions by training-prediction cases

hyper-parameters obtained from manual search, the R^2 scores obtained from different ML models are presented in table 3. For case 1, the ML models were trained with data for friction Reynolds numbers 550, 1000, 2000 and tested for 5200. The maximum value of train R^2 is obtained from the results of case 4 with MLP. The loss convergence history of MLP for the four different cases are shown in figure.6. In most of the training cases, the R^2 magnitude is greater than 0.95, which signifies a good correlation between the actual and predicted value of pressure strain term (ϕ) for training samples (A R^2 value of 0 and 1 correspond to no-correlation and perfect correlation between the actual and predicted values of the pressure strain term). In figure.7, the results of DNS simulation samples versus data driven model predictions of the pressure strain correlation are presented. There is very good match between the DNS results and DNS model predictions. There are very few points in the data-set with high values of the pressure strain correlation, this signifies the fact that, the relative inaccuracy of the model predictions for high values of the pressure strain correlation is due to very few such training data samples(figure5).

9. Testing of the trained ML models

The trained models were tested against testing data for all the four cases as discussed in table.1. The results of testing of the ML models are presented in 8. From all four figures of figure 8 it is clear that, the GBDT predictions are better in comparison to the predictions of MLP and RF. The neural networks failed to predict the pressure strain correlation for $y+$ values greater than 25 for case 4. The testing results are presented in table 3. In most of the testing cases the R^2 value was found to be greater than 0.95, but in few cases like case 4, the R^2 value falls below 0.9 for the testing case. This is because of over-fitting of the models with the training data. The over-fitting problem is common in random forests. The advantage of GBDT over MLP is that the former needs less number of weights in developing the correlation between the inputs and the output. Since the ultimate aim of turbulence modeler is to apply the ML based based into CFD solver, models with less number of weights are always preferred. Since model with large number of weights, may end with divergence in CFD solver. So we have tried to enhance the predictive capability of the GBDT with Bayesian optimization of hyper-parameters as discussed in section 7.. The results of predicted values ϕ with the optimized GBDT is presented in 9. The solid line in the figure correspond to the predictions of the optimized GBDT and the dashed line correspond the prediction of the GBDT with the hyper-parameters obtained from manual tuning. from the figure it is clear that, the predictions of the optimized GBDT matches well with the DNS results. Finally, we have tested the GBDT model against a fully unknown flow e.g. Couette flow in channels. The model predictions are presented in figure.10. It is noticed that the GBDT predictions are matching well the DNS results of (Lee and Moser 2018).

10. Conclusions

In this article, we discuss rationale for the application of machine learning with high-fidelity turbulence data to develop models at the level of Reynolds stress transport modeling. Then, we ascertain the efficacy of different machine learning algorithms at creating surrogate models for the pressure strain correlation. We have modeled the pressure strain correlation of turbulence using three different machine learning approaches, those are Random forests, gradient boosted tress and artificial neural networks. The input features to the ML models were chosen from the traditional modeling basis of the pressure strain correlation, those are mean strain, turbulence kinetic energy, Reynolds stress anisotropy and turbulence dissipation. The ML models were trained and tested for DNS data of turbulent channel flow at different friction Reynolds numbers. The optimal values of ML model hyper-parameters were optimized using manual search and Bayesian optimization approaches. The feature importances of different input features of the random forest were obtained using mean decrease

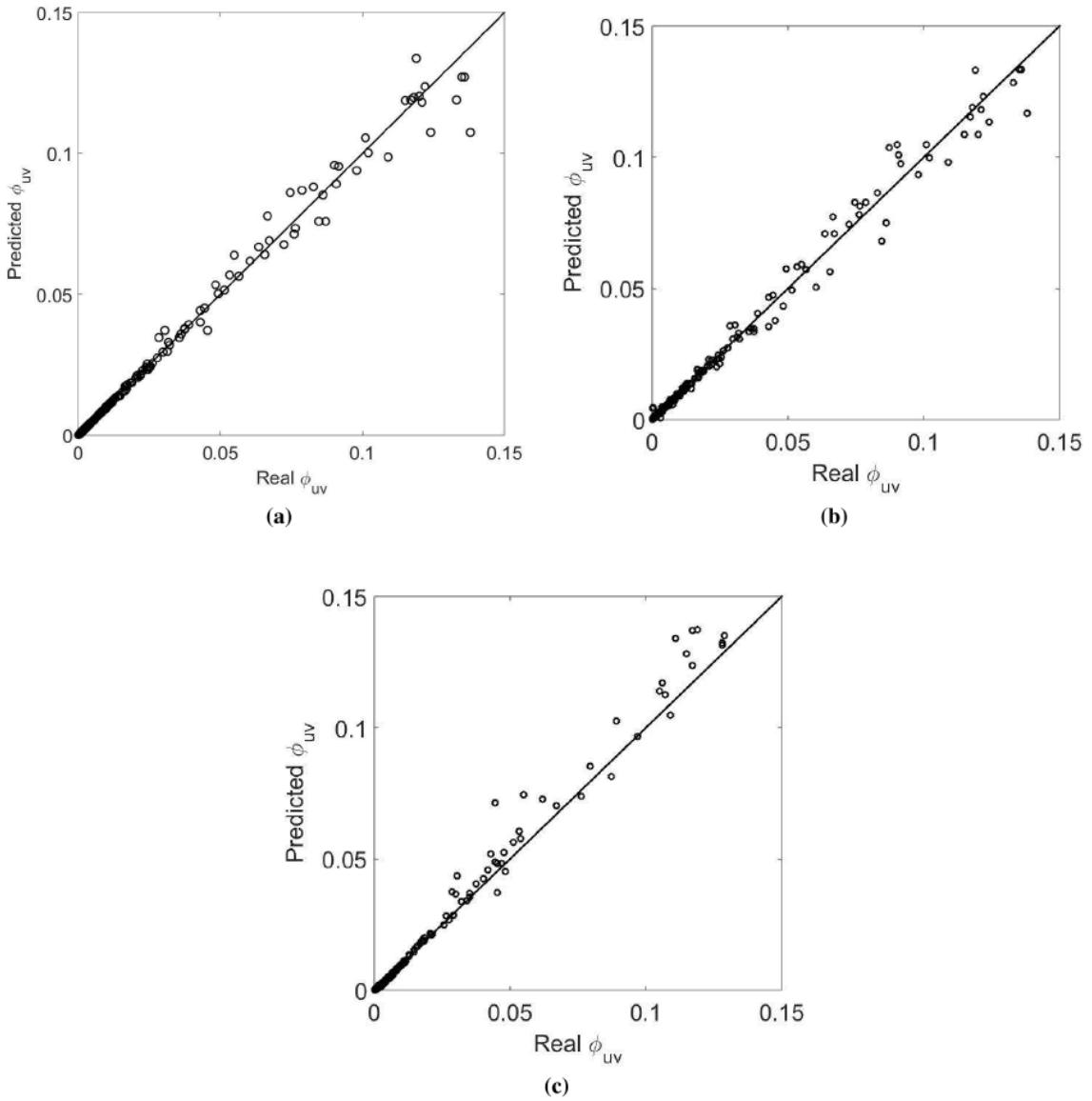


Fig. 7. DNS simulation samples versus data driven model predictions of the pressure strain correlation for the ML models training for case 4, a) Random forest b) Gradient boosted trees c) Neural network.

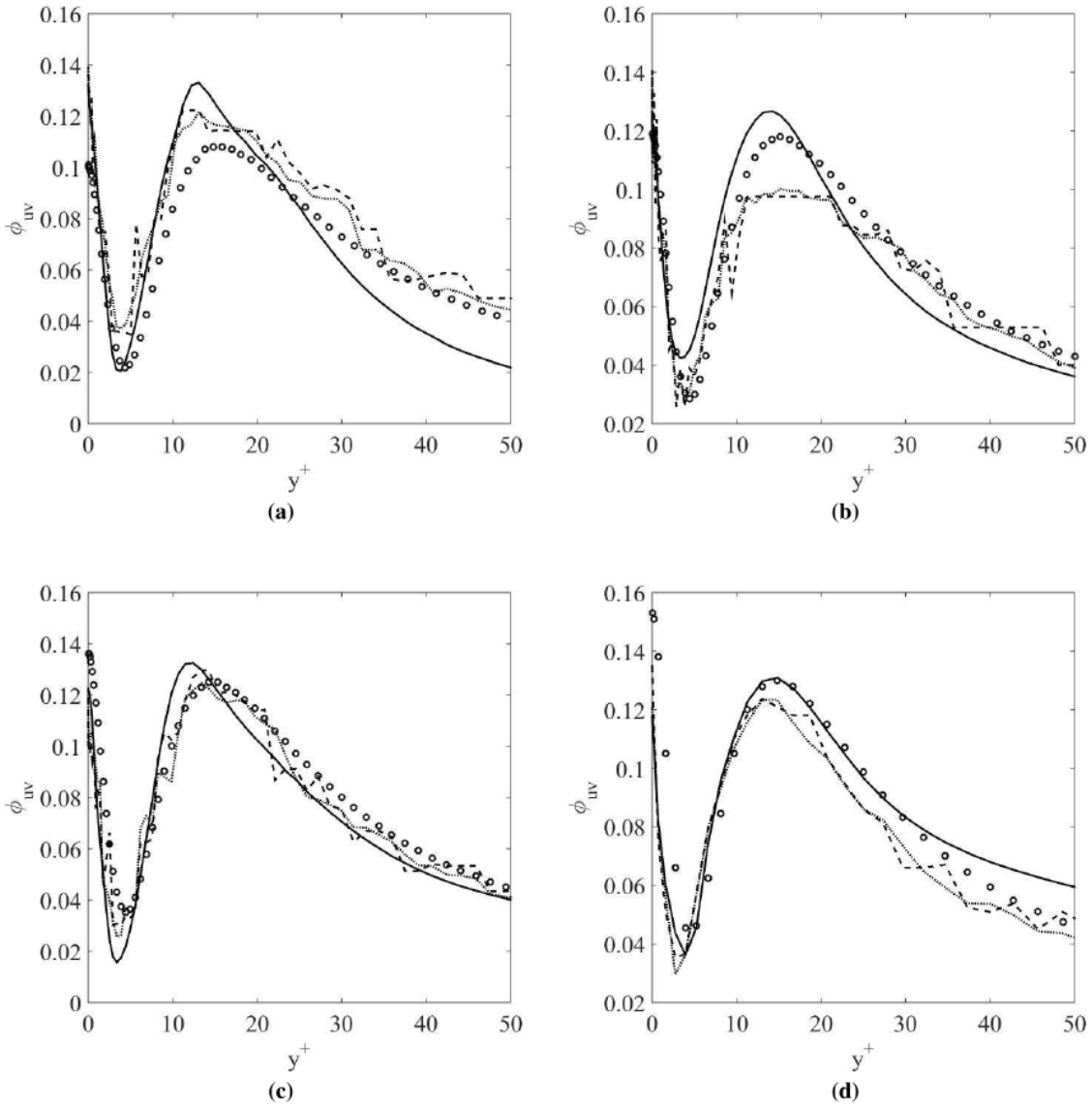
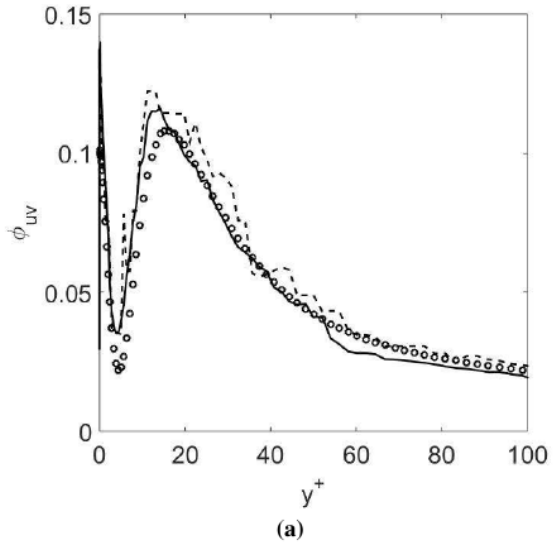
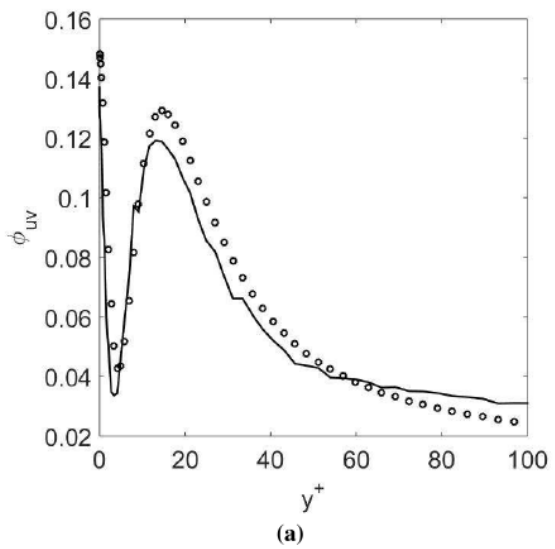


Fig. 8. Prediction of the pressure strain correlation using different machine learning algorithms.Solid line, dashed line and dotted line represent predictions of neural network, gradient boosted trees and random forests respectively. a) case 4, b) case 3, c) case 2, d) case 1.



(a)

Fig. 9. Prediction of the pressure strain correlation for case 4, using the optimized GBDT. The hyper-parameters of the GBDT were tuned using Bayesian optimization.



(a)

Fig. 10. Prediction of the pressure strain correlation for turbulent Couette flow at $Re_\lambda = 550$.

in impurity method. It was noticed that the mean strain has larger correlation with the pressure strain term. The ML models developed by using data-driven approaches can be utilized in computational fluid dynamics solvers for improved flow predictions in turbulent channel flows. in future course of work, for achieving universality, large amount of DNS or experimental data can be fed to the ML models during training and development.

REFERENCES

Bardino, J., J. H. Ferziger, and W. C. Reynolds (1983). Improved turbulence models based on large eddy simulation of homogeneous, incompressible turbulent flows. *Report Stanford Univ.*.

Beck, A., D. Flad, and C.-D. Munz (2019). Deep neural networks for data-driven les closure models. *Journal of Computational Physics 398*, 108910.

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Duraisamy, K., G. Iaccarino, and H. Xiao (2019). Turbulence modeling in the age of data. *Annual Review of Fluid Mechanics 51*, 357–377.

Edeling, W. N., G. Iaccarino, and P. Cinnella (2018). Data-free and data-driven rans predictions with quantified uncertainty. *Flow, Turbulence and Combustion 100*(3), 593–616.

Fang, R., D. Sondak, P. Protopapas, and S. Succi (2020). Neural network models for the anisotropic reynolds stress tensor in turbulent channel flow. *Journal of Turbulence 21*(9-10), 525–543.

Gatski, T. B. and C. G. Speziale (1993). On explicit algebraic stress models for complex turbulent flows. *Journal of fluid Mechanics 254*, 59–78.

Huijing, J. P., R. P. Dwight, and M. Schmelzer (2021). Data-driven rans closures for three-dimensional flows around bluff bodies. *Computers & Fluids*, 104997.

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

Lee, M. and R. D. Moser (2015). Direct numerical simulation of turbulent channel flow up to $re_tau = 5200$. *Journal of Fluid Mechanics 774*, 395–415.

Lee, M. and R. D. Moser (2018). Extreme-scale motions in turbulent plane couette flows. *Journal of Fluid Mechanics 842*, 128–145.

Ling, J., A. Kurzawski, and J. Templeton (2016). Reynolds averaged turbulence modelling using deep neural networks with embedded invariance. *Journal of Fluid Mechanics*, 155–166.

Luan, Y. and R. P. Dwight (2020). Influence of turbulence anisotropy on rans predictions of wind-turbine wakes. In *Journal of Physics: Conference Series*, Volume 1618, pp. 062059. IOP Publishing.

Maulik, R., O. San, A. Rasheed, and P. Vedula (2018). Data-driven deconvolution for large eddy simulations of kraichnan turbulence. *Physics of Fluids 30*(12), 125109.

Mishra, A. A., K. Duraisamy, and G. Iaccarino (2019). Estimating uncertainty in homogeneous turbulence evolution due to coarse-graining. *Physics of Fluids 31*(2), 025106.

Mishra, A. A. and S. Girimaji (2019). Linear analysis of non-local physics in homogeneous turbulent flows. *Physics of Fluids 31*(3), 035102.

Mishra, A. A. and S. S. Girimaji (2010). Pressure–strain correlation modeling: towards achieving consistency with rapid distortion theory. *Flow, turbulence and combustion 85*(3-4), 593–619.

Mishra, A. A. and S. S. Girimaji (2013). Intercomponent energy transfer in incompressible homogeneous turbulence: multi-point physics and amenability to one-point closures. *Journal of Fluid Mechanics 731*, 639–681.

Mishra, A. A. and S. S. Girimaji (2014). On the realizability of pressure-strain closures. *Journal of fluid mechanics 755*, 535.

Mishra, A. A. and S. S. Girimaji (2015). Hydrodynamic stability of three-dimensional homogeneous flow topologies. *Physical Review E 92*(5), 053001.

Mishra, A. A. and S. S. Girimaji (2017). Toward approximating non-local dynamics in single-point pressure–strain correlation closures. *Journal of Fluid Mechanics 811*, 168–188.

Panda, J. (2019). A review of pressure strain correlation modeling for reynolds stress models. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 0954406219893397.

Panda, J. (2020). A reliable pressure strain correlation model for complex turbulent flows. *J Appl Fluid Mech 13*, 1167–1178.

Panda, J., K. Sasmal, S. Maity, and H. Warrior (2020). A simple nonlinear eddy viscosity model for geophysical turbulent flows. *J. Appl. Fluid Mech. 13*, 1167–1178.

Panda, J. and H. Warrior (2018). A representation theory-based model for the rapid pressure strain correlation of turbulence. *Journal of Fluids Engineering 140*(8).

Panda, J., H. Warrior, S. Maity, A. Mitra, and K. Sasmal (2017). An improved model including length scale anisotropy for the pressure strain correlation of turbulence. *ASME Journal of Fluids Engineering 139*(4), 044503.

Parashar, N., B. Srinivasan, and S. S. Sinha (2020). Modeling the pressure-hessian tensor using deep neural networks. *Physical Review Fluids 5*(11), 114604.

Parish, E. J. and K. Duraisamy (2016). A paradigm for data-driven predictive modeling using field inversion and machine learning. *Journal of Computational Physics 305*, 758–774.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Pope, S. (1975). A more general effective-viscosity hypothesis. *Journal of Fluid Mechanics 72*(2), 331–340.

Pope, S. (2000). *Turbulent Flows*. New York: Cambridge University Press.

Pope, S. B. (2001). Turbulent flows.

Rotta, J. (1951). Statistische theorie nichthomogener turbulenz. *Z. Phys. 129*, 547–572.

Sarkar, S. and C. G. Speziale (1990). A simple nonlinear model for the return to isotropy in turbulence. *Physics of Fluids A: Fluid Dynamics 2*(1), 84–93.

Schmelzer, M., R. P. Dwight, and P. Cinnella (2020). Discovery of algebraic reynolds-stress models using sparse symbolic regression. *Flow, Turbulence and Combustion 104*(2), 579–603.

Singh, A. P., S. Medida, and K. Duraisamy (2017). Machine-learning-augmented predictive modeling of turbulent separated flows over airfoils. *AIAA journal 55*(7), 2215–2227.

Speziale, C. G., S. Sarkar, and T. B. Gatski (1991). Modelling the pressure–strain correlation of turbulence: an invariant dynamical systems approach. *Journal of fluid mechanics 227*, 245–272.

Tracey, B. D., K. Duraisamy, and J. J. Alonso (2015). A machine learning strategy to assist turbulence model development. In *53rd AIAA aerospace sciences meeting*, pp. 1287.

Versteeg, H. K. and W. Malalasekera (2007). *An introduction to computational fluid dynamics: the finite volume method*. Pearson education.

Wang, J.-X., J.-L. Wu, and H. Xiao (2017). Physics-informed machine learning approach for reconstructing reynolds stress modeling discrepancies based on dns data. *Physical Review Fluids 2*(3), 034603.

Warrior, H., S. Mathews, S. Maity, and K. Sasmal (2014). An improved model for the return to isotropy of homogeneous turbulence. *ASME Journal of Fluids Engineering 136*(3), 034501.

Weatheritt, J. and R. Sandberg (2016). A novel evolutionary algorithm applied to algebraic modifications of the rans stress–strain relationship. *Journal of Computational Physics 325*, 22–37.

Weatheritt, J. and R. Sandberg (2017). The development of algebraic stress models using a novel evolutionary algorithm. *International Journal of Heat and Fluid Flow 68*, 298–318.

Wu, J., J. Li, X. Qiu, and Y. Liu (2020). A comparative analysis of multi-machine learning algorithms for data-driven rans turbulence modelling. In *Journal of Physics: Conference Series*, Volume 1684, pp. 012043. IOP Publishing.

Wu, J.-L., H. Xiao, and E. Paterson (2018). Physics-informed machine learning approach for augmenting turbulence models: A comprehensive framework. *Physical Review Fluids 3*(7), 074602.

Zhao, Y., H. D. Akolekar, J. Weatheritt, V. Michelassi, and R. D. Sandberg (2020). Rans turbulence model development using cfd-driven machine learning. *Journal of Computational Physics 411*, 109413.

Zhu, L., W. Zhang, J. Kou, and Y. Liu (2019). Machine learning methods for turbulence modeling in subsonic flows around airfoils. *Physics of Fluids 31*(1), 015105.