Proceedings of the ASME 2024
VVUQ Symposium
VVS2024
May 15 – 17, 2024, College Station, TX, USA

VVS2024-132964

CREDIBILITY ASSESSMENT OF MACHINE LEARNING-BASED SURROGATE MODEL PREDICTIONS ON NACA 0012 AIRFOIL FLOW

Jared Kirsch[1,2], William Rider[2], Nima Fathi[1]

[1]Texas A&M University, TX, USA
[2]Sandia National Laboratories, NM, USA

ABSTRACT

The use of surrogate models in computational mechanics is an area of high interest due to the potential for significant savings in computational cost. However, assessment and presentation of evidence for surrogate model credibility has yet to reach a standard form. The present study utilizes a deep neural network as a surrogate for a computational fluid dynamics simulation in order to predict the coefficients of lift and drag on a NACA 0012 airfoil for various Reynolds numbers and angles of attack. Using best practices, the credibility of the underlying simulation predictions and of the surrogate model predictions are analyzed. Conclusions are drawn which should better inform future uses of surrogate models in the context of their credibility.

Keywords: Surrogate modeling, Machine learning, CFD, Credibility, Aerospace, Verification and Validation

NOMENCLATURE

α	angle of attack
Re	Reynolds number
C_l	Coefficient of lift
C_d	Coefficient of drag
U_{val}	Validation uncertainty
U_{num}	Numerical uncertainty
U_{input}	Input uncertainty
U_D	Experimental uncertainty
U_{surr}	Surrogate model uncertainty
GCI	Grid Convergence Index
F_s	Factor of safety
f_i	QoI value on i^{th} grid
r	Refinement factor
p	Order of accuracy
E	Validation comparison error
S	Simulation result
D	Experimental result
δ_{model}	Model form error

1. INTRODUCTION

The use of computational mechanics in engineering is widespread and impactful in many ways, from informing details of physical mechanisms of complex phenomena to high-level design decisions. Computational fluid dynamics (CFD) is no exception. CFD models exist at multiple levels of physics fidelity, the most common of which in many engineering applications is Reynolds-averaged Navier–Stokes (RANS). Despite significant computational cost savings over Direct Numerical Simulation (DNS) and Large Eddy Simulation (LES), RANS can still be quite costly, especially in the context of design optimization, in which relatively large parameter spaces are simulated over. The use of a surrogate model in this context can provide large computational cost reduction without sacrificing significant predictive accuracy. From another point of view, RANS models sometimes exist as the highest-fidelity model that is available for routine aerodynamic analysis, with lower-fidelity models being computationally less expensive, but also less accurate [1, 2]. Slotnick et al. noted in their "CFD Vision 2030 Study" that surrogate models had the potential to make multidisciplinary design optimization with data corresponding to high-fidelity model predictions more feasible [3].

Surrogate models for computational mechanics can come in a variety of forms but some of the most popular are machine learning (ML) models. In the paper "Statistical Modeling: The Two Cultures", Breiman notes that classically, statisticians have sought to use data models that have a high level of interpretability, but not always sufficient accuracy. He notes that interpretability versus accuracy is a false paradigm and that predictive accuracy is the goal of modeling. As an example of what he calls an "A+ predictor", Breiman mentions Random Forrests, which have relatively low interpretability but can have high accuracy [4]. Hemming considered several models including kriging and machine learning (ML) algorithms such as radial basis function, neural networks, random forests, and gradient boosting. He decided on using the latter three in his study due to training costs. In this study, the surrogates were used

to predict several features of the flow field around a 2D hypersonic compression ramp. The problem was broken up into two main parts, the first of which was prediction of the location of the shock wave and boundary layer and the second of which was prediction of quantities of interest in the flow field. Gradient boosting performed well on the first task but was too expensive for the second, which left random forests and neural networks, which had comparable performance. Random forests were faster to train and were generally more accurate [5]. The optimal activation function for the neural networks was ReLu, which was also the case in the study of Zelong et al. who used a two-layer convolutional neural network [6].

One challenge with surrogate modeling is the selection of an appropriate sampling strategy to fully cover the parameter space such that the surrogate model does not predict at points that are far from training points. The most simple but expensive method is known as full factorial sampling, in which every value of each parameter is used. Alternatives include Monte Carlo and Latin Hypercube sampling [7]. Dupuis et al. used what they termed the Local Decomposition Method to classify different regions in a flow field around aircraft, vary the sampling density depending on the nonlinearities present in each region, and predict quantities of interest (QoIs) in the flow field. The surrogate model in this case incorporated proper orthogonal decomposition (POD) and Gaussian process regression (GPR). The method was able to correctly identify specific regions of the flow field for multiple aircraft and accurately predict QoIs including the pressure and friction coefficients [8, 9].

In the CFD community, credibility is often discussed in terms of predictive accuracy (or error) and uncertainty. These concepts are highly applicable in surrogate modeling for computational mechanics applications. In order to obtain a wholistic understanding of ML surrogate model credibility, however, it is useful to incorporate ideas from the broader ML community. The book "Trustworthy Machine Learning", by Kush Varshney is a popular reference, and its breadth of exploration immediately presents itself in a cursory glance at the back cover, which states that "accuracy is not enough when you're developing machine learning systems for consequential application domains" [10]. Safety is defined in terms of the minimization of aleatoric and epistemic uncertainty. An undesired outcome is defined as a "harm" if its cost exceeds some threshold. Varshney goes into great depth in this book, and several additional concepts are relevant, but the above summary hopefully indicates the value of such a resource for broadening understanding of credibility in this field and for providing terminology around which discussions and standards can form.

In the present study, 2D CFD simulations of airflow over the NACA 0012 airfoil are done in Ansys Fluent using the RANS model with the $k-\omega$ shear stress transport (SST) two-equation eddy-viscosity model. The simulations are performed over a range of Reynolds numbers and angles of attack. Solution verification was performed on these simulation results using Roache's GCI approach. A deep neural network (DNN) is used as a surrogate for the turbulence model in order to predict the lift and drag coefficients based on the Reynolds number and angle of attack. A datasheet for this dataset is created. The credibility of the RANS simulation results is assessed using the ASME VVUQ 20-2009 methodology. The credibility of the machine learning model's predictions is then assessed in light of the RANS predictions and experimental data, and the contribution of error and uncertainty to the machine learning model's predictions is presented. This work contains a relatively broad study of surrogate model credibility in the context of a known validation case, highlighting important aspects to overall credibility of surrogate model predictions in engineering applications involving computational mechanics.

2. METHODOLOGY

2.1 NASA Validation Case

The physical system chosen for analysis in this study corresponded to the NASA NACA 0012 turbulence model validation case [11]. This case was created to provide a context to perform a rigorous validation analysis of turbulence models, and is convenient for such purposes due to the thorough description on the NASA site and the experimental data that has been compiled and described there. The 2D geometry is specified and boundary conditions are given. The default set of conditions are for air with standard properties at a bulk flow Mach number of 0.15, a Reynolds number per chord of 6 million, and a reference temperature of 300 K. Experimental data exist for multiple Reynolds numbers and angles of attack from around -15° to 15°. However, the NASA page notes that the Ladson tripped data are most appropriate for comparison with fully turbulent simulations, and this data extends from around -4° to around 20° at roughly every 2° (the experimental values of α are not whole numbers). For the purposes of validation in the present study, Ladson data at Ma = 0.15 and a fixed (tripped flow) transition from -4° to 10° are used [12]. At angles of attack of higher magnitude than 10°, flow nears separation and the trend in the coefficient of lift C_l becomes highly nonlinear. Reynolds numbers of 2 million, 6 million and 8.95 million are simulated and experimental data at these values is used for validation, while data at Re = 4 million is compared against ML model predictions at this value without any simulation informing the surrogate at that point.

2.2 RANS Simulations

The geometry and meshes for the present study were created in Ansys Design Modeler and Mesher, and the simulations were run in Ansys Fluent. All tools were academic version 2022 R2.

2.2.1 Geometry and Mesh

The geometry of the NACA 0012 airfoil was obtained from the NASA validation page and the computational domain was created in Ansys Design Modeler. The base mesh was then created, which had 240,000 quadrilateral elements, a smooth transition inflation from the airfoil surface, and a growth rate of 1.2. The domain extents were large (>10c) to avoid

contamination of the solution from domain extent effects. The overall mesh structure is similar to that given on the NASA validation page and by Eleni et al. [13]. Two meshes were created from this base mesh with uniform refinement ratios of 2 each with respect to the next coarsest mesh. Thus, the base mesh was labeled the 1X mesh and the two additional meshes were labeled the 2X and 4X meshes. In addition, a 0.5X mesh was created and used for numerical uncertainty quantification in the tool StREEQ. The 2X and 4X meshes both had an initial cell height less than y+ for this problem. While the 1X mesh did not, the average percent difference in predictions for both coefficients was less than 5% with respect to the predictions on the 2X mesh, and differences in predictions between the 2X and 4X meshes were of similar magnitude. This indicated sufficient refinement levels for the purpose of the study, which is focused more on a process of credibility assessment than achieving the highest level of accuracy possible. Moreover, computational cost constraints associated with running the simulations affected this decision.

2.2.2 Boundary Conditions, Material Properties, and Models

The boundary conditions of this problem consisted of a velocity inlet, a pressure outlet, and a no-slip wall (on the surface of the airfoil). On the velocity inlet boundary, the velocity magnitude and vector direction were specified, the turbulent intensity was set to 5%, and the turbulent viscosity ratio was set to 10. The velocity magnitude was computed as 52.08 m/s from the Mach number of 0.15 and the properties of ambient air. The angle of attack of the airfoil was adjusted by changing the velocity vector. On the pressure outlet, gauge pressure was specified as zero, backflow turbulent intensity was 5%, and backflow turbulent viscosity ratio was 10. Air was the fluid used in the simulations, and it had a density of 1.177 kg/m^3 and a dynamic viscosity that varied depending on the desired Reynolds number, from 6.85×10^{-6} kg/(m s) to 3.07×10^{-5} kg/(m s). This method was exemplified in a related study [14] and proved to change the Reynolds number in such a way as to accurately align with experimental results in the predicted QoIs, as the incompressible nature of the fluid and the Mach number were preserved. The Reynolds numbers simulated were 2 million, 6 million, and 8.95 million, corresponding to three experimental Reynolds numbers from Ladson's study [12]. Throughout this report, "Reynolds number" and "Reynolds number per chord" are used alternatively, as the chord length was 1 m.

The turbulence model used in the present study was the Reynolds-averaged Navier–Stokes (RANS) model with the standard $k-\omega$ shear stress transport (SST) two-equation eddy-viscosity model as implemented in Ansys Fluent [15, 16]. This choice was informed by a previous study [14] as well as a preliminary calculation that compared predictions of the coefficients of lift C_l and drag C_d for the RANS-SST and RANS-SA (Spallart–Allmaras) models. In this calculation, RANS-SST was found to be more accurate. The coefficients of this model were left at their default values for the present study, and these values can be readily found in the Fluent interface.

2.3 Surrogate Model

The surrogate model implemented in the present study was a deep neural network (DNN) implemented in Python using the Keras API within Tensorflow. This choice of model was made based on the fact that it offered a good combination of simplicity, efficiency, and accuracy for the application. These attributes are true in a relative sense when comparing alternative models as mentioned in the introduction. The optimizer used was Adam [17, 18] and the kernel initializer was he_uniform [19, 20]. The network had an input layer with 28 nodes and two hidden layers, with 128 and 256 nodes, respectively. This architecture was found to be more accurate and efficient than alternatives. Predictive accuracy increased significantly with two hidden layers over one hidden layer (the ability to capture nonlinearity in the QoI trends factored into this), while it stayed roughly the same with three hidden layers. The number of nodes per layer resulted in good predictive accuracy while allowing the model to train relatively quickly (within 20% or 3 seconds to train over 5,000 epochs as compared with a network with 28, 56, and 112 neurons in each layer, respectively). Data from the simulations for angles of attack of $-10° \leq \alpha \leq 10°$ was extracted from the simulation reports and put into CSV files. The sampling plan is shown in Figure 1. This range of α-values was chosen because 1) separated flow significantly affects the coefficient of lift at α near ±15°, 2) not all simulations converged at these α-values, and 3) the range of experimental data in the Ladson report was fairly limited.

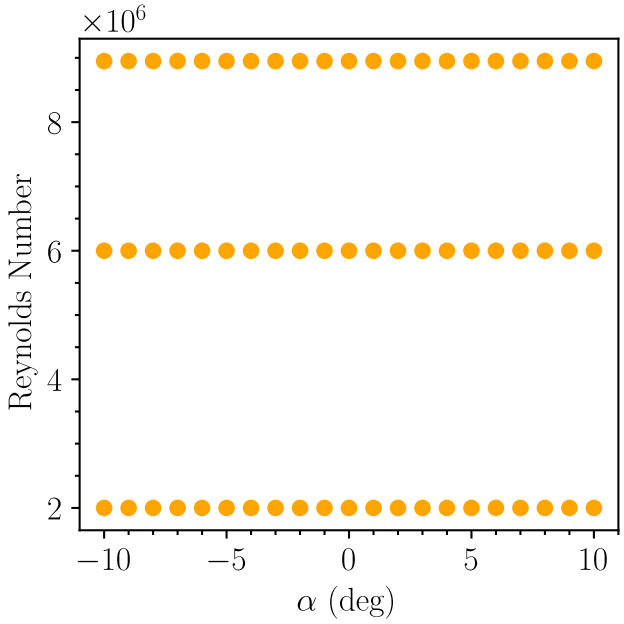

Figure 1. Sampling plan for construction of dataset.

The features of the ML model were the Reynolds number and angle of attack in degrees. Labels of the ML model were the coefficients of lift and drag. In machine learning, features are model inputs and labels are model outputs. Thus, the DNN model took Reynolds number and angle of attack as inputs and

predicted the coefficients of lift and drag. Predictive accuracy with unscaled data was poor, so scaling was pursued. The Reynolds number was divided by 1×10^6 and C_d was multiplied by 100. This improved predictive accuracy significantly. The use of normalization tools within the Tensorflow package did not result in an increase in accuracy relative to this simple scaling. Thus, the simple scaling method was used, and resulting predictions were scaled back before comparison with simulation and experimental data. 70/30, 80/20, and 90/10 training/testing splits were explored, with the 80/20 split producing similar but slightly lower accuracy to the 90/10 split. The 90/10 split was used for the final predictions. The model was trained using 5000 epochs, a number settled upon after testing at several lower numbers. A plot of loss versus number of epochs from this study is shown in Figure 2 using mean absolute error (MAE). 5000 epochs produced low loss and approached asymptotic behavior on the loss curve. 10,000 epochs was also tested and resulted in negligible improvement in predictive accuracy. K-fold cross-validation was used with 5,000 epochs for each of 10 splits. The resulting average root mean square error (RMSE) for both QoIs was on the order of 1×10^{-4}. A plot showing loss using RMSE for a training run is shown in Figure 3. The spikes in loss are likely due to the adaptive learning rate used in the Adam optimizer by default. Additional analysis of this is ongoing. After cross-validation, the model was used to predict the QoIs.

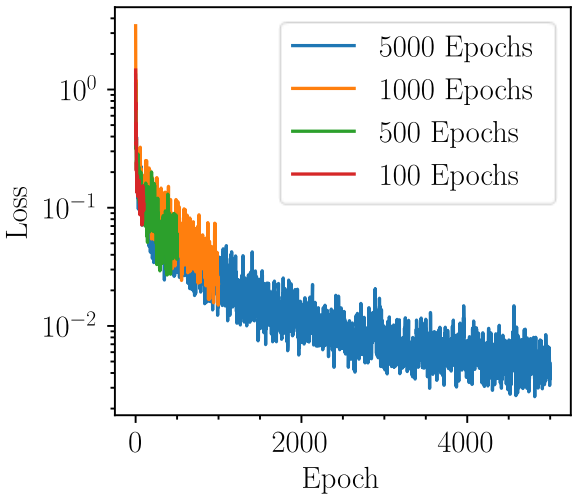

Figure 2. Determining the proper number of training epochs.

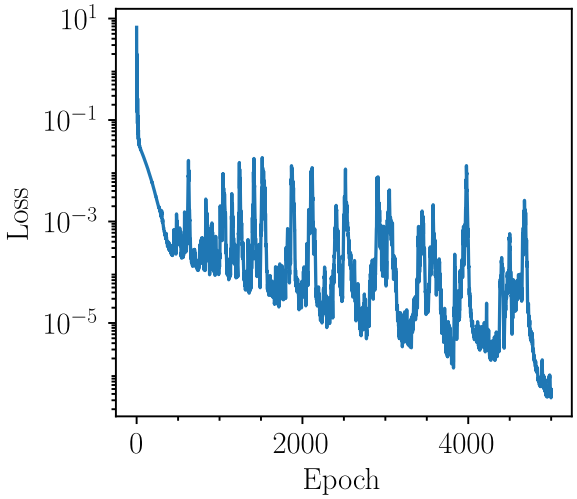

Figure 3. Typical loss curve for training run. Note Logarithmic scale.

2.6 Uncertainty Quantification Theory

The uncertainty quantification methodology in the present report follows that of the ASME VVUQ 20-2009 standard [17]. According to this standard, the validation uncertainty of a simulation, U_{val} , is written as

$$U_{\text{val}} = \sqrt{U_{\text{num}}^2 + U_{\text{input}}^2 + U_D^2} \tag{1}$$

where U_{val} represents uncertainty at the 95% confidence level, which is typically denoted by capitalization. In Equation 1, U_{num} is the numerical uncertainty associated with the simulation prediction, U_{input} is the uncertainty on simulation inputs, and U_D is the experimental uncertainty. In the scope of the present study, numerical and experimental uncertainties are quantified. Input uncertainty is not explored due to time and resource constraints. However, it could be significant and should be included in a comprehensive credibility assessment. In the present study, the numerical uncertainty is quantified using the grid convergence index (GCI). Additionally, for a test case, a method employed by a recently-developed tool named StREEQ at Sandia National Laboratories is used. Documentation on this method is currently limited and will be further investigated in future study. The input uncertainty is not quantified but would be worthy of future investigation. The experimental uncertainty was not provided in detail in the Ladson report [12], but a (likely conservative) estimate is made from a statement in the report.

For validation uncertainty associated with ML model predictions, Equation 1 is modified to the form shown in Equation 2. In this equation, an additional term U_{surr} is added, which represents the surrogate model uncertainty. Below follows a description of how each type of uncertainty was quantified.

$$U_{\text{val}} = \sqrt{U_{\text{num}}^2 + U_{\text{input}}^2 + U_D^2 + U_{\text{surr}}^2} \quad (2)$$

The GCI was computed in the present study using Equation 3, where f_1, f_2, and f_3 are simulation results on the fine, medium, and coarse grids, respectively. The order of accuracy, p, is the result of passing the observed order of accuracy (Equation 4) through a filter. In Equation 4, r is the refinement ratio, which in the present study is 2. The filter applies a ceiling of 2 to the observed order of accuracy and a floor of 0.5. These limits correspond to "reasonable" limits of code order of accuracy for scientific codes. F_s is 1.25 when the difference between the observed and theoretical orders of accuracy is less than 10% and 3.0 when the difference is greater than or equal to 10%. In contrast to the GCI, StREEQ uses four simulation predictions at four corresponding levels of mesh refinement and computes an estimate mesh-converged value with corresponding levels of numerical uncertainty.

$$GCI = F_s \frac{|f_2 - f_1|}{(r^p - 1)} \quad (3)$$

$$p_{\text{obs}} = \frac{\ln\left(\frac{f_3 - f_2}{f_2 - f_1}\right)}{\ln(r)} \quad (4)$$

In order to quantify the experimental uncertainty, the Ladson report was consulted. No specific uncertainties were given, but it was mentioned that a repeatability study was conducted which found that for two points nominally at $\alpha = 0^\circ$ and within 0.01° of each other, the drag coefficient varied by 0.0002 or less and the normal-force coefficient varied by 0.004 or less. The normal-force coefficient variability was applied to the lift coefficient, and these values were taken as experimental uncertainties.

Surrogate model uncertainty was quantified by comparing the mean value of the QoI to the maximum value at each angle of attack. The data for this exercise came from 100 runs of the model, which produced 3-7 values at each angle of attack. The difference between the maximum and mean at each angle of attack was taken to be the surrogate model uncertainty at that angle of attack.

2.6 Validation Theory

Model validation can be described as the process of analysis of the extent that a model represents physical phenomenon for its intended uses [21]. Validation is the anchor to reality for computational predictions, and involves comparison with experimental results. Though historically, comparison of contour plots and other qualitative measures were considered validation, the ASME VVUQ 20-2009 standard places an emphasis on quantitative assessment by defining the validation comparison error as in Equation 5. In this equation, E is the validation comparison error, S is the simulation result, and D is the experimental data.

$$E = S - D \quad (5)$$

The validation comparison error includes possible errors from measured data and simulation predictions. The actual model form error, δ_{model}, is bounded by the validation uncertainty as shown in Equation 6. The validation comparison error and validation uncertainty are shown in Chapter 7.

$$\delta_{\text{model}} \in [E - U_{\text{val}},\ E + U_{\text{val}}] \quad (6)$$

3. RESULTS AND DISCUSSION

In this chapter, the predictions of the CFD model and surrogate model are presented. The numerical uncertainty associated with the CFD simulations is computed and presented. A validation analysis is performed using the ASME VVUQ 20-2009 methodology. Finally, the credibility analysis is discussed from a vantagepoint that seeks to collect key overall takeaways in surrogate model credibility analysis.

3.1 CFD Predictions

The full flow field was predicted using RANS-SST k–ω model, with the QoIs being the coefficient of lift C_l and coefficient of drag C_d. Most of the results that follow are focused on these QoIs. Simulation predictions of both QoIs for $-10^\circ \leq \alpha \leq 10^\circ$ are shown in Figures 4 and 5. Because the change in C_l over α-space is relatively large compared to the difference between curves, the data was transformed. A line was fit to the experimental data and this line was subtracted from each curve. This process was repeated for each Reynolds number using the associated experimental data. The resulting curves are shown in Figure 6. The C_l predictions for $Re = 8.95 \times 10^6$ deviate less from the experimental data than those at other Reynolds numbers. RANS-SST simulation results from the NASA validation study at two angles of attack for $\alpha = 2^\circ$, $Re = 6 \times 10^6$ are also shown in these figures for reference, and agree reasonably well. These results were obtained using NASA's CFL3D code [22].

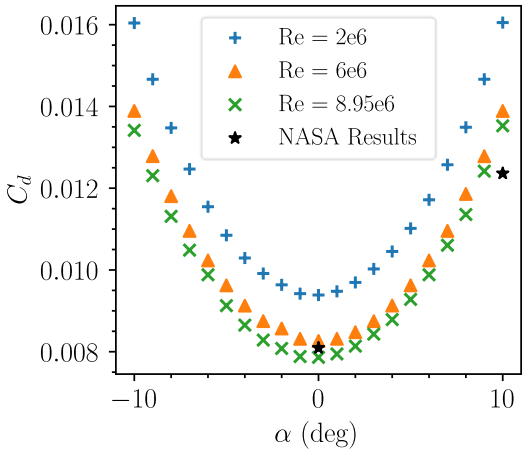

Figure 4. C_d from simulations.

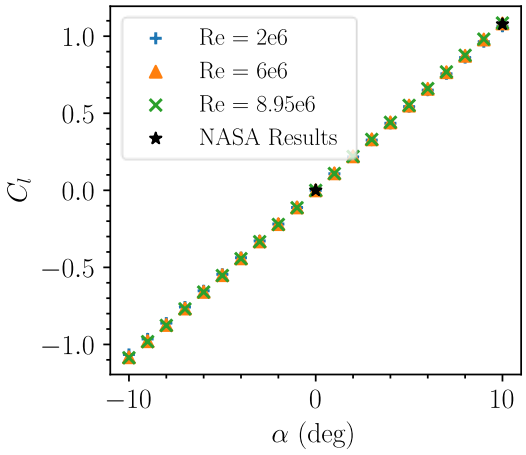

Figure 5. C_l from simulations.

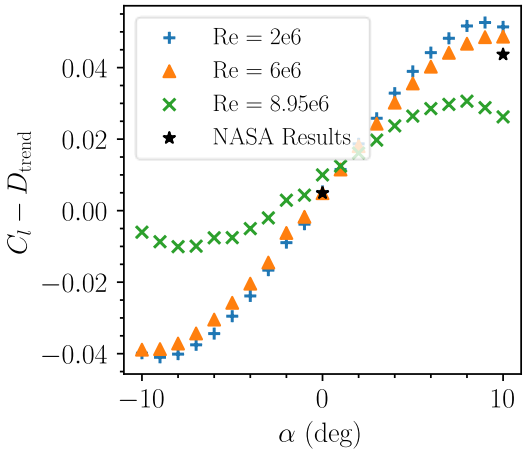

Figure 6. Transformed C_l simulation results.

3.2 ML Model Predictions

A given run of the ML model resulted in a prediction of five values each of C_d and C_l at five angles of attack. A sample result showing the predicted and true values from a run is shown in Figure 7. Note that the C_d-values are scaled. The predictions closely approximate the true values at all points. Though this figure gives a clean visual representation of the predictions of the model, it cannot be used for rigorous credibility assessment. The first step in the credibility assessment process was the use of repeated k-fold cross-validation using scikit-learn's RepeatedKFold method with 10 splits and 3 repeats. The resulting RMSE for C_l was 5.68×10^{-4} and for was 2.16×10^{-4}.

Because each run of the ML model resulted in only five predictions at random angles of attack, the model was run 100 times resulting in 3-7 predictions at each angle of attack. Analysis was done comparing the average and the maximum values at each angle of attack from these runs. This analysis showed that the percent difference between the average and maximum values was nearly always below 10% for both QoIs, with two exceptions. The first was one data point with slightly higher percent difference for C_d. The second was a peak in percent difference due to the low magnitude of the quantity for C_l. Since predictions were consistent from run to run (see Figures 8 – 9), the average value of the QoIs was used at each angle of attack. However, the run-to-run variability was included in the overall validation uncertainty as shown in the next section.

Figures 10 and 11 show the ML model's predictions for all three Reynolds numbers as well as an intermediate Reynolds number (Re= 4×10^6) that simulations were not performed at. Figure 11 shows the linearly transformed C_l trends. The trend at Re = 4×10^6 is smoother than that at the other Reynolds numbers. This reflects the fact that more runs per angle of attack were performed for this Reynolds number, since 100 runs were used but only one Reynolds number was predicted at. In the larger picture, this reflects the fact that the more runs are used, the better the average ML model prediction will converge to the true average value. However, given a quantification of the run-to-run variability and the reflection of that variability in the validation uncertainty, this is a known phenomenon that can be controlled or accounted for in the use of the ML model. Moreover, if the model used significantly more data (e.g., 100 values each of Re and α), training costs for such an exercise would be significant.

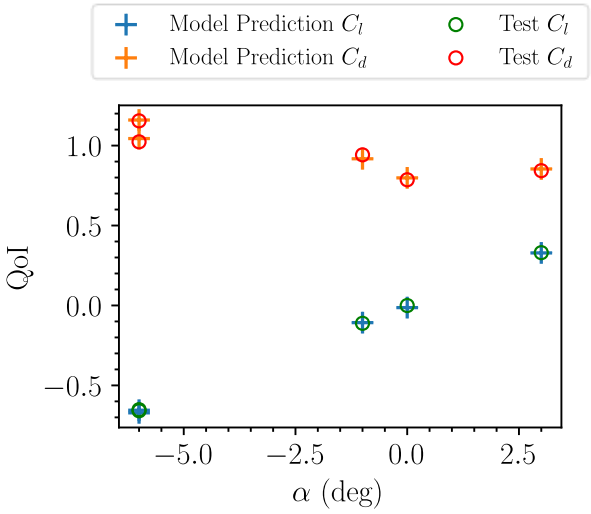

Figure 7. Example of ML predictions from run.

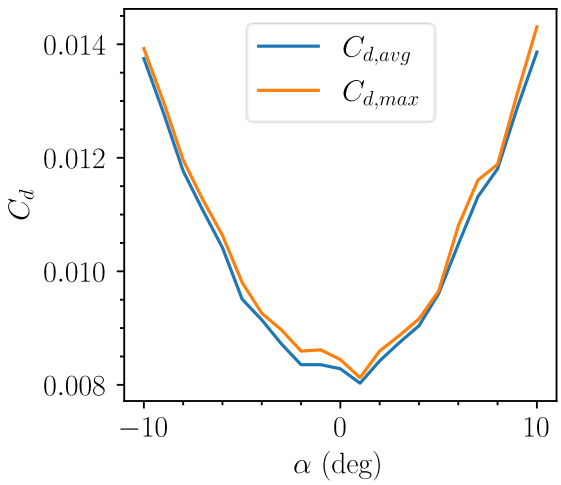

Figure 8. Average and max. values for C_d from 100 runs of ML model at $Re = 6 \times 10^6$.

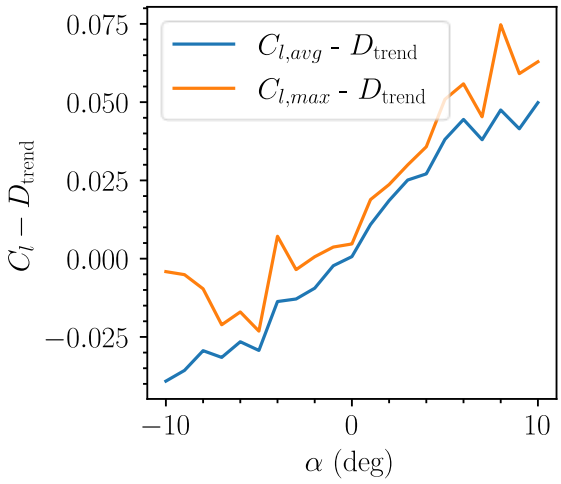

Figure 9. Transformed average and max. values of C_l from 100 runs of ML model at $Re = 6 \times 10^6$.

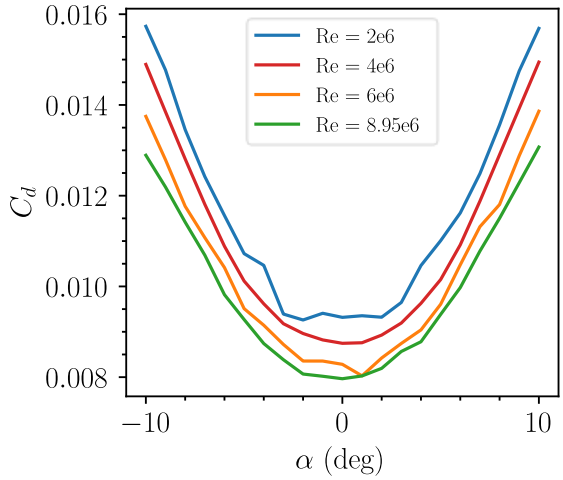

Figure 10. Average ML model-predicted values for C_d for all Reynolds numbers.

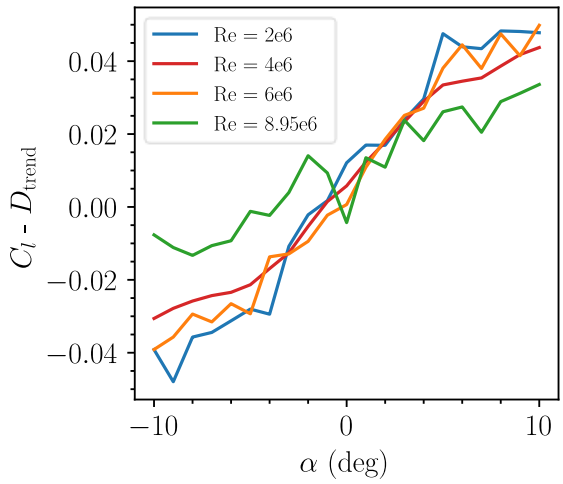

Figure 11. Transformed average ML model-predicted values for C_l for all Reynolds numbers.

3.3 Uncertainty Quantification

Uncertainty quantification in the present study consisted of the quantification of numerical (U_{num}), experimental (U_D), and surrogate (U_{surr}) uncertainties. The methodology used to quantify these uncertainties is discussed in Section 2.6. The numerical uncertainty was quantified using the GCI, and results for each of the QoIs are shown in Figures 12 and 13. The GCI for C_d is generally below 10^{-3} and that for C_l is generally below 10^{-2}. The highest values are generally at the bounds of the α-domain, where the flow is more complex and simulation predictions were more mesh dependent.

The GCI is shown applied to QoI trends as uncertainty bounds in Figure 14. Numerical uncertainty is reasonably large for the Cd predictions at $Re = 2 \times 10^6$ and 8.95×10^6, and smaller

for $Re = 6 \times 10^6$. It also does not appear as a large uncertainty band on the C_l plots, largely due to the magnitude of the QoI.

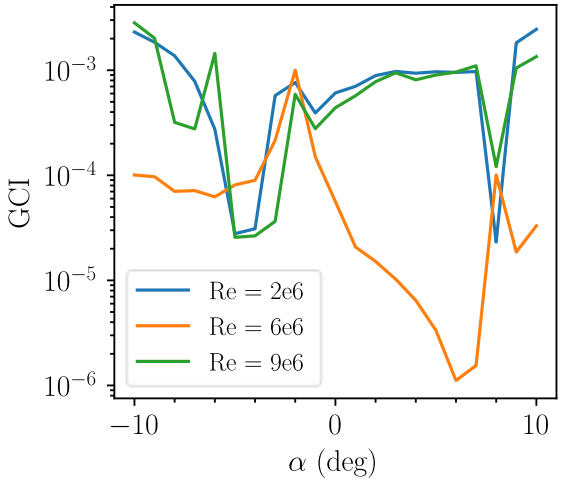

Figure 12. GCI computed for C_d simulation results at $Re = 2$, 6, and 8.95×10^6.

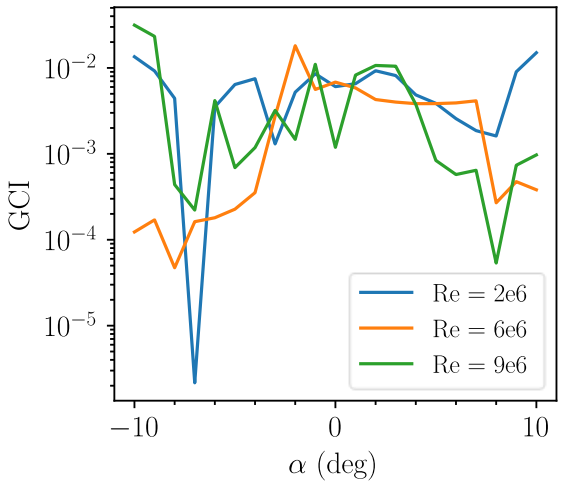

Figure 13. GCI computed for C_l simulation results at $Re = 2$, 6, and 8.95×10^6.

Predictions were also made using the surrogate model at $Re = 4 \times 10^6$. Since simulations were not run at this Reynolds number, numerical uncertainty was calculated as the interpolated GCI (using the GCI at $Re = 2 \times 10^6$ and 6×10^6). The experimental uncertainty was the same as for other Reynolds numbers. Surrogate model uncertainty was computed as for other Reynolds numbers. These calculations resulted in an uncertainty estimate that is reasonable but contains influence from neighboring points in the GCI.

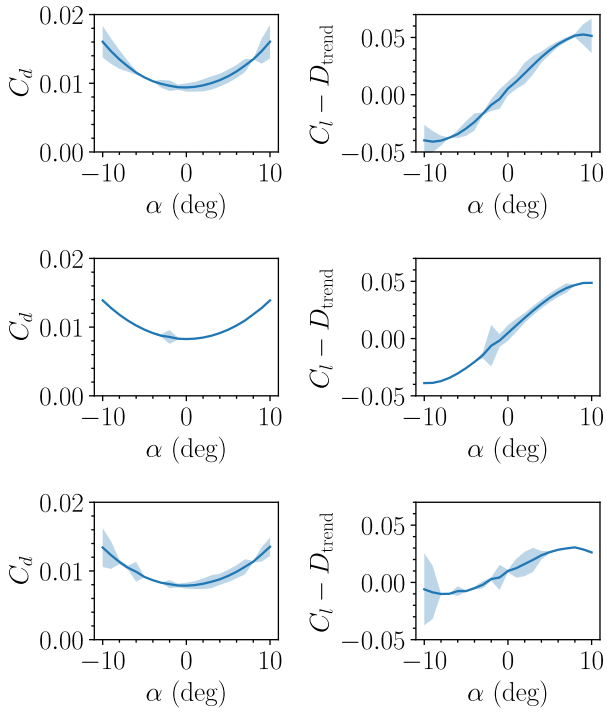

Figure 14. GCI applied to QoI trends at $Re = 2 \times 10^6$ (top), 6×10^6 (middle), and 8.95×10^6 (bottom).

3.4 Validation

Validation comparison error was computed using the methodology described in Section 2.7 for simulation and surrogate model predictions. This error and the corresponding validation uncertainty were scaled by the experimental values and presented as relative error and uncertainty. The results for simulation error and validation uncertainty are shown in Figure 15. Error and validation uncertainty for the surrogate model is shown in Figure 16. Validation comparison error and validation uncertainty are not shown for the simulation results at $Re = 4 \times 10^6$ because simulations were not run at this Reynolds number. In each case, validation comparison error was computed as a difference between the prediction of the model of interest and the corresponding experimental value, before being scaled. The validation uncertainty of the surrogate model is different from that of the simulations by the inclusion of the surrogate model variability as described in Section 2.6. Error levels are generally higher for the surrogate model predictions than for the simulations, but this is not always the case, as the surrogate model deviates from the simulation predictions both above and below. Validation uncertainty is higher for the surrogate model, which makes sense as the surrogate model variability is only additive in its impact on the validation uncertainty. In general, two observations can be made. First, the error is generally distinguishable from uncertainty, which indicates that actual error exists in the model predictions. When uncertainty ex-tends to the horizontal line at zero, it is an indication that the model

predictions may in fact have zero error at that point, and that reduction in uncertainty is the priority. Because error is distinguishable, it is possible that an application of the surrogate model could correct for the expected bias of the model's predictions. The second observation is that despite error being distinguishable from uncertainty, the relative error is reasonable in general. It is below 20% over most of the domain of analysis for both QoIs, with exceptions primarily in the C_l predictions near $\alpha = 0$, where the magnitude of the QoI is small. While the surrogate model was known to perform well in terms of accuracy from the cross validation and comparison of predicted to test values during prediction, it is helpful to compare the validation comparison error of the surrogate model to that of the simulation predictions. This allows one to see how much of the overall error in a surrogate model prediction is inherited from the parent simulations and how much is due to the surrogate model itself. Figure 17 shows the validation comparison error corresponding to the simulations and that of the surrogate model. In general, the surrogate model error follows the simulation error closely. This indicates that most of the validation comparison error and model form error of the surrogate model comes from the parent simulations. The surrogate model has lower error than the numerical model for some angles of attack. This reflects minor deviations in the surrogate model predictions from the numerical model predictions. As percent error, the values for C_l can be significantly lower near zero angle of attack for the surrogate model due to the small magnitude of the QoI there.

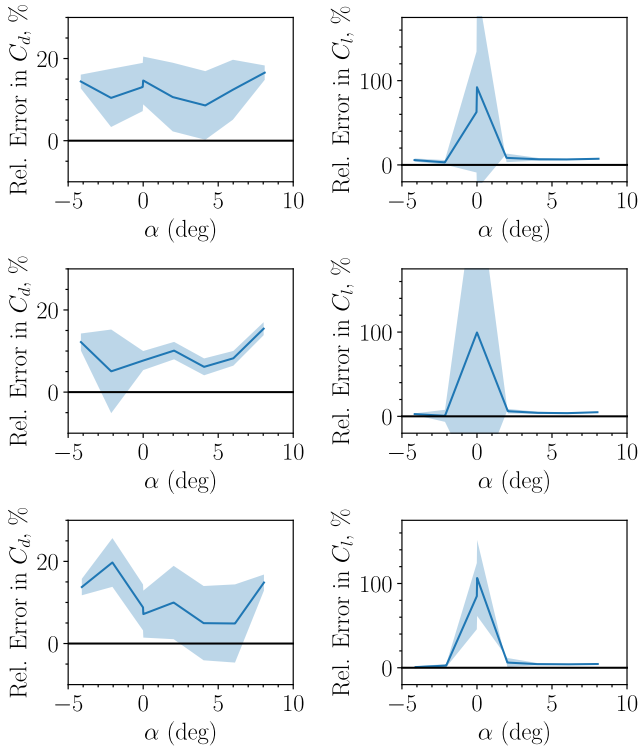

Figure 15. Relative validation comparison error and validation uncertainty of simulations.

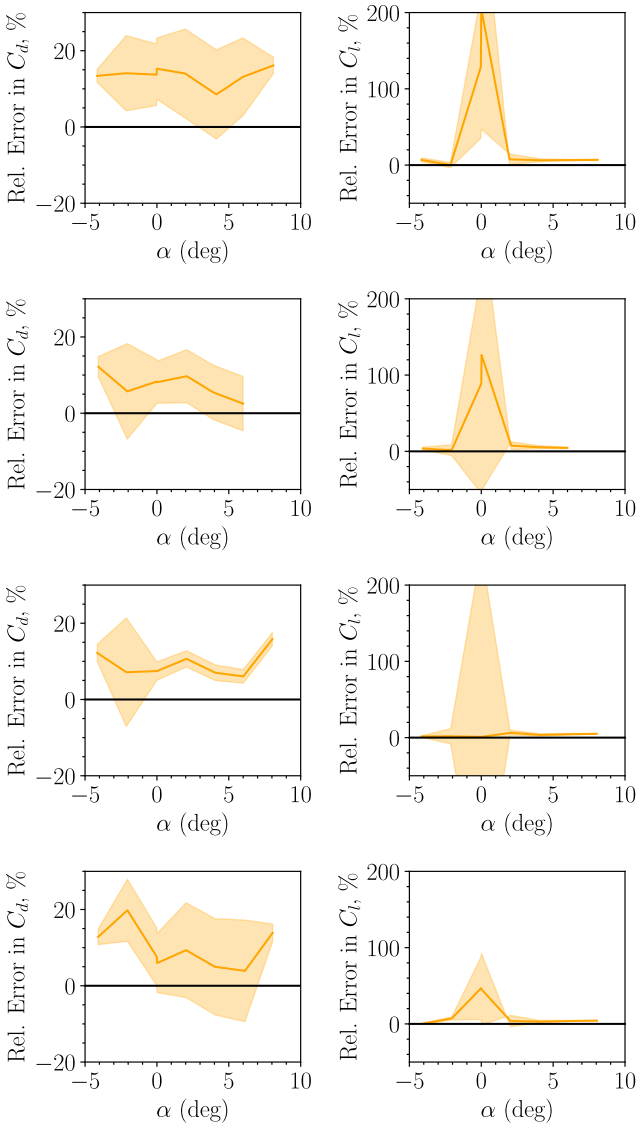

Figure 16. Relative validation comparison error and validation uncertainty of surrogate model.

3.5 The Big Picture

Ultimately, credibility assessment must result in a statement of whether the model of interest and its predictions are credible in the application of interest. If so, the assessment should provide an indication of how credible the model and its predictions are in that application. The above results showed quantitatively the error and uncertainty in the surrogate model's predictions. Figure 18 shows this information for $Re = 2 \times 10^6$, 6×10^6, and 8.95×10^6 tied up into plots containing simulation predictions, surrogate model predictions with accompanying numerical uncertainty, and experimental data with accompanying uncertainty. Figure 19 shows the surrogate model predictions with their accompanying numerical uncertainty as well as the experimental data and uncertainty.

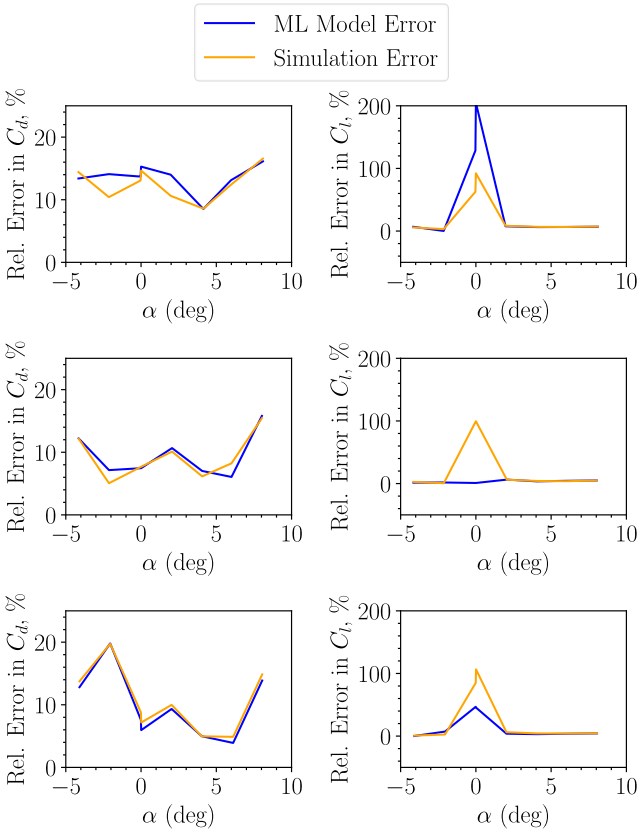

Figure 17. Validation comparison error corresponding to simulation and surrogate model predictions.

The data for C_l is transformed in order to highlight the difference between the simulation predictions, surrogate model predictions, and experimental data. Figures 18 and 19 summarize the performance of the model nicely and show that there is a substantial but arguably manageable error in the surrogate model predictions, stemming primarily from the parent CFD simulations. The estimated uncertainty associated with the surrogate model's predictions capture some of the experimental data but appear less conservative than necessary to capture a majority of it. Uncertainty is relatively high at the edges of the domain, where the model struggles relatively more to predict values accurately. This is a well-known weakness of ML models – difficulty of prediction under extrapolative and near-edge conditions. The predictions at $Re = 4 \times 10^6$ are essentially as accurate as those at Reynolds numbers for which simulation results exist, and this should be the case for any prediction between training points in the parameter space. Any use of the model should take these results into consideration with appropriate correction or conservatism in subsequent design.

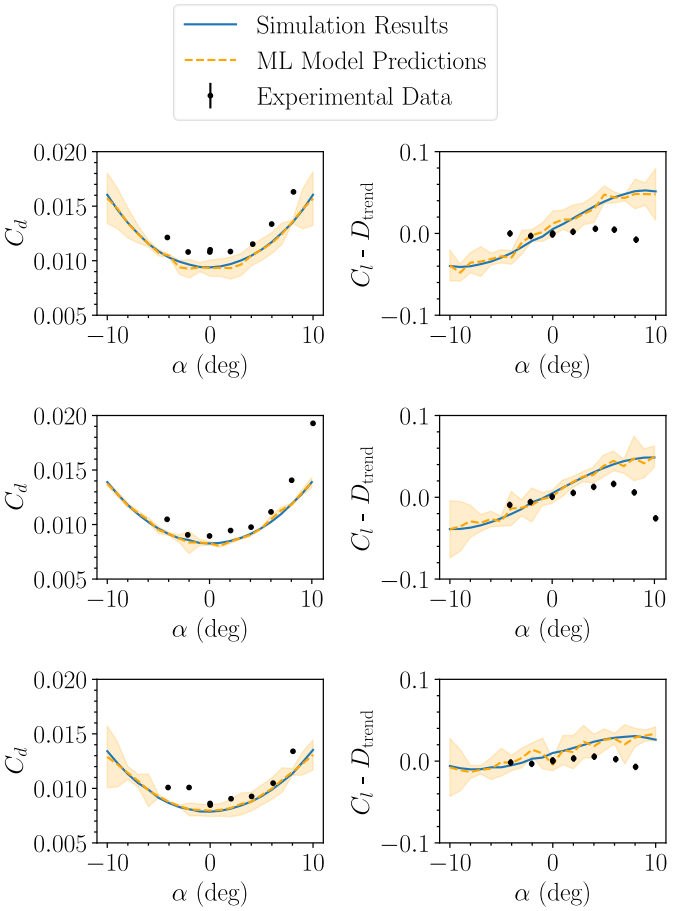

Figure 18. Surrogate model predictions (***transformed*** C_l) with numerical uncertainty, simulation predictions, and experimental data. $Re = 2 \times 10^6$ (top), $Re = 6 \times 10^6$ (middle), $Re = 8.95 \times 10^6$ (bottom).

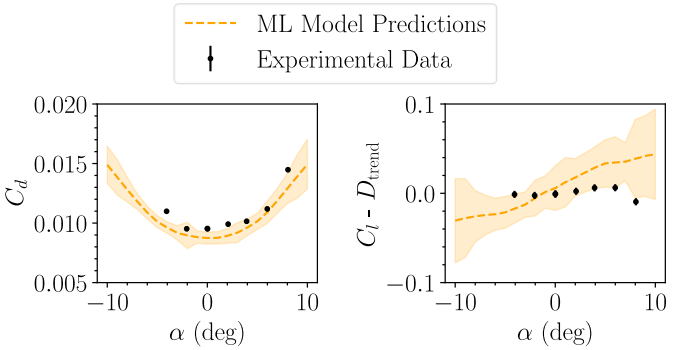

Figure 19. Surrogate model predictions (***transformed*** C_l) with numerical uncertainty for $Re = 4 \times 10^6$.

4. FUTURE WORK

Although the analysis contained in the present study is believed to be thorough and to build on best practices, there are many potentials for further analysis in such a credibility assessment. The accuracy of the surrogate model predictions was

limited by that of the parent CFD simulations. This points to the fact that input uncertainty quantification for these simulations could 1) help to identify input parameters that should be more precisely set for the application (e.g. turbulence levels, $k-\omega$ model coefficients, etc.) and 2) expand the computed validation uncertainty to a more conservative level. StREEQ could be used for numerical uncertainty quantification on all cases, or the comparative assessment between StREEQ and GCI could be expanded to all cases and a decision made. The surrogate model could be updated to correct for simulation bias. Efforts could be made to quantify error and uncertainty on predictions made outside of the training space.

5. CONCLUSION

In the present study, a DNN-based surrogate model was used to predict coefficients of lift and drag for a NACA 0012 airfoil at various angles of attack. Building on best practices for credibility assessment including the PCMM, datasheets for datasets, and the VVUQ approach in ASME V&V 20-2009, the predictive accuracy and uncertainty of the surrogate model was analyzed. Distinguishable but moderate model form error was found to be present in the surrogate model predictions. This could potentially be addressed by bias correction in the surrogate model. Additional uncertainty quantification of the parent CFD model could also be done, focusing on input uncertainty and leading to correction of the most important inputs. Input uncertainty quantification would also increase the estimated validation uncertainty, which did not capture much of the experimental data. The numerical uncertainty estimates computed using the GCI were conservative compared to those of the Sandia National Laboratories UQ tool StREEQ. Further work could analyze numerical uncertainty estimates from GCI and StREEQ in more depth. The credibility of surrogate model predictions between training points was assessed and found to be comparable to that very near to training points. Future work could potentially examine credibility assessment of surrogate model pre-dictions outside of the training space. Overall, the present study showed a start-to-finish process for robust credibility assessment of surrogate model predictions resulting in a statement of model credibility. It is the authors' belief that the present study demonstrates the necessary elements for surrogate model credibility assessment and can be built off of in order to establish more clearly defined standards in this field.

ACKNOWLEDGEMENTS

The authors would like to thank Erin Acquesta for her helpful input.

This work was performed at Texas A&M University through contract with Sandia National Laboratories.

This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan https://www.energy.gov/downloads/doe-public-access-plan.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

REFERENCES

[1] Kroo, I., Willcox, K., March, A., Haas, A., Rajnarayan, D., and Kays, C., 2010, "Multifidelity Analysis and Optimization for Supersonic Design," NASA, Tech. Rep. NASA/CR–2010–216874.

[2] Kirsch, J., Krueger, A., Freno, B., and Lance, B., 2023, "Expanded Verification and Validation Studies of Hypersonic Aerodynamics with Multiple Physics-Fidelity Models," Sandia National Laboratories, Tech. Rep. SAND2023-05305.

[3] Slotnick, J., Khodadoust, A., Alonso, J., Darmofal, D., Gropp, W., Lurie, E., and Mavriplis, D., 2014, "CFD Vision 2030 Study: A Path to Revolutionary Computational Aerosciences," NASA, Tech. Rep. NASA/CR–2014-218178.

[4] Breiman, L., 2001, "Statistical Modeling: The Two Cultures," Statistical Science, 16.

[5] Hemming, N., 2018, "A methodology for rapid hypersonic flow predictions via surrogate modeling with machine learning and deep learning," Master's thesis, Iowa State University, Ames, Iowa.

[6] Yuan, Z., Wang, Y., Qiu, Y., Bai, J., and Chen, G., 2018, "Aerodynamic Coefficient Prediction of Airfoils with Convolutional Neural Network," 2018 Asia-Pacific International Symposium on Aerospace Technology (APISAT 2018), Chengdu, China.

[7] Forrester, A. I. J., Sóbester, A., and Keane, A. J., 2008, "Engineering Design via Surrogate Modeling," Wiley, University of Southampton, UK.

[8] Dupuis, R., Jouhaud, J.-C., and Sagaut, P., 2018, "Aerodynamic Data Predictions for Transonic Flows via a Machine-Learning-based Surrogate Model," Proc. 2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Nagoya, Japan.

[9] Dupuis, R., Jouhaud, J.-C., and Sagaut, P., 2018, "Surrogate Modeling of Aerodynamic Simulations for Multiple Operating Conditions Using Machine Learning," AIAA Journal of Aeronautics and Astronautics, 56.

[10] Varshney, K. R., 2022, Trustworthy Machine Learning, Chappaqua, New York, USA.

[11] NASA Langley Research Center, T. M. R., 2022, "2DN00: 2D NACA 0012 Airfoil Validation Case," https://turbmodels.larc.nasa.gov/naca0012_val.html

[12] Ladson, C. L., 1988, "Effects of Independent Variation of Mach and Reynolds Numbers on the Low-Speed

Aerodynamic Characteristics of the NACA 0012 Airfoil Section," NASA, Tech. Rep. NASA-TM-4074.

[13] Eleni, D. C., Athanasios, T. I., and Dionissios, M. P., 2012, "Evaluation of the turbulence models for the simulation of the flow over a National Advisory Committee for Aeronautics (NACA) 0012 airfoil," Journal of Mechanical Engineering Research, 4.

[14] SimFlow, "NACA 0012 Airfoil - CFD Simulation: SimFlow Validation Case," https://help.sim-flow.com/validation/naca-0012-airfoil

[15] Turbulence Modeling Resource, NASA Langley Research Center, 2023, "The Menter Shear Stress Transport Turbulence Model," https://turbmodels.larc.nasa.gov/sst.html

[16] Menter, F. R., 1994, "Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications," AIAA Journal, 32(8).

[17] Keras 3 API Documentation, Keras, "Adam", https://keras.io/api/optimizers/adam/

[18] Kingma, D. P., and Ba, J., 2015, "Adam: A Method for Stochastic Optimization", digital preprint, https://arxiv.org/abs/1412.6980.

[19] Keras 3 API Documentation, Keras, "Layer Weight Initializers", https://keras.io/api/layers/initializers/

[20] He, K., Zhang, X., Ren, S., Sun, J., Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", digital preprint, https://arxiv.org/abs/1502.01852

[21] ASME, "V&V20-2009: Standard for Verification and Validation in Computational Fluid Dynamics and Heat Transfer," American Society of Mechanical Engineers, 2009.

[22] Turbulence Modeling Resource, NASA Langley Research Center, 2021, "2D NACA 0012 Airfoil Validation Case: SSTm Model Results," https://turbmodels.larc.nasa.gov/naca0012_val_sst.html