Derivation of Integrated Load Distributions from Resampled Computational Data

T.J. Wignall* and Michael W. Lee†
NASA Langley Research Center, Hampton, VA 23666

Principal component analysis (PCA) has been the center of many surrogate models used to characterize fluid flows in recent years. However, little work has been done to characterize the uncertainty in the PCA transform itself and its effect on derived surrogate models. To explore the uncertainty, a typical interpolated surrogate model is constructed for a representative aerodynamic body from computational data. The computational data is then resampled to explore the robustness of the PCA transformation. The variations of the model predictions during this resampling are analyzed to get a measure of confidence in the PCA transformation, which is then applied to the interpolated surrogate model to get uncertainty on integrated force predictions. An initial test case has been explored with promising results.

Nomenclature

CFD	=	Computational Fluid Dynamics
PCA(I)	=	Principal Component Analysis (with Interpolation)
POD	=	Proper Orthogonal Decomposition
STF	=	Simplified Tactical Fighter
A	=	modal coefficient matrix at snapshots
a	=	modal coefficient
B	=	modal coefficient matrix at interpolated points
b	=	bootstrap coefficient residual matrix
C_D	=	drag coefficient
C_L	=	lift coefficient
C_m	=	pitching moment coefficient
M_∞	=	Mach number
M	=	number of modes
N	=	number of snapshots
p	=	parameter domain
s	=	surface mesh domain
U	=	left singular vector matrix
V	=	right singular vector matrix
X	=	snapshot matrix comprised of CFD data
Y	=	data matrix at interpolated parameter points
y	=	arbitrary function
α	=	aerodynamic angle of attack
β	=	continuously defined probability function matrix
λ	=	maximum likelihood estimator
Σ	=	diagonal singular value matrix
σ	=	standard deviation
Υ	=	data distribution matrix at interpolated points
Φ	=	POD mode matrix
Ψ	=	PCA mode matrix
ψ	=	mode defined on entire surface mesh

*Research Aerospace Engineer, Configuration Aerodynamics Branch, thomas.j.wignall@nasa.gov.
†Research Aerospace Engineer, Configuration Aerodynamics Branch, AIAA Member.

$\square_B$ = denotes a matrix derived from a bootstrapped snapshot matrix
$\square_{\#\#}$ = denotes a mode number, e.g., ψ_{24}
$\square_*$ = denotes approximated full-basis equivalent in bootstrapping process
$\square'$ = denotes a matrix with columns removed from its base definition
$\hat{\square}$ = denotes columnar mean of a matrix
$\tilde{\square}$ = denotes a matrix with its columnar mean subtracted

I. Introduction

Quantifying uncertainty in surrogate models is a difficult task when the source data does not provide any measure of confidence as is typical in standard computational fluid dynamics (CFD) solutions. Surrogate models based on the proper orthogonal decomposition (POD) [1] and the principal component analysis (PCA) [2] are often used to approximate CFD data where none has been computed. Some such methods use empirical, orthogonal modes derived from a data set and use those for interpolation at new parameter points [3, 4]. It utilizes system-defining modes, which in this case represent and are derived from fluid flow solution snapshots, that are separate from the parametric information of the problem at hand.

Although uncertainty quantification *via* surrogate models is a rich field [5, 6], including in aerospace and fluid flow applications [7], uncertainty quantification *of* surrogate models for similar applications is less regularly studied. For example, Ref. [8] discusses posterior error estimation of surrogate models, but the process relies on known prior distributions and the authors even acknowledge the difficulty of application to highly nonlinear problems (e.g., configuration-scale fluid simulations).

Based on work in the 1970s and 80s and by using asymptotic theory and assuming random variables, it is possible to calculate an asymptotic standard error for interpolated modal coefficients, which would allow for a confidence interval on predicted low-cost data without the need for a prior uncertainty distribution [2]. However, in practice the narrow assumptions necessary for the standard error to be accurate, for example a normally distributed data set, are rarely satisfied. To help address the shortcomings of these assumptions, bootstrapping (or resampling with substitution) is commonly used. In particular bootstrapping can help find confidence intervals on statistics of interest and has been applied to the transformation matrix [9]. This method has been successfully used to characterize surrogate model confidence intervals in a variety of datasets [10, 11]. Bootstrapping is a technique where the available data are resampled and analysis is repeated to give an idea of the limitations of the available data. The bootstrapping method developed here builds off of work that uses the magnitude of the residuals between bootstrap predictions and full data predictions as a stopping criteria for generating new data for an interpolation surrogate model.[12] The new method expands the bootstrapping to provide a confidence interval on the prediction of integrated loads. A proof of concept of this technique was developed in Ref. [13] to predict distributions on lineload predictions for the Space Launch System, but in-depth analysis was not performed.

In this work, the bootstrapping process is refined to yield confidence intervals on distributed surface pressures of a full aerospace configuration, computed via a modal interpolation surrogate model. Since the analysis focuses on integrated force predictions, only the surface values (and not the volume data) are considered. Also to simplify things, only pressure is used during this analysis, but the technique can easily be extended to the entire flow and any flow variable of interest.

II. Methodology

In the most general sense, this class of surrogate models represents an arbitrary function via a modal sum:

$$y(s,p) = \sum_{i=1}^{M} \psi_i(s) a_i(p) \tag{1}$$

where in this application a snapshot of the presure field on a surface s varies in parameter space p, the modes ψ and their coefficients a respectively only depend on s and p, and a total of M modes are utilized. Different surrogate models will employ different techniques to define the modes ψ and the coefficients a, but this very general structure usually holds. In this work, the modes will be defined via an empirical eigendecomposition based on trusted snapshots of the function of interest x. The coefficients will be found via interpolation, based on the values of those coefficients that are anchored at the trusted snapshots. Bootstrapping is then used to quantify the fidelity of this surrogate model, independent of the fidelity of the underlying snapshots.

A. POD and PCA

Before discussing the bootstrapping method itself, it is worth briefly differentiating the similar modal decomposition techniques known as the proper orthogonal decomposition (POD) and the principal component analysis (PCA) [14]. Both yield an orthogonal basis representation of a snapshot matrix X, which in this case comprises columnar snapshots of stationary flow solutions at different parametric points of interest. POD modes Φ can be defined as the scaled eigenvectors of the autocorrelation matrix [15]

$$XX^T = \Phi\Phi^T \tag{2}$$

whereas PCA modes Ψ can be defined as the scaled eigenvectors of the autocovariance matrix [2]

$$XX^T - \hat{X}\hat{X}^T = \Psi\Psi^T \tag{3}$$

where the vector $\hat{X}$ contains the mean values in each column of X. Note that in both cases the modes, as defined here, span the column space of the original snapshot matrix; and in the context of Eq. 1, the equivalent row-spanning modes are here referred to as *coefficients*. The two modal bases are thus only equivalent when the columns of the snapshot matrix are already mean-centered, which is rarely the case with fluid flow snapshots. POD bases more efficiently project onto the snapshot space, which make them ideal for dynamical systems objectives; PCA bases more efficiently project onto the variance space, which make them ideal for statistical analysis objectives. Since the focus of this effort is on the confidence, rather than the quality, of the resulting transformation, PCA modes were used in this effort.

B. Modal Basis Construction

The singular value decomposition (SVD) and the method of snapshots [16] enables a more computationally efficient modal decomposition (relative to the full-matrix decomposition). The columnar mean-centered snapshot matrix is decomposed into its left and right singular vectors:

$$\tilde{X} \equiv X - \hat{X} = \begin{bmatrix} | & & | \\ \tilde{x}(s,p_1) & \cdots & \tilde{x}(s,p_N) \\ | & & | \end{bmatrix} = U\Sigma V^T \tag{4}$$

where the singular vector matrices U and V are orthonormal by construction and the singular value matrix Σ is diagonal. By construction, the following matrix can be decomposed

$$\tilde{X}^T\tilde{X} = V\Sigma^2 V^T \tag{5}$$

to obtain only the first N right singular vectors, where N is the number of snapshots. In the event that the snapshot matrix is rank-deficient, only the singular vectors with non-zero (positive) singular values are retained. The modes spanning the (usually tall) snapshot matrix kernel are thus not considered, thereby yielding a modal decomposition of cost scaled by snapshot number rather than snapshot size (viz. mesh). The PCA modes Ψ can be found next by using the orthonormal right singular vectors as a transform matrix from snapshot space into modal space.

$$\Psi = \begin{bmatrix} | & & | \\ \psi_1 & \cdots & \psi_N \\ | & & | \end{bmatrix} = U\Sigma = \tilde{X}V \tag{6}$$

Note that these modes are orthogonal but not orthonormal. Each mode is a linear combination of all provided snapshots and is defined, in this application, on the entire surface mesh s.

The matrix representation of Eq. 1, for several (column) snapshots at once comprising the snapshot matrix X, is

$$X = \Psi A^T \tag{7}$$

$$\begin{bmatrix} | & & | \\ x(s,p_1) & \cdots & x(s,p_N) \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \psi_1(s) & \cdots & \psi_M(s) \\ | & & | \end{bmatrix} \begin{bmatrix} a_1(p_1) & -- & a_1(p_N) \\ & \vdots & \\ a_M(p_1) & -- & a_M(p_N) \end{bmatrix}$$

which, given the now derived modal matrix Ψ and its known inverse $\Psi^{-1} = \Sigma^{-1}\Psi^T$, yields the coefficient matrix A at all parameter points discretely provided in the original snapshot matrix.

$$A^T = \Sigma^{-1}\Psi^T X \tag{8}$$

Obtaining a confidence interval on these coefficient quantities is a common effort in PCA applications [2]. The modal decomposition is now complete and defined at all originally observed points in parameter space. The number of modes needed for an adequate representation of the observed system variance is optimal (i.e., small), which is itself the objective of the PCA decomposition [2].

Eq. 7 is thus defined at all N originally observed, trusted points in the parameter space. Now, however, the spatial dependence has been isolated to the (independent) modes ψ and the parametric dependence has been isolated to the (independent) coefficients a. As such, the surrogate model can represent the surface data anywhere in the parameter space – not just where the original snapshots are defined – by interpolating the coefficients to the parameter point of interest and then performing the modal sum as before. Mathematically,

$$Y = \Psi B^T \tag{9}$$

where Y is a snapshot matrix at points other than those in X and B is the interpolated coefficient matrix approximating the data at Y based on the (non-varying) principal components contained in Ψ.

The critical assumption of these kinds of surrogate models is that the available snapshots X are representative of intermediate physics observable in Y; namely, the principal components of the spatially complicated function represent not only the original snapshots but also the system behavior at other points. Such an assumption is common in the literature and has shown to be acceptable for even nonlinear fluid flows [17, 18]. Bootsrapping quantifies the validity of this assumption by highlighting how the modal representation changes when random snapshots are withheld from the original decomposition.

C. Bootstrapping

As discussed in the introduction, it is difficult in practice to compute the error distribution of a discrete, interpolated coefficient set B. Bootstrapping approximates such a quantity by using resampling to assign a measure of accuracy to sample estimates. In this application, bootstrapping is performed by first computing modes and coefficients for a snapshot matrix with randomly dropped columns:

$$\Psi_B,\ A_B \leftarrow X' \tag{10}$$

via Eqs. 5, 6, and 8 where there are fewer than N columns in X'. These bootstrapped coefficients A_B are then interpolated to B_B to reside at the parameter points contained in Y.

$$Y_B = \Psi_B B_B^T \tag{11}$$

There are now two representations of the interpolated data: Y, derived from the full-snapshot basis Ψ; and Y_B, derived from the reduced-snapshot basis Ψ_B. Y and Y_B have the same matrix dimensions (surface mesh data in each column and parameter points in each row, in this application), but will, as bootstrapping will quantify, differ in what values are predicted. This is the point of interest: how reliably the modes based on a reduced snapshot matrix characterize the same interpolated values. Y and Y_B differ due to different modes and different coefficients, but the difference in modes can be moved to only the coefficients via a second representation of Y_B using the full-snapshot modes.

$$Y_B = \Psi B_*^T \tag{12}$$

Though B_* could be computed via simple inversion of the modal matrix as was done for A, this can lead to erroneous behavior. While the analytical modal basis is formally complete, in practice the discrete number of modes used cannot span all possible solutions to all possible interpolated values. The potential disparity in the number of modes retained between Ψ and Ψ_B further muddies these waters. B_* was thus computed via a linear least-squares algorithm, which accepts that the problem may be more or less constrained.

To solve via least-squares for B_*, a residual matrix is defined

$$b \equiv B - B_* \tag{13}$$

where b is a matrix with rows equal to the number of interpolated parameter points and columns equal to the number of retained modes. Subtracting the two full-basis representations of Y (Eqs. 9 and 12) yields the equation solved directly via least-squares:

$$\Psi b^T = Y - Y_B \ . \tag{14}$$

The residual matrix b is how the trusted, full-snapshot modes change in significance when some trusted snapshots are removed from the data stream. Said another way, solving for b will effectively give the variations in the original coefficient matrix, B, necessary to account for the missing data. This difference is a representation of the modal basis sensitivity, and by propagation its reliability in interpolating unknown values.

Computing b many times with different snapshots removed from the system yields a discrete, emergent distribution of values of b for each mode and each parameter point. For uncertainty quantification, this discrete distribution is fit to a one-dimensional, continuouly defined probability density function

$$\beta \leftarrow b \tag{15}$$

that can be sampled arbitrarily. This yields an interpolated quantity of interest matrix Υ (data at each surface mesh location, in this application) akin to Y in Eq. 9 but with distributed values at each entry instead of just nominal values.

$$\Upsilon = \Psi \left[B + \beta\right]^T \tag{16}$$

In this application, these distributions of surface data were then integrated to yield distributions of integrated forces and moments on an aerospace vehicle. Since the modal decomposition effectively splits the spatial and parametric information, Ψ can be spatially integrated ahead of time to allow for efficient calculations of forces and moment distributions.

III. Results

A test case is used to explore this methodology for a simple problem. The geometry used is a simplified tactical fighter (STF) configuration developed under the Air Force Wright Lab [19]. It is a simplified semispan planform used in studies for fighter craft-like configurations [20]. The results were generated using the NASA CFD solver USM3D-ME. USM3D is a finite volume, cell centered, unstructured grid solver developed at NASA Langley [21–23]. The problem is a standard test case provided by the developers of USM3D and the only inputs that were changed from the example setup were angle of attack, α, and Mach, M_∞. The grid is made up of 1,511,114 mixed elements and was run on the NASA Langley K-Cluster using 80 cores and took about 15 minutes each. Simulations used the Spalart-Allmaras turbulence model and were evaluated for 1000 iterations, which dropped the initial residuals by at least 10 orders of magnitude. Figure 1 shows the pressure on the STF at M_∞ =0.75, $\alpha = 3°$.

Two parameters were varied to collect data. The first being α, which was varied between -7° and 7°, and the second being M_∞, which varied from 0.3 to 0.9. Data were generated at $\Delta M_\infty = 0.05$ and at $\Delta\alpha = 1°$. These data were split into three subsets of varying sparsity. The first being the full set of all 195 solutions, which represents a situation where data are plentiful, which is rarely the case in practice. The second being about a third of the data set with $\Delta M_\infty = 0.1$ and α = -7°, -5°, -3°, -1°, 0°, 1°, 3°, 5°, and 7°, which results in 63 solutions used to construct the snapshot matrix. This medium case is representative of situations generating databases where most conditions have direct observations but surrogate modeling is used to fill in some of the gaps. The final subset changes ΔM_∞ to 0.2 and α = -7°, -3°, -1°, 0°, 1°, 3°, and 7° giving 28 input solutions. This coarse case is representative of either initial exploration studies or configurations were each additional data point is very expensive necessitating the use of surrogate models. Model evaluations were at $\Delta M_\infty = 0.025$ and $\Delta\alpha = 0.5°$, which gives 725 parameter points.

The bootstrapping processes was ran three times on each data subset with 300 trials. The first where all modes are retained or in other words 100% of the information is retained. The second and third being where 99.9% and 99% of the information is retained based on the singular values. These thresholds are chosen based on typical usage of surrogate models of this form. If the number of modes to reach that limit was even, it was increased to the next highest mode to prevent cut off of dual modes. Table 1 summarizes the number of solution snapshots used in each case and the number of modes retained in each model. During the bootstrapping process, approximately 20% of the solution snapshots were removed. This is a relatively arbitrary choice and the sensitivity to it should be explored in the future. During the resampling process, the percentage of information retained is a constant. As such the resampled models and the full model may have different number of modes retained. For simplicity during interpolation, the 4 corner points at α = -7, 7 and M_∞ = 0.3 and 0.9 were always retained to prevent extrapolation. To remove parameter points where the resulting value of b is trivial (usually due to querying at a snapshot), any value less than 10^{-14} was omitted from the analysis.

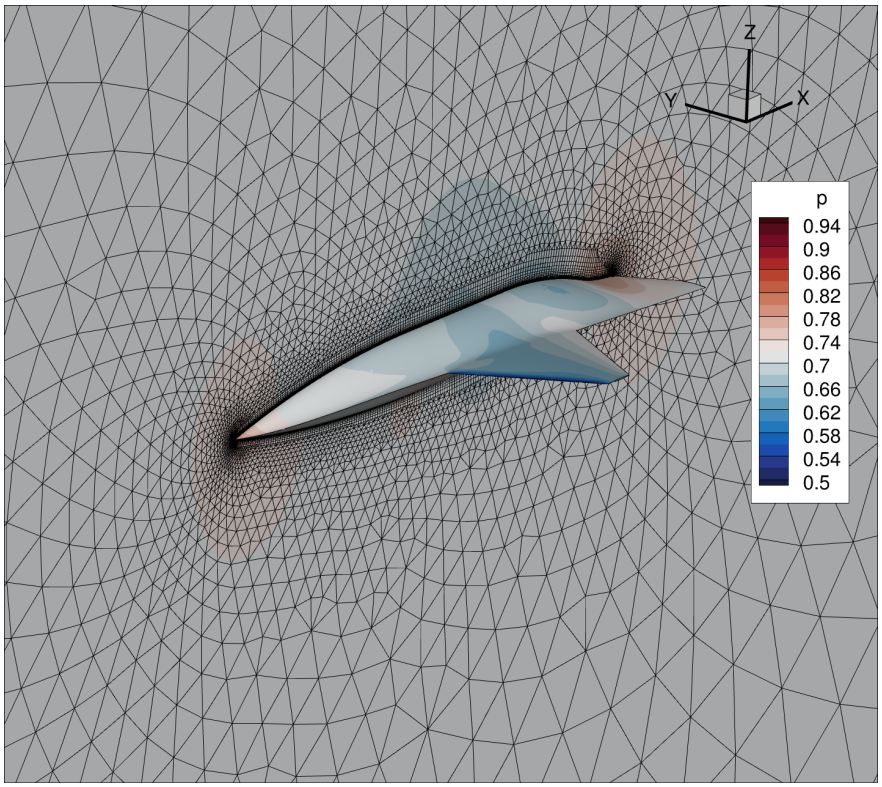


Fig. 1 Surface pressures on the STF at M_∞ =0.75, α = 3°.

Table 1 Table Summarizing the Amount of Snapshots and Modes Used.

Case	% Information	Number of Solutions	Modes Retained
Fine	100	195	195
Fine	99.9	195	41
Fine	99	195	13
Medium	100	63	63
Medium	99.9	63	25
Medium	99	63	11
Coarse	100	28	28
Coarse	99.9	28	15
Coarse	99	28	9

The values of b from the least squares problem (Eq. 14) are plotted in the next couple of figures. These results are pooled across the entire parameter space; however, there is suspected to be a dependence on the input parameters, which is not explored in this work. Figure 2 shows the distribution of b for various mode numbers. At the top, Fig. 2a shows b_1, which is the variations of the coefficient associated with the first mode for the 100% information models for all three cases. On the left is the absolute values of b with the probability distribution on a log scale, while the right shows the natural log of the absolute value. The limits of the natural log plot is chosen to have x_{min} be the cut off value of 10^{-14}. The blue represents the fine case, the orange the medium case and the coarse case is in green. One thing to keep in mind when looking at these plots is that the definition of Ψ includes singular values, which means the values of B are mathematically bounded from -1 to 1 due to the originating eigenvectors being normalized to have a magnitude of 1. The initial observation is this is a distribution characterized by long tails and the distributions explored in this work were chosen to support this observation. As expected, the subsets with less starting information are more effected by the resampling procedure which show more values of b at higher values and a wider spread. The log transformed data on the right, gives a fairly good argument to use a log-normal distribution to fit the data.

As mode number increases, a secondary cluster appears in the log space. The distributions of b_{24} seen in Fig. 2b shows this secondary cluster, which makes the results appear bimodal in log space. This secondary cluster is seen around $\ln|b_{24}| = -30$. It is unclear if this double cluster is a real phenomenon or an artifact of the process. Ideally when doing sensitivity analysis, perturbations are of controlled sizes and done with a continuous distribution of possible values. However, the removal of snapshots is a discrete perturbation that can only be made so small and so there are effectively missing data between the mean of the perturbation and 0. It is likely these tailing values in the secondary cluster to the left are representative of smaller perturbations to the system.

As the mode number continues to increase, the secondary cluster continues to move to the right and grows in prominence. Figure 2c shows the observed scatter for b_{57}. Note that since the coarse dataset had only 28 snapshots, there is not a corresponding 57th mode for that case. When comparing these distributions, it is important to remember that these plots organize b by mode number and depending on the model and case, b_i could have a significantly different contribution to the final predictions.

These trends broadly hold as well for the models where information thresholds are used with some significant differences. As seen in Fig. 3, which examines the medium case at the various information thresholds, the secondary cluster appears at a lower mode number, and as mode number increases it moves closer to the primary cluster much faster and eventually merges. Figure 3a shows the distribution of b_1 for the three information thresholds on the medium case. As can be seen clearly, the 99.9% and the 99% models in orange and green respective already display a secondary cluster. The values of this secondary cluster are already around -20 in the log space, which is a significant difference compared to the distribution seen in 100% information distributions (Fig. 2b) where the secondary cluster starts at edge of the plot. This also means that the models with information cut off have a distribution of b_1 with a higher probability of values closer to 0 compared to the 100% information model. The likely reason behind this is that by removing the higher modes, the system becomes less sensitive to what data are used. This increase in robustness is expected because the information in the highest modes tends to be noisy, which is why they are often dropped.

From here the trends seen in the 100% information models continue. The secondary cluster for the models with information removed continues to the right and eventually merges with the primary cluster in log space. This is seen in Fig. 3b with b_3 where the 99% information model has already merged clusters while the 99.9% has moved closer.

If mode number continues to increase, the distributions of the 99% model starts to reliably return higher values than the 100% model. Figure 3c, shows the value of b_{11} where the 99% model now has most values greater than the 100% values and the two clusters for the 99.9% model have merged. The secondary cluster for the 100% model has not yet appeared at this mode number.

The similarity of the distribution as different subsets of the data were used gives confidence that techniques to fit the distributions can be examined in detailed on one model and case and then tested on the others. The multimodal nature of the log transformed data will not be address explicitly but does prove a challenge that still needs to be overcome. It is implicitly addressed through a truncated log-normal distribution discussed in the next section.

A. Fitting Distributions

Now that the values of b have been calculated the distribution β as defined in Eq. 15 can be fitted. Three distributions are explored in fitting the distributions. The first is a Laplace distribution, the second a log-normal distribution and the third a truncated log-normal distribution. The Laplace distribution is simply a double-sided exponential fit with both positive and negative values. In the plots below, the distribution is labelled exponential because the absolute value is

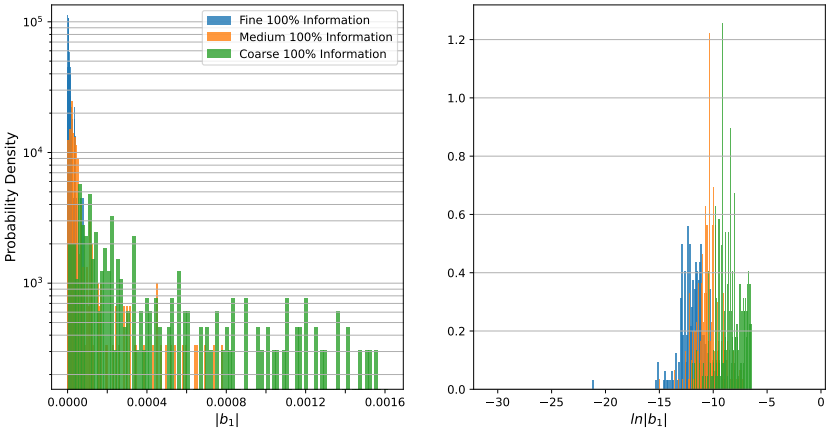
used. Unlike the log-normal distribution it is defined for 0, which by design should be the center of the distributions. It is also relatively easy to sample from, which makes doing Monte Carlo draws easier. The canonical maximum likelihood estimator for the scale factor is used, which is $\lambda_i = 1/|\hat{b_i}|$, where $|\hat{b_i}|$ is the average of the absolute value of the observations of b_i.

The log-normal distribution was explored due to the characteristics of the transformed data. The double bump nature of the data lead to exploration on how best to fit the data and two log-normal distributions were used to explore which is better. One uses all the data and another uses a truncated dataset. The right most cluster is considered more real and so an attempt to focus on that is done by removing any data less than the truncation limit. The data were truncated at $e^{-10}(\approx 4.54x10^{-5})$, which lines up with -10 in the log transformed space. The resulting fit could of course could be improved by tuning this parameter and dynamically picking it; however, that is left for follow up work. In the case where no data were greater than the cut off, the cutoff was reduced to e^{-20}.

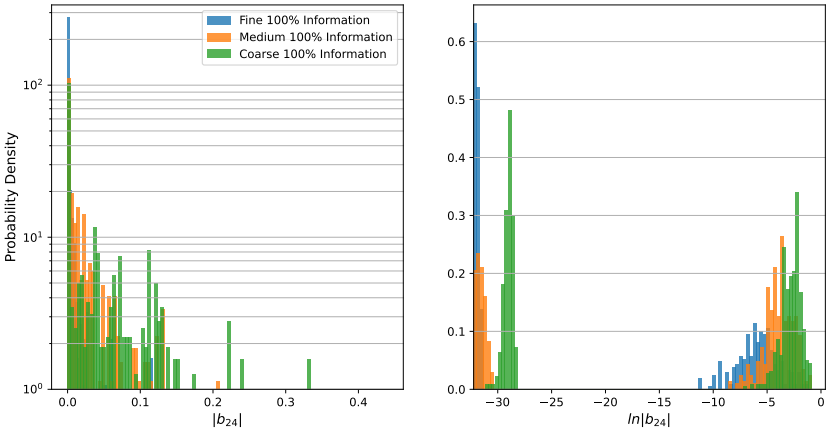
The probability density functions (PDFs) of the potential fits can be seen in Fig. 4 which shows the methods being fitted to the 99% information model for the medium case. The analysis begins with b_1, which is seen in Fig. 4a As shown previously, the trends and behavior of b are consistent across size of input space as well as retained modes and so this makes the results representative. The exponential distribution is seen in orange with the two log-normal distributions in green and red. The red being fitted using the truncated dataset, which can clearly be seen in the log space plot. A first analysis could lead one to believe the exponential in orange is poorly fitting the data; however; it is relatively close to the peak of the right cluster in log space and it is capturing the sharp increase as b goes to 0 in the real space. The red truncated log-normal shows the limitations of a hard cut-off. The peak values are well away from 0 in normal space and falling to almost nothing relatively early. While the green log-normal is not capturing either of the clusters in log space, the resulting distribution in real space aligns fairly well the observations of b

Going to b_2 in Fig. 4b, the benefit of the truncated data set is seen. The red truncated log-normal distribution captures that right cluster well in log space as designed and again the green log-normal distribution captures neither very well. The log-normal distributions while initially promising when looking at the log transformed data, suffers in real space. As the values get close to 0, the PDFs also goes to 0 and being incapable of returning a sample value of 0. However, with the focus being on prediction intervals, the critical performance tends to be behavior of the 90% interval and so this may not be a concern.

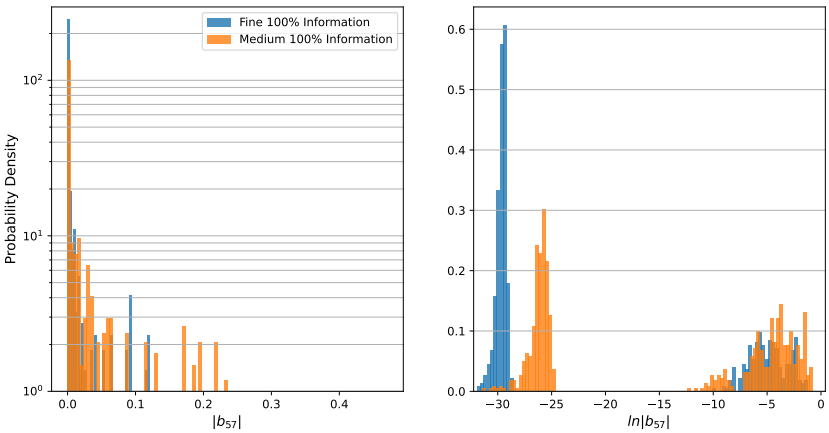
To confirm the findings, the fine data case with 100% of information retained is briefly examined. One difference with the models that retain all the modes, is that the secondary cluster moves to the right slowly as mode number increase. It also spends a long time being on the edge of the data space. This results in the full log-normal fit being relatively flat across log space as is plotted in Fig. 4c, which shows the fits on b_{16}. This causes significant problems while sampling. While the plotted PDF limits match the observed maximum value, it is clear looking at the log space plot that values as high as 5 would be sampled from the log-normal distribution. Converting 5 to real space would give a value of almost 150 which would cause completely inaccurate predictions when used to predict forces and moments. This will be seen in the next section where these distributions will be used to generate predictions of integrated forces and moments.



(a) Histograms of b_1.

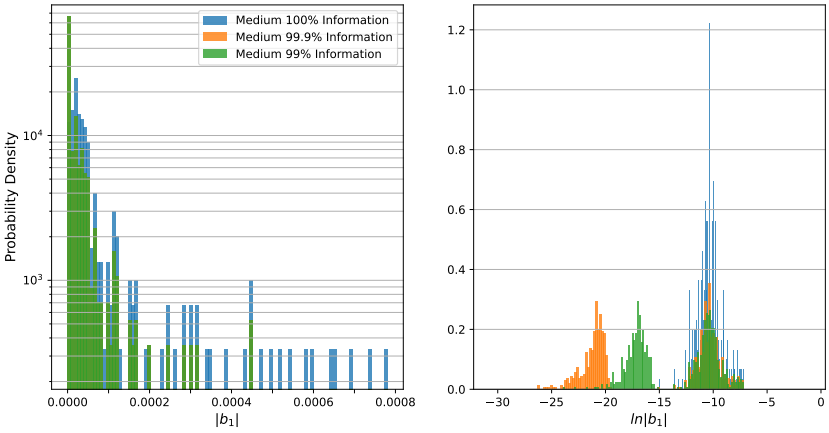


(b) Histograms of b_{24}.

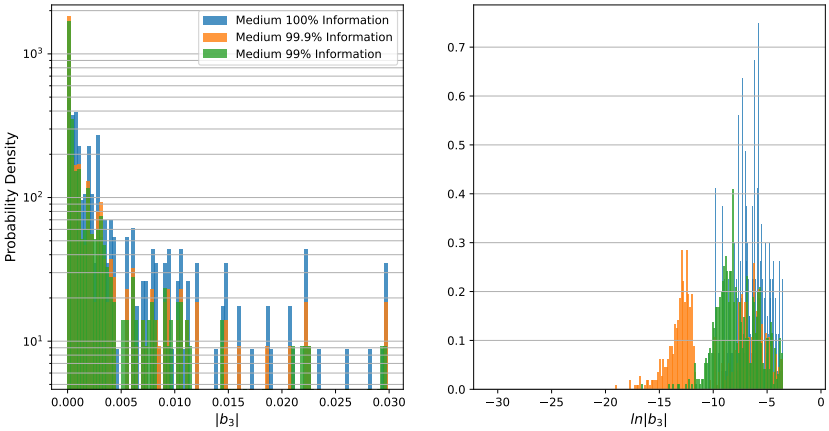


(c) Histograms of b_{57}.

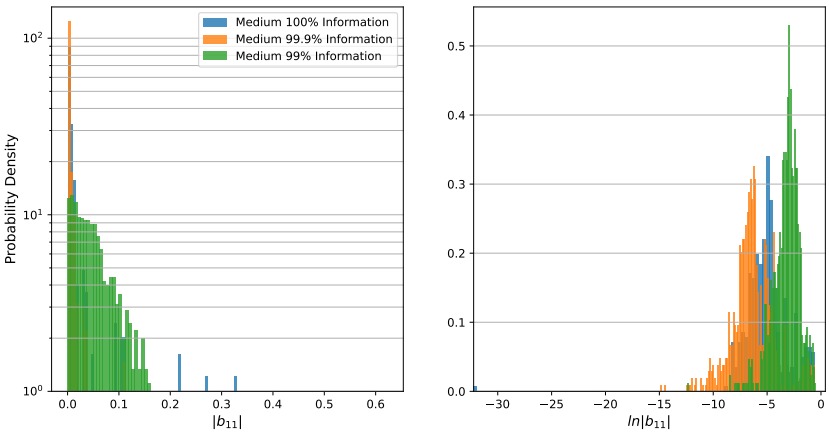
Fig. 2 Histograms of b_1 (a), b_{24} (b), and b_{57} (c) for the 100% information retained cases. Left plots show the absolute value and right plots show the natural log of the absolute value.



(a) Histograms of b_1.

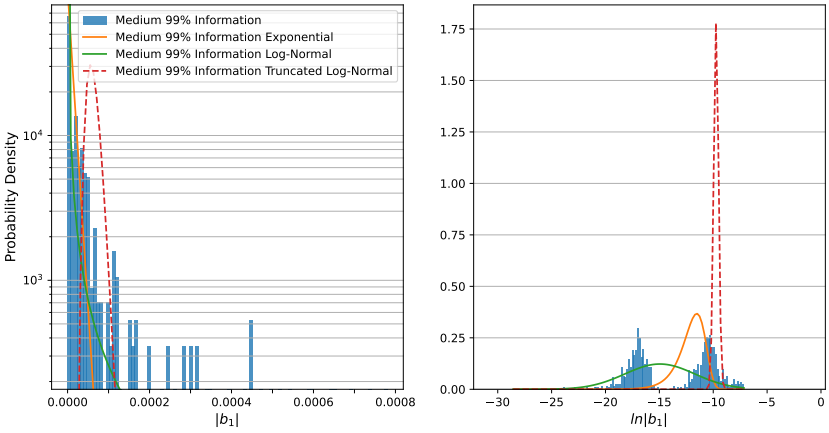


(b) Histograms of b_3.

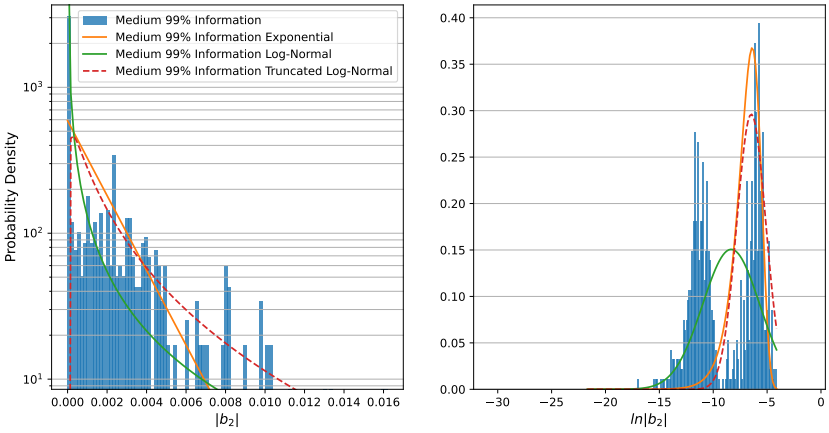


(c) Histograms of b_{11}.

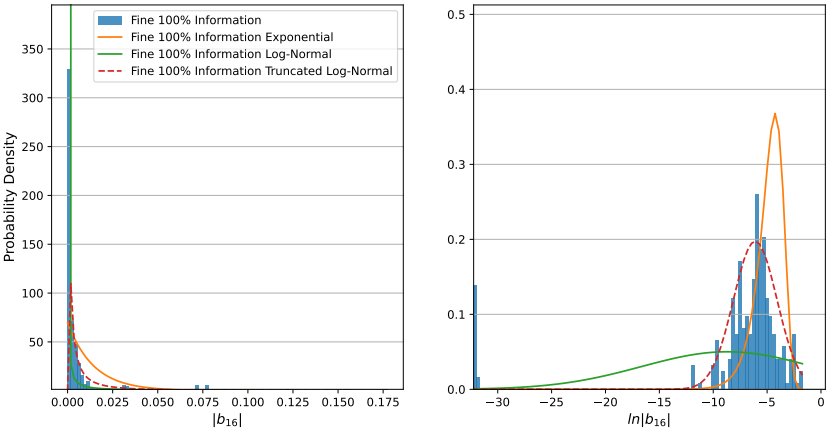
Fig. 3 Histograms of b_1 (a), b_3 (b), and b_{11} (c) for the medium case. Left plots show the absolute value and right plots show the natural log of the absolute value.



(a) b_1 of the 99% information model for the medium case.



(b) b_2 of the 99% information model for the medium case.



(c) b_{16} of the 100% information model for the fine case.

Fig. 4 Histograms with fitted PDFs. Left plots show the absolute value and right plots show the natural log of the absolute value.

B. Results on Test Case

Now that observations are fitted and β defined, the resulting distributions can be used. The same models used to generate the samples are now queried in accordance to Eq. 16. The models are queried on the same parameter grid as before with 1000 observations each. To sample both positive and negative from the log-normal distributions, there is a 50% chance the returned sample is negative otherwise sampling from the Laplace and log-normal distributions are consistent with usual process. These distributions are centered about 0 and symmetric; however, due to adding many distributions with long tails, it is possible that even with the 1000 observations the results may be asymmetric.

The focus of the results are on the lift coefficient (C_L), drag coefficient (C_D) and pitching moment coefficient (C_m) of the models. Since the source data are from a semispan model, the other coefficients are not significant. The results are explored in two ways. The first is a series of plots looking at derived distributions with α and M_∞ sweeps and the second is a series of tables summarizing the statistics.

These plots are organized at a constant value of either M_∞ or α with the other parameter being varied across its range. The first figure (Fig. 5) shows the resulting distribution of the Laplace fit on the medium data with 99.9% information retained at $M_\infty = 0.7$. The nominal line is in black, with the shaded region representing 90% coverage of the observations. The blue squares are the integrated values from the data used to train the model while the orange circles are the test data, which is the available CFD data that were not used in model generation. Since the fine case uses all available data, there is no test data. The plot shows C_L on the left, C_D in the middle, and C_m on the right all as a function of α. Since a linear interpolation is used for the baseline model, the linear connection between training points is expected as clearly seen in the C_m plot on the right. Because this is a model with an information threshold, the nominal is not expected to go precisely through the source data, but as can be seen at these scales the difference is negligible. This means the observed distribution is significantly greater than the effect of dropping modes. The jagged edges of the 90% interval is expected since these are long tail distributions, and it is not unusual to need more than the 1000 observations to create smooth results.

Looking in the other parametric direction, Fig. 6 shows the same model at a constant $\alpha = 2°$. As seen from the lack of blue squares this is a condition without any training data. It also shows the limitations of using a linear interpolation scheme with the resulting zig-zag seen in C_L and C_D. While the resulting nominal often misses the mark of the CFD, the distribution covers most of the points. This observation gives confidence that the methodology is working as intended. Here since the plotting range is smaller, the jagged edges are even more prominent.

When looking at the other possible fits, Fig. 7 shows the truncated log-normal fit for the 100% information model in the coarse case at α = -7° This is a much more jagged distribution than the Laplace, which is expected due to the even longer tails. Also as clearly seen with the jagged edge, the nominal is not at the center of the distribution. While in theory it should be, the chance for extreme tails on only one side is relatively high. Consistent with the other model and case, the CFD data not used in training are typically captured by the interval.

The final plot in Fig. 8 shows the log-normal distribution on the fine case with 99% of the information retained. This figure is at $M_\infty = 0.8$. Of immediate note is the extreme outliers that skew the scale of the plot. This shows the cost of including those tiny values when fitting the log-normal distributions. Future work should develop some technique to handle such data if a log-normal fit is desired.

While the individual figures are informative, evaluating the performance across the whole range is important. To that end, the size of the resulting distribution is tabulated. Each table looks at one case and one coefficient of interest.

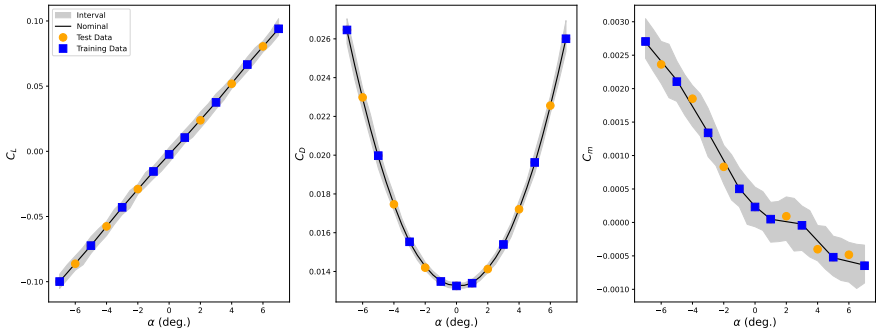


Fig. 5 The medium case with 99.9% information using a Laplace distribution for an α sweep at $M_\infty = 0.7$ (a) C_L (b) C_D (c) C_m.

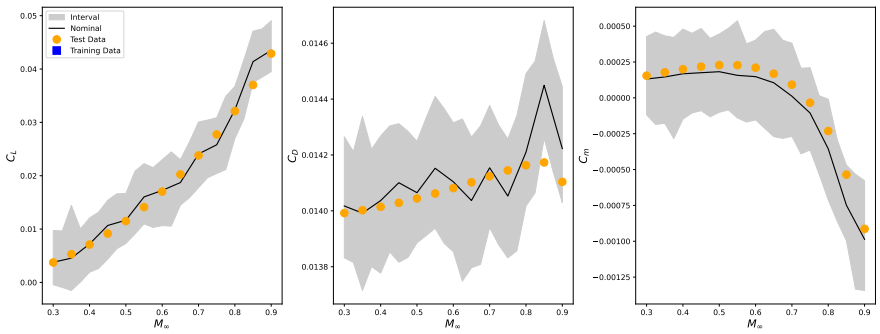


Fig. 6 The medium case with 99.9% information using a Laplace distribution for a M_∞ sweep at α = 2.0° (a) C_L (b) C_D (c) C_m.

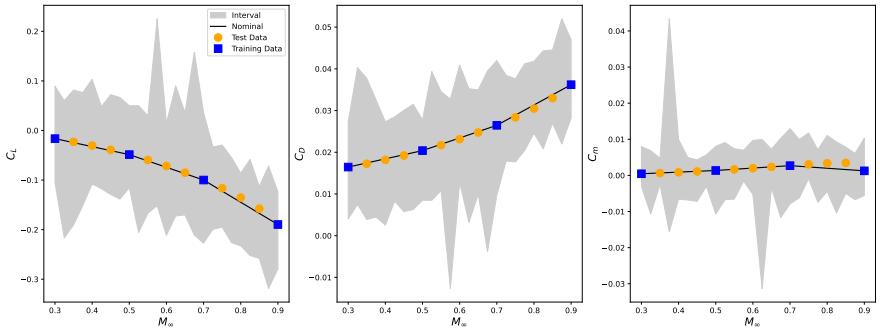


Fig. 7 The coarse case with 100% information using a truncated log-normal distribution for a M_∞ sweep at α = -7.0° (a) C_L (b) C_D (c) C_m.

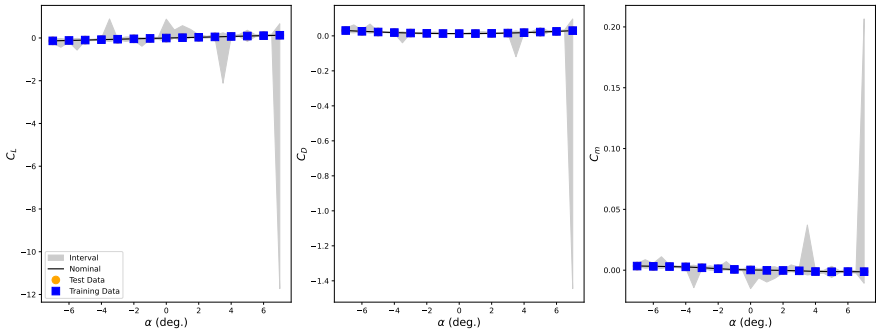


Fig. 8 The medium case with 99.9% information using a Laplace distribution for a M_∞ sweep at α = 2.0° (a) C_L (b) C_D (c) C_m.

To generate the tables, the differences from the nominal values are first taken. Since the distributions as fitted are not a function of α or M_∞, these residuals are pooled across the parameter space. The table then reports on the standard deviation, σ, and the range need to cover 50% and 90% of the observations. For example, Table 2 shows the performance of the three fits in the medium case at the 3 different information thresholds for C_L. First thing to note is the predicted σ for the 100% information log-normal fit is extremely high. This is a symptom of those "flattened" distributions as seen in Fig. 4c. Even though the chance is relatively small the huge outliers cause issues. As designed, by truncating the training set this is overcome and the σ value is in family with the other models. The σ values for the log-normal at other information thresholds are also higher then their peers but less significantly. Also of note is that these high σ values do translate to high ranges to cover 50% and 95% of the dispersed data seen in the next two columns; however only by a couple orders of magnitude instead of dozens.

The distributed statistics for the Laplace distribution and the truncated log-normal are relatively similar with minor differences. Based on the previous section, the difference between the two would be most different at small values of b; however these small values are rarely important when examining confidence and prediction intervals.

Table 3 shows the results of the medium case when looking at C_D. These results are in line with the results of C_L. Because the PCA + Interpolation model works by predicting the full surfaces it is unsurprising that the forces and moments have similar distributions. One of the advantages of this methodology operating on the underlying surfaces, is it is able to capture correlations between parameters of interests without extra considerations. Table 4 covers the results looking at C_L on the models trained on the fine case. As expected as the amount of source data increases, the resulting distributions are smaller. The same problems with the log-normal fit are seen; however, the 99.9% and 99% models have a 90% interval value similar to the others showing that while removing modes can increase the spread, it may introduce more robustness into the system. The tables for the remaining cases and pitching moment are in the Appendix and show similar trends. The tables confirm what was seen in the plotting where the Laplace distribution performs the best. The log-normal distribution has potential once the problem of the secondary cluster is overcome; however, until then there is no reason to use it over the Laplace.

IV. Conclusions

The developed method for resampling in the context of PCA + Interpolation has shown promising results. The method of least squares to find the distribution in the coefficients of the model is successful. It allows for starting the process of quantifying the uncertainty of these types of surrogate models. Future work would like to explore the value of the least square residuals and, if significant, incorporate them into the final model. After fitting and sampling the distributions, a Laplace distribution was found as the best fit for the data. While log-normal distribution also shows some promise, the amount of tuning in comparison for about the same performance will likely disqualify it from future analysis. The resulting integrated force predictions tend to cover CFD data not used in the creation of the models, while still having many of the limitation inherent in the PCA + Interpolation family of models.

To further improve the models, it could be possible to improve the interpolation method of the model; however, that was outside the scope of the work. The secondary cluster in values of b warrant further investigation. Based on

Table 2 Distribution of Medium Case Fits for C_L.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	2.259e-03	2.512e-03	7.312e-03
100	Log-Normal	3.981e+14	8.373e-01	4.041e+03
100	Truncated Log-Normal	3.617e-03	2.695e-03	9.033e-03
99.9	Laplace	1.212e-03	1.333e-03	3.922e-03
99.9	Log-Normal	2.480e+02	9.099e-04	2.800e-02
99.9	Truncated Log-Normal	6.131e-03	2.697e-03	9.852e-03
99	Laplace	1.392e-03	1.618e-03	4.518e-03
99	Log-Normal	7.067e-02	1.327e-03	9.241e-03
99	Truncated Log-Normal	4.790e-03	2.154e-03	9.614e-03

Table 3 Distribution of Medium Case Fits for C_D.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	5.477e-05	7.152e-05	1.795e-04
100	Log-Normal	1.724e+16	5.169e-01	2.012e+03
100	Truncated Log-Normal	2.548e-04	1.168e-04	4.206e-04
99.9	Laplace	3.990e-05	5.148e-05	1.306e-04
99.9	Log-Normal	5.198e-01	8.998e-05	5.299e-04
99.9	Truncated Log-Normal	1.573e-04	7.820e-05	2.714e-04
99	Laplace	3.677e-05	4.403e-05	1.199e-04
99	Log-Normal	2.824e-04	5.142e-05	1.783e-04
99	Truncated Log-Normal	5.354e-05	4.947e-05	1.542e-04

Table 4 Distribution of Fine Case Fits for C_L.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	8.290e-04	8.919e-04	2.686e-03
100	Log-Normal	4.867e+14	9.203e+00	1.113e+04
100	Truncated Log-Normal	1.222e-03	9.953e-04	3.287e-03
99.9	Laplace	2.987e-04	3.250e-04	9.669e-04
99.9	Log-Normal	7.711e+00	1.426e-04	3.003e-03
99.9	Truncated Log-Normal	1.191e-03	9.676e-04	3.220e-03
99	Laplace	3.134e-04	3.614e-04	1.015e-03
99	Log-Normal	7.166e-02	2.301e-04	1.343e-03
99	Truncated Log-Normal	1.015e-03	8.692e-04	2.813e-03

the theory that it is an artifact of the discrete resampling process used, a methodology based on work proposed in [24] could lead to a variable weighting of snapshots instead of the binary inclusion or exclusion in the current work. This continuous weighting would allow for exploration of small perturbations in the system and not just the large ones of removing entire snapshots. Also of interest is in the work of Lahiri [25], which has explored resampling with dependent data. If successful in incorporating his techniques, it could be possible to further improve this work by bringing in some of the other techniques to evaluate the accuracy of PCA based models.

References

[1] Berkooz, G., Holmes, P., and Lumley, J. L., "The proper orthogonal decomposition in the analysis of turbulent flows," *Annual review of fluid mechanics*, Vol. 25, No. 1, 1993, pp. 539–575.

[2] Jolliffe, I., *Principal Component Analysis*, 2nd ed., Springer, 2002.

[3] Bui-Thanh, T., Damodaran, M., and Willcox, K., *Proper Orthogonal Decomposition Extensions for Parametric Applications in Compressible Aerodynamics*, No. 0 in Fluid Dynamics and Co-located Conferences, American Institute of Aeronautics and Astronautics, 2003. https://doi.org/doi:10.2514/6.2003-4213, URL https://doi.org/10.2514/6.2003-4213.

[4] Boncoraglio, G., and Farhat, C., "Active manifold and model-order reduction to accelerate multidisciplinary analysis and optimization," *AIAA Journal*, Vol. 59, No. 11, 2021, pp. 4739–4753.

[5] Butler, T., Dawson, C., and Wildey, T., "Propagation of uncertainties using improved surrogate models," *SIAM/ASA Journal on Uncertainty Quantification*, Vol. 1, No. 1, 2013, pp. 164–191.

[6] Sudret, B., Marelli, S., and Wiart, J., "Surrogate models for uncertainty quantification: An overview," *2017 11th European conference on antennas and propagation (EUCAP)*, IEEE, 2017, pp. 793–797.

[7] Burkhead, A. C., and Dalle, D. J., "Statistically Consistent Dispersion of Line Loads to Uncertain Integrated Forces and Moments," *AIAA SCITECH 2024 Forum*, 2024, p. 1844.

[8] Chen, P., Quarteroni, A., and Rozza, G., "Reduced order methods for uncertainty quantification problems," *ETH Zurich, SAM Report*, Vol. 3, 2015.

[9] Efron, B., *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, 1982. https://doi.org/10.1137/1.9781611970319, URL https://epubs.siam.org/doi/abs/10.1137/1.9781611970319.

[10] J.J. Daudin, C. D., and Trecourt, P., "Stability of principal component analysis studied by the bootstrap method," *Statistics*, Vol. 19, No. 2, 1988, pp. 241–258. https://doi.org/10.1080/02331888808802095, URL https://doi.org/10.1080/02331888808802095.

[11] Aaron Fisher, B. S., Brian Caffo, and Zipunnikov, V., "Fast, Exact Bootstrap Principal Component Analysis for p > 1 Million," *Journal of the American Statistical Association*, Vol. 111, No. 514, 2016, pp. 846–860. https://doi.org/10.1080/01621459.2015.1062383, URL https://doi.org/10.1080/01621459.2015.1062383, pMID: 27616801.

[12] Braconnier, T., Ferrier, M., Jouhaud, J.-C., Montagnac, M., and Sagaut, P., "Towards an adaptive POD/SVD surrogate model for aeronautic design," *Computers & Fluids*, Vol. 40, No. 1, 2011, pp. 195–209. https://doi.org/https://doi.org/10.1016/j.compfluid.2010.09.002, URL https://www.sciencedirect.com/science/article/pii/S0045793010002306.

[13] Wignall, T., "Development of a Data Fusion Methodology for Lineload Aerodynamic Databases for a Launch Vehicle during Liftoff and Transition." Ph.D. thesis, NCSU, 2024.

[14] Wu, C., Liang, Y., Lin, W., Lee, H., and Lim, S., "A note on equivalence of proper orthogonal decomposition methods," *Journal of Sound and Vibration*, Vol. 265, No. 5, 2003, pp. 1103–1110.

[15] Taira, K., Brunton, S. L., Dawson, S. T., Rowley, C. W., Colonius, T., McKeon, B. J., Schmidt, O. T., Gordeyev, S., Theofilis, V., and Ukeiley, L. S., "Modal analysis of fluid flows: An overview," *AIAA Journal*, Vol. 55, No. 12, 2017, pp. 4013–4041.

[16] Sirovich, L., "Turbulence and the dynamics of coherent structures. I. Coherent structures," *Quarterly of applied mathematics*, Vol. 45, No. 3, 1987, pp. 561–571.

[17] Amsallem, D., and Farhat, C., "Interpolation method for adapting reduced-order models and application to aeroelasticity," *AIAA journal*, Vol. 46, No. 7, 2008, pp. 1803–1813.

[18] Edwards, C., Lee, M. W., Ramezanian, D., and Smith, R., "Quantifying Emergent Fluid Dynamics using Reynolds-interpolated Fluid Reduced-order Models," *AIAA SCITECH 2023 Forum*, 2023, p. 1376.

[19] O'Neil, P. J., Krekeler, G. C., Billman, G. M., and Creasman, F., "Aero Configuration / Weapons Fighter Technology (ACWFT) - Summary Technical Report," Tech. Rep. WL-TR-95-3002, Wright Labs, December 1994.

[20] Hunter, C., Viken, S., Wood, R., and Bauer, S., *Advanced aerodynamic design of passive porosity control effectors*, No. 0 in Aerospace Sciences Meetings, American Institute of Aeronautics and Astronautics, 2001. https://doi.org/doi:10.2514/6.2001-249, URL https://doi.org/10.2514/6.2001-249.

[21] Pandya, M. J., Diskin, B., Thomas, J. L., and Frink, N. T., "Assessment of USM3D Hierarchical Adaptive Nonlinear Method Preconditioners for Three-Dimensional Cases," *AIAA Journal*, Vol. 55, No. 10, 2017, pp. 3409–3424. https://doi.org/10.2514/1.J055823, URL https://doi.org/10.2514/1.J055823.

[22] Pandya, M. J., Frink, N. T., Ding, E., and Parlette, E., *Toward Verification of USM3D Extensions for Mixed Element Grids*, No. 0 in Fluid Dynamics and Co-located Conferences, American Institute of Aeronautics and Astronautics, 2013. https://doi.org/doi:10.2514/6.2013-2541, URL https://doi.org/10.2514/6.2013-2541.

[23] Pandya, M. J., Jespersen, D. C., Diskin, B., Thomas, J., and Frink, N. T., *Accuracy, Scalability, and Efficiency of Mixed-Element USM3D for Benchmark Three-Dimensional Flows*, No. 0 in AIAA SciTech Forum, American Institute of Aeronautics and Astronautics, 2019. https://doi.org/doi:10.2514/6.2019-2333, URL https://doi.org/10.2514/6.2019-2333.

[24] Lee, M., and Dowell, E. H., *Fluid Galerkin Reduced-order Models: Computational Efficiency and Sensitivity to Methods of Flow Decomposition*, 2020.

[25] Lahiri, S. N., *Resampling methods for dependent data*, Springer Science & Business Media, 2013.

Appendix: Tables of Distribution Performance

Table 5 Distribution of Coarse Case Fits for C_L.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	1.327e-02	1.311e-02	4.312e-02
100	Log-Normal	1.648e+13	2.856e-02	4.059e+00
100	Truncated Log-Normal	1.554e-02	1.360e-02	4.351e-02
99.9	Laplace	8.580e-03	8.435e-03	2.801e-02
99.9	Log-Normal	4.734e+02	3.122e-03	2.096e-01
99.9	Truncated Log-Normal	5.494e-02	1.436e-02	4.920e-02
99	Laplace	8.177e-03	8.071e-03	2.660e-02
99	Log-Normal	1.114e+00	2.993e-03	5.975e-02
99	Truncated Log-Normal	3.777e-01	3.550e-03	5.814e-02

Table 6 Distribution of Coarse Case Fits for C_D.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	1.637e-04	2.011e-04	5.339e-04
100	Log-Normal	1.707e+13	5.526e-03	4.260e+00
100	Truncated Log-Normal	3.071e-04	2.443e-04	7.934e-04
99.9	Laplace	9.454e-05	1.152e-04	3.086e-04
99.9	Log-Normal	2.371e+00	2.166e-04	4.401e-03
99.9	Truncated Log-Normal	1.167e-02	1.713e-04	1.616e-03
99	Laplace	1.254e-04	1.518e-04	4.099e-04
99	Log-Normal	2.705e-03	1.743e-04	8.077e-04
99	Truncated Log-Normal	8.997e-04	1.672e-04	7.192e-04

Table 7 Distribution of Coarse Case Fits for C_m.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	3.560e-04	4.193e-04	1.157e-03
100	Log-Normal	8.971e+12	4.777e-03	2.067e+00
100	Truncated Log-Normal	7.451e-04	4.618e-04	1.675e-03
99.9	Laplace	2.197e-04	2.562e-04	7.131e-04
99.9	Log-Normal	5.424e+00	3.932e-04	8.155e-03
99.9	Truncated Log-Normal	1.349e-02	3.722e-04	2.481e-03
99	Laplace	2.699e-04	2.943e-04	8.740e-04
99	Log-Normal	1.278e-02	2.730e-04	1.760e-03
99	Truncated Log-Normal	4.372e-03	2.734e-04	1.681e-03

Table 8 Distribution of Medium Case Fits for C_m.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	1.427e-04	1.741e-04	4.644e-04
100	Log-Normal	3.270e+15	3.217e-01	1.532e+03
100	Truncated Log-Normal	6.251e-04	1.941e-04	8.455e-04
99.9	Laplace	7.892e-05	9.783e-05	2.568e-04
99.9	Log-Normal	3.806e+00	1.248e-04	1.341e-03
99.9	Truncated Log-Normal	1.079e-03	1.219e-04	5.879e-04
99	Laplace	1.644e-04	1.686e-04	5.342e-04
99	Log-Normal	1.116e-03	1.628e-04	7.188e-04
99	Truncated Log-Normal	2.492e-04	1.634e-04	6.212e-04

Table 9 Distribution of Fine Case Fits for C_D.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	3.009e-05	3.948e-05	9.869e-05
100	Log-Normal	6.122e+14	1.622e+01	1.985e+04
100	Truncated Log-Normal	1.455e-04	7.686e-05	2.578e-04
99.9	Laplace	1.369e-05	1.788e-05	4.490e-05
99.9	Log-Normal	1.268e-02	2.309e-05	1.245e-04
99.9	Truncated Log-Normal	5.251e-05	3.457e-05	1.062e-04
99	Laplace	5.605e-06	6.909e-06	1.827e-05
99	Log-Normal	1.181e-04	5.207e-06	2.577e-05
99	Truncated Log-Normal	1.850e-05	2.229e-05	4.212e-05

Table 10 Distribution of Fine Case Fits for C_m.

% Information	Fit	σ	50% Interval	90% Interval
100	Laplace	5.038e-05	6.341e-05	1.645e-04
100	Log-Normal	1.090e+15	4.952e+00	6.439e+03
100	Truncated Log-Normal	1.401e-04	7.508e-05	2.708e-04
99.9	Laplace	2.011e-05	2.561e-05	6.570e-05
99.9	Log-Normal	1.344e-01	2.654e-05	2.166e-04
99.9	Truncated Log-Normal	1.372e-04	5.364e-05	2.195e-04
99	Laplace	4.991e-05	5.509e-05	1.616e-04
99	Log-Normal	1.250e-03	5.109e-05	2.119e-04
99	Truncated Log-Normal	8.042e-05	5.383e-05	1.946e-04