Aerodynamic Data Fusion Towards the Digital Twin Paradigm

S. Ashwin Renganathan †
Argonne National Laboratory
Lemont, IL 60439

Kohei Harada ‡ and Dimitri N. Mavris §
Georgia Institute of Technology
Atlanta, GA 30332

We consider the fusion of two aerodynamic data sets originating from differing fidelity physical or computer experiments. We specifically address the fusion of: 1) noisy and incomplete fields from wind tunnel measurements and 2) deterministic but biased fields from numerical simulations. These two data sources are fused in order to estimate the true field that best matches measured quantities that serves as the ground truth. For example, two sources of pressure fields about an aircraft are fused based on measured forces and moments from a wind-tunnel experiment. A fundamental challenge in this problem is that the true field is unknown and can not be estimated with 100% certainty. We employ a Bayesian framework to infer the true fields conditioned on measured quantities of interest; essentially we perform a statistical correction to the data. The fused data may then be used to construct more accurate surrogate models suitable for early stages of aerospace design. We also introduce an extension of the Proper Orthogonal Decomposition with constraints to solve the same problem. Both methods are demonstrated on fusing the pressure distributions for flow past the RAE2822 airfoil and the Common Research Model wing at transonic conditions. Comparison of both methods reveal that the Bayesian method is more robust when data is scarce while capable of also accounting for uncertainties in the data. Furthermore, given adequate data, the POD based and Bayesian approaches lead to similar results.

Nomenclature

\[\mathcal{N}\] = normal distribution
\[\mu\] = mean of multivariate normal distribution
\[\sigma, \tau\] = standard deviation

*Submitted to the editors
†Postdoctoral Appointee, Division of Mathematics & Computer Science, 9700 S Cass Ave. Member
‡Ph.D. Candidate, School of Aerospace Engineering, 270 Ferst Dr NW. Student Member
§Regents Professor, School of Aerospace Engineering, 270 Ferst Dr NW. Fellow
The Digital Twin (DT) concept in aerospace systems design is a vision that aims to achieve paradigm shift in flight certification, fleet management and sustainment [1]. The basic idea is to develop a simulator integrating high-fidelity physics models, historical flight-test data and sensor updates amongst others, to mirror the operation of the corresponding flying twin. This would in turn enable real-time health monitoring of those vehicles during their operation that could potentially result in reliable and efficient designs that meet the design requirements of the future. The trend in aircraft design over the past couple decades have resulted in an abundance of data being generated from three main sources namely, numerical simulations, wind-tunnel measurements and flight testing; thereby laying the foundation towards realizing an aerodynamic digital twin. Particularly, the steady growth in computational capabilities in the past few decades have assisted aircraft manufacturers to numerically simulate the aerodynamics of an aircraft under real flight conditions. This comes at a significantly cheaper cost than wind-tunnel or flight tests and thereby replacing or supplementing them to a good extent [2]. On the other hand, flight test and wind tunnel data, though not abundant are available in non-negligible amounts. Numerical simulations at the practical scale are seldom exact and suffer from model bias, whereas measurements are noisy, incomplete and could be contaminated by errors introduced due to model scale and instrumentation among other factors. Therefore, as a means for realizing the DT vision, there is a need to take advantage of available data by combining them in a fashion that is likely to make them more accurate than
their individual selves. We refer to this problem as multifidelity data fusion and within the context of this work, we consider aerodynamic pressure fields arriving from computational fluid dynamics (CFD) simulations and wind-tunnel measurements. We distinguish the present work from some of the past work where the term data fusion has been used in the context of using variable fidelity simulation codes to accelerate design optimization; for instance see Keane (2003) [3]. On the contrary, the present work is purely data-driven and uses only domain knowledge about the actual models involved as will be demonstrated in later sections.

An associated challenge with the data fusion problem is to also quantify the confidence in the fused data. As mentioned earlier, neither the CFD predictions nor the wind-tunnel field measurements are devoid of uncertainties and therefore their fusion is expected to inherit those uncertainties. Uncertainty in CFD results could originate from inadequate physical models (such as turbulence closures [4,5]), geometry discrepancy, and discretization errors that lead to numerical diffusion, just to name a few. The wind-tunnel measurements are affected by model installation effects (such as due to wall and sting), simulation of proper boundary-layer development, and flow angle correction that could lead to errors in directional static stability. Quantifying these uncertainties is a challenge in itself; however we assume they are known and show how to properly account for them in the proposed approach. The goal of data fusion therefore, is to synergistically combine datasets from multiple fidelity sources, to improve the overall prediction in addition to quantifying uncertainty in the overall prediction.

Combining aero/fluid dynamic data from computer and physical experiments is commonly done with the Proper Orthogonal Decomposition (POD) originally introduced by Lumley [7] and Sirovich [8] in the context of turbulent fluid flow, where the combined flow snapshots are stacked in a matrix and a set of orthonormal modes are extracted via the singular value decomposition (SVD); the state of the flow is then assumed to be a linear combination of these modes. In a way, the POD leads to a dimensionality reduction from the original state-space dimension to the number of POD modes which are typically much fewer. Taking advantage of this fact, the gappy POD introduced by Sirovich [9] imputes missing data in fluid dynamic datasets where the POD modes are first extracted from the complete datasets. Then the missing data from the incomplete sets are estimated as the best linear combination of the POD modes nearest to the incomplete dataset in the least-squares sense. See [10] for further details. Ruscher et al [11] supplement the gappy POD with wavelet methods to ensure continuity in the reconstructed data, while Wen et al [12] apply it when there are two sets of missing data but in a pattern that complement each other. Another technique that has a very similar goal as data fusion is the so-called Data Assimilation (DA), which refers to combining physical measurements with mathematical models particularly in the context of dynamical systems. DA originated in the geosciences field for weather prediction [13,14]. The methodology proposed in this work differs from DA by the fact that it is purely data-driven. Although domain knowledge about the models that generated the data is leveraged, the models themselves don't feature in the proposed method, as in DA. Overall, our method is more generic, robust to small data and applies to any situation where there are multiple sources of data and some measurement of the ground truth.
As a first step, this work proposes a methodology to fuse two datasets that could be corrupted by noise, bias and/or incompleteness. As mentioned before, we consider field data that represent the state of the system; such as the distribution of pressure on the surface of an aircraft wing at a specific flight condition. Using this information and additional measurements of quantities of interest (QoIs) such as forces and moments, that are considered the ground truth, the inverse problem of estimating the true field by fusing the available datasets that best matches the measurements is posed and solved. The measurements of QoIs, although corrupted by noise and wind-tunnel scaling effects like the field measurements are extracted from more mature and reliable instrumentation such as force balance. Furthermore, these measurements are independent of the field measurements, i.e. they are not calculated as a function of the fields but are directly measured. For these reasons, the QoI measurements are treated as the ground truth in this work. An schematic of the overall methodology is provided in Figure 1.

The inversion of the pressure field from the measured forces and moments is an ill-posed problem which is easy to see. For instance an arbitrary shift in the pressure distribution over a closed aerodynamic surface does not change the the value of its integral; i.e. \( \int_S p \, ds = \int_S (p - p_{\infty}) \, ds \). In this regard, we incorporate a Bayesian framework such that regularizing priors can be specified to tackle the ill-posedness of the inverse problem, in addition to accounting for uncertainties associated with the datasets; this method was originally introduced in [15]. This way, the method performs a statistical correction on the datasets and infers the probability distribution of the fused data. The output of the proposed method may then be used to construct more accurate surrogate models of the fields which can be queried cheaply to solve problems in aerospace design such as optimization and uncertainty quantification. Additionally, we extend the work in [15] and introduce a POD-based method to solve the same problem, namely the POD with constraints (CPOD). In contrast to the Bayesian approach, this method searches the POD subspace generated from the data to find the fused data corresponding to a given flight condition. However, the CPOD, similar to the Bayesian method also treats the measured forces and moments as the ground truth.

This work makes the following assumptions. Firstly, we consider only two levels of fidelity for the sake of simplicity although it extends to a hierarchy of fidelities without any modification except for the notations. Secondly, we assumes identical geometric fidelity for all the sources. For instance, it does not fuse wing and airfoil pressure distributions. Thirdly, we assume that a pair of fidelity data are available for all flight conditions. In practice, the case might be that the wind-tunnel and CFD date are do not correspond to identical flight conditions. To meet this requirement, the CFD data used in this work is generated to match the flight conditions of a pre-existing wind-tunnel dataset. Fourthly, we assume that the uncertainty in the available data sources are known. Whereas in the wind-tunnel measurements an estimate of these uncertainties are available from replications and sensor calibration tests, quantifying all forms of uncertainty in CFD data is not trivial. While this assumption makes the proposed method subjective, it also adds the flexibility of leveraging expert domain knowledge when available. Finally, viscous effects are not accounted for in the methodology because field measurements of shear stress are typically hard to obtain. For this reason, the QoIs considered in this work
are restricted to the lift and moment coefficients where the contributions from viscosity are relatively marginal.

The rest of the article is organized as follows. The Bayesian methodology is outlined in the next section. The two test cases and their associated detail are discussed in section III. Demonstration of the proposed Bayesian method is shown in section IV and the CPOD method and comparison of both methods are discussed in section V. Finally the paper concludes with a summary of main outcomes of the study and some directions for future work.

II. Data Fusion via Bayesian Inference

In section IV we demonstrate the methodology on the coefficient of pressure ($C_p$) distribution on an aircraft wing-section as well as the entire aircraft. However, here we keep the exposition general and consider two independent fidelity sources of data $y_1 \in \mathbb{R}^n$ and $y_2 \in \mathbb{R}^n$ represented by random variable $Y$ with densities

$$
\pi_y(y_1) = N(\mu_1, \sigma_1^2 I_n)
$$

$$
\pi_y(y_2) = N(\mu_2, \sigma_2^2 I_n)
$$

where $\sigma_1^2$ and $\sigma_2^2$ are the variance in the datasets and are assumed known. Furthermore, $\mu_1$ and $\mu_2$ are the expected values of the datasets which in the case of computer experiments are the direct predictions and in the case of physical experiments could be the ensemble averaged measurements of the fields. Since the source of $\mu_1$ and $\mu_2$ could be vectors...
of disparate lengths, we interpolate them onto a common grid of length \( n \). Let \( \mathbf{z} = [z_1, \ldots, z_m]^\top \), \( z_i \in \mathbb{R} \) be \( m \) QoI’s which are a function of \( \mathbf{y} \) and whose direct measurements are also available. Furthermore, we assume that the forward problem \( \mathbf{z} = f(\mathbf{y}) \) is known and we are interested in solving the inverse problem \( \mathbf{y} = f^{-1}(\mathbf{z}) \) given \( \mathbf{z} \). Finally we restrict the current scope of the problem to be linear and hence \( \mathbf{z} = f(\mathbf{y}) = \mathbf{H}\mathbf{y} \), where \( \mathbf{H} \) is a linear operator \( \in \mathbb{R}^{n \times m} \) that maps the field to the QoI. In the field of applied aerodynamics such a linear assumption is valid since quantities such as the lift, drag and moment coefficients are linear functions of the pressure distribution.\(^\ast\) The \( \mathbf{z} \) are assumed to be noisy and are related to the forward model via the following relationship

\[
\mathbf{z} = \mathbf{H}^\top \mathbf{y} + \epsilon
\]

where \( \epsilon \) represents additive Gaussian white noise with probability density \( \pi_E \) given by \( N(0, \tau^2 \mathbf{I}_m) \), where \( \tau^2 \) is again assumed known. Our goal is now to estimate the probability distribution of the unknown \( \mathbf{y} \) given the measurements \( \mathbf{z} \) and this is done via Bayes’ rule explained as follows.

A. Bayes’ Rule

The cornerstone of the present methodology is the Bayes’ rule, stated in (3), whose highlight is that it operates on the probability densities of the variables rather than the variables themselves, thereby incorporating the uncertainty in the data naturally. The main idea is that, to infer the distribution of some unknown parameter conditioned on available data, \( \pi(\mathbf{y}|\mathbf{z}) \)^† the likelihood of observing the data given the model and its inputs, \( \pi(\mathbf{z}|\mathbf{y}) \) can be consolidated with whatever prior belief is available about the unknown parameter itself, denoted by \( \pi(\mathbf{y}) \) and are related to each other as follows

\[
\pi(\mathbf{y}|\mathbf{z}) = \frac{\pi(\mathbf{z}|\mathbf{y}) \pi(\mathbf{y})}{\pi(\mathbf{z})} \tag{3}
\]

In the equation above, \( \pi(\mathbf{y}|\mathbf{z}) \) is called the posterior density and \( \pi(\mathbf{z}) \) is the marginal density of \( \mathbf{z} \) which can be evaluated by integrating out \( \mathbf{y} \) from the joint density \( \pi(\mathbf{z}, \mathbf{y}) \). For practical purposes we are interested in some moment of the posterior density such as its mode or expected value. Since \( \pi(\mathbf{z}) \) evaluates to a constant, its evaluation can be ignored in such situations. In this work we shall work with the value of the parameter estimate that maximizes the posterior probability which is called the maximum *aposteriori* (MAP) estimate given by

\[
\mathbf{y}^* = \arg\max_{\mathbf{y}} \pi(\mathbf{z}|\mathbf{y}) \pi(\mathbf{y}) \tag{4}
\]

which can be evaluated via gradient-based non-linear optimization methods. We now proceed to give specific details on the prior and likelihood models.

\(^\ast\)note that we do not account for viscous effects in this context since they are typically unavailable from physical experiments

\(^\dagger\)Note that we remove the subscripts in the density for convenience of notation
B. Prior Distribution

A prior distribution on the unknown $y$ can be thought of as a regularizer that restricts the posterior distribution to physically valid solutions. This is where we take advantage of the available information in the form of $y_1$ and $y_2$ and combine them to define the prior. In this work, we introduce a fusion parameter $\theta \in \mathbb{R} \subseteq [0, 1]$ which combines $y_1$ and $y_2$ as

\[
\hat{\mu} = \theta \times \mathbb{E}(y_1) + (1 - \theta) \times \mathbb{E}(y_2)
\]

\[
= \theta \times \mu_1 + (1 - \theta) \times \mu_2
\]

where $\theta$ is chosen as the solution of

\[
\theta = \arg\min_{\theta} \|H^T \hat{\mu} - z\|^2_2
\]

The prior distribution on $y$ is then set as a multivariate Gaussian distribution centered on $\hat{\mu}$ given by

\[
\pi(y) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (y - \hat{\mu})^T \Sigma^{-1} (y - \hat{\mu}) \right\}
\]

In other words, we first chose the linear combination of the two sources of information that best matches the observed data (in the least squares sense) and define a Gaussian distribution centered around this estimate. This is built on the belief that the true value of $y$ lies somewhere in the neighborhood of $y_1$ and $y_2$ which is approximated as a linear weighted combination. Although the current way of estimating $\hat{\mu}$ is not unique, it provides a very simple and intuitive way of specifying the prior; i.e. it is a linearly weighted average of the available data. Furthermore, such prior specification are called sample based priors \[10\] where prior belief is expressed via a combination of sample solutions of the unknown. Note that alternatively, one can treat $\theta$ as a random variable and use a hierarchical Bayes \[17\] framework and instead infer the posterior $\pi(\theta|z)$ with some prior on $\pi(\theta)$. In that case, with a uniform prior on theta and a Gaussian likelihood, using the MAP estimate of $\pi(\theta|z)$ is equivalent to the present approach. However we do not take that route in the present work to favor simplicity of exposition. In \[7\] we treat $y$ to be spatially correlated and hence define the covariance matrix $\Sigma$ as

\[
\Sigma_{ij} = \text{Cov}(y(x_i), y(x_j)) = \sigma^2 \exp \left( -\frac{\|x_i - x_j\|^2}{2\ell^2} \right)
\]

where, $x$ denote the spatial coordinates with $\| \cdot \|_2$ denoting the Euclidean distance, the parameter $\ell$ represents the length scale that is assumed to be known and (due to independence of $y_1$ and $y_2$) $\sigma^2 = \theta^2 \sigma_1^2 + (1 - \theta)^2 \sigma_2^2$. The choice of the squared-exponential kernel in \[8\] is to specify smoothness in the prior realizations although other choice of kernels may be considered depending on the problem. See Ch.4 of \[18\] for a compendium of covariance kernels.
C. Likelihood Model

Assuming that $\varepsilon$ and $y$ are mutually independent, the probability density of $z$, conditional on $Y = y$ is obtained by shifting the density $\pi_E$ around $H^T y$ leading to the likelihood density

$$
\pi(z|y) \sim N(H^T y, \tau^2 I_m)
$$

where again, $\tau^2$ is the measurement noise.

D. Maximum a-Posteriori (MAP) Estimation

We are interested in solving for the inverse problem of estimating the true field given the measurements in terms of its probability density function, $\pi(y|z)$, which is given by

$$
\pi(y|z) \propto \pi(z|y) \times \pi(y)
$$

$$
\pi(y|z) \propto \exp \left\{ -\frac{1}{2\tau^2} ||z - H^T y||^2 \right\} \times \exp \left\{ -\frac{1}{2}(y - \tilde{\mu})^T \Sigma^{-1} (y - \tilde{\mu}) \right\}
$$

By (10) what we mean is that we estimate the true $y$ distribution that best fits the measured value of the quantity of interest while also being similar to the prior elicited for $y$ via (7). The mode of the resulting posterior probability distribution is the MAP estimate we are interested in. It should be noted that since the posterior distribution in the present context is Gaussian, the mean, median and mode are identical and hence the choice does not matter. Therefore we are interested in solving the following optimization problem

$$
y_{MAP} = \arg \max_y \exp \left\{ -\frac{1}{2\tau^2} ||z - H^T y||^2 \right\} \times \exp \left\{ -\frac{1}{2}(y - \tilde{\mu})^T \Sigma^{-1} (y - \tilde{\mu}) \right\}
$$

which is equivalent to solving

$$
y_{MAP} = \arg \min_y \frac{1}{2\tau^2} ||z - H^T y||^2 + \frac{1}{2}(y - \tilde{\mu})^T \Sigma^{-1} (y - \tilde{\mu})
$$

$$
y_{MAP} = \arg \min_y J(y) \text{ (say)}
$$

where $J$ denotes the entire cost function. Since (12) is differentiable everywhere we evaluate its gradient and set it to zero. Furthermore, $J(y)$ is a symmetric positive-definite quadratic form which has a unique minimizer. Therefore we write the gradient as

$$
\frac{\partial J}{\partial y} = \frac{1}{\tau^2} (H^T y - z) + \frac{1}{2} (y - \tilde{\mu}) \Sigma^{-1} (y - \tilde{\mu})
$$

and solve

$$
y_{MAP} = \arg \min_y \frac{1}{\tau^2} (H^T y - z) + \frac{1}{2} (y - \tilde{\mu}) \Sigma^{-1} (y - \tilde{\mu})
$$

8
\[
\frac{\partial J}{\partial y} = \frac{1}{\tau^2} \left[ -(z - H^Ty)^T H^{-1} y + (y - \tilde{\mu})^T \Sigma^{-1} \right] = 0
\]  \hspace{1cm} (13)

Rearranging above equation gives

\[
y^T_{MAP} \left[ \frac{1}{\tau^2} H H^T + \Sigma^{-1} \right] = \frac{1}{\tau^2} z^T H^T + \tilde{\mu}^T \Sigma^{-1}
\]

which gives

\[
y^T_{MAP} = \left[ \frac{1}{\tau^2} z^T H^T + \tilde{\mu}^T \Sigma^{-1} \right]^{-1} \left[ \frac{1}{\tau^2} H H^T + \Sigma^{-1} \right] \]

It can then be shown that the posterior distribution is given by

\[
\pi(y|z) \sim N(y^T_{MAP}, \Gamma)
\]

where,

\[
\Gamma = \left[ \frac{1}{\tau^2} H H^T + \Sigma^{-1} \right]^{-1}
\]

The diagonal elements of \( \Gamma \) contain the pointwise variance of \( y|z \) which may be used to construct confidence intervals on the predictions. The hyperparameters of the methodology are \( \gamma = \{\sigma_1^2, \sigma_2^2, \tau^2, \ell\} \). Among them, \( \sigma_1^2, \sigma_2^2, \tau^2 \) represent the uncertainty in the available dataset including the measured QoIs. Quantifying these uncertainties in the data is a very elaborate task which is not undertaken in this work. Instead we present a method that fuses the datasets by accounting for these uncertainties. As for the length-scale parameter \( \ell \), although it can be estimated from data, in the present work we fix its value for each of the test cases listed in section III. It is chosen from trial-and-error such that the realizations from the prior look physically reasonable. Note that one of the primary advantages of the Bayesian framework is the ability to specify subjective priors that leverage domain knowledge. The method is summarized in Algorithm 1.

**Algorithm 1:** Bayesian Data Fusion

**Result:** \( y_{MAP}, \Gamma, \) confidence intervals

**Data:** fields: \( \mu_1, \mu_2 \), measurements: \( z \), parameters: \( \gamma \)

1) Estimate \( \theta \) from (6)

2) Compute \( y_{MAP} \) from (14)

3) Compute posterior covariance matrix from (15)

4) Extract standard deviation to construct confidence intervals; \( \pm \sqrt{\text{diag}(\Gamma)} \)
III. Test Cases and Data

The two main test cases for the methods used in this work are the viscous transonic flow past the RAE2822 airfoil and the NASA Common Research Model (CRM). The corresponding wind-tunnel data are extracted from [19] and [20] respectively. Specifically, for the CRM test case, the pressure distributions from pressure-sensitive paint (PSP) [21, 22] measurements are obtained from the NASA Ames 11ft Transonic Wind Tunnel provided in [23]. All the CFD simulations were performed using the commercial, finite-volume based unstructured code STAR-CCM+ [24].

RAE-2822

The RAE2822 is a commonly used benchmark test case for transonic flight conditions. The airfoil shape and the mesh for the RANS analysis are shown in Figure 2. A mesh with approximately 122,500 polyhedral mesh elements is generated with 41 prism layers to capture the boundary layer. The first layer of the prism layer is placed approximately $1.6 \times 10^{-6}$ m away from the wall such that the wall $y^+ \approx 1$ for the Reynolds number ranges considered in this work. The AGARD [19] wind-tunnel measurements are available for both the pressure distribution as well as the lift and moment coefficients and the operating conditions are summarized in Table 1. The airfoil surface is discretized into an $n = 128$ size equally-spaced grid on which both the CFD and wind-tunnel data are interpolated before performing the fusion.

| Case | Mach  | Reynolds number (millions) | Angle of attack (degrees) |
|------|-------|---------------------------|---------------------------|
| 1    | 0.676 | 5.7                       | 2.40                      |
| 2    | 0.676 | 5.7                       | -2.18                     |
| 3    | 0.600 | 6.3                       | 2.57                      |
| 4    | 0.725 | 6.5                       | 2.92                      |
| 5    | 0.725 | 6.5                       | 2.55                      |
| 6    | 0.725 | 6.5                       | 3.22                      |
| 7    | 0.730 | 6.5                       | 3.19                      |
| 8    | 0.750 | 6.2                       | 3.19                      |
| 9    | 0.730 | 2.7                       | 3.19                      |
| 10   | 0.745 | 2.7                       | 3.19                      |
| 11   | 0.740 | 2.7                       | 3.19                      |

Fig. 2 The RAE2822 airfoil shape and near-field mesh with polyhedral elements.
**Common Research Model**

The clean wing-body configuration of the CRM without tail-planes is used in this work. A mesh with approximately 24 million trimmed hexahedral elements were generated as shown in Figure 3 where the boundary layer is resolved with 41 prism layers and the wall $y^+ \approx 1$. In order to keep the computational costs of the method tractable, the CFD and PSP data are interpolated onto a coarse grid consisting of $n = 8688$ cells (further details are provided in the Appendix B). The viscous contributions to the lift coefficient is in $O(10^{-5})$ and the moment coefficient is in $O(10^{-3})$ across all the test cases; while this is negligible for the lift, it is not so for the moments. In this work, the viscous contributions are ignored to keep the exposition simple although they can be easily included by adding an offset parameter in (6) and (9) of the form $\|z - H^T y - \delta\|$ where $\delta > 0$ is a discrepancy parameter and is in the same order as the viscous contributions. The test data for the CRM are set at a Reynolds number of 5 million and a full-factorial design of Mach = \{0.70, 0.85, 0.87\}, angle of attack = \{0, 0.5, 1.0, 1.5, 2.0, 3.0, 4.0\}. The results are demonstrated in the next section for a selected combination of operating conditions.

**IV. Method Demonstration**

Here we demonstrate the methodology on inferring the true $C_P$ distributions given measurements. Here $z_1 = C_L/C_I$, $z_2 = C_M/C_m$, $\mu_1 = C_{P,PSP/WT}$, $\mu_2 = C_{P,CFD}$ and $\tilde{\mu} = \tilde{C}_P$. The matrix $H \in \mathbb{R}^{n \times 2}$ where the two columns contain information about the local surface normals, cell areas and free-stream flow necessary to numerically integrate the pressure distributions to compute $C_L/C_I$ and $C_M/C_m$ respectively.
Consider the Figure 4a where the CFD prediction and the wind-tunnel measurement of the $C_P$ distribution (for flight condition $Mach = 0.676$, $Re = 5.7M$, $\alpha = -2.18 \ deg.$) match quite closely. However, the QoI’s ($C_l$, $C_m$) obtained by integrating the $C_P$ curves (evaluating the forward model) from the wind-tunnel and CFD $C_P$’s are $(-0.122, -0.074)$ and $(-0.101, -0.075)$ respectively. Both QoI’s are off from the measured QoI’s $(-0.121, -0.028)$ although, the discrepancy is more pronounced for $C_m$. In this case, no matter what value of $\theta$ is chosen, the resulting $C_P$ is not expected to produce outputs that match the measurements. Therefore, we are interested in adjusting the curves in Fig. 4a such that the QoI’s derived from the adjusted curve matches the measured values in the least-squares sense. We set $\tau^2 = 10^{-6}$, $\sigma_1^2 = 10^{-2}$, $\sigma_2^2 = 10^{-2}$ based on the belief that the the pressure distributions from CFD predictions and wind-tunnel measurements have greater uncertainty than the force and moment measurements; while $\ell$ is set to $10^{-4}$. The statistically adjusted solution is then given by the posterior distribution (15) and 500 draws from this distribution are shown in Fig 4b. The MAP estimate is the expectation of the posterior distribution and is shown in Fig 4c overlaid with plots for the CFD and wind-tunnel distributions. Notice that the current approach shrinks the best combination of the two distributions in order to minimize the misfit between the derived QoIs and the measurements. It should be noted that the MAP estimate should not be treated as the true $C_P$ distribution that uniquely defines the state of the system at the given operating conditions. This is because the MAP estimate depends on the chosen values of the hyperparameters. Additionally, it tries to minimize the misfit between noisy data and a deterministic model while not accounting for viscous effects. Therefore the fused $C_P$ should be analyzed in terms of its probability densities which account for the uncertainty rather than treating it as a deterministic estimate. Having said that, the MAP estimate provides an interpretable visualization of the posterior distribution and is used while comparing various curves.

A point worth mentioning is the impact on the results due to the parameters. In Fig 4 the parameter $\tau^2$ was set to $10^{-6}$ which implies that we trust the measured QoIs to possess very high signal-noise ratio. As a result the proposed approach tries to match them as closely as possible leading to relatively more adjustment on the original $C_P$ curves. On the other hand if we admit our ignorance about the actual value of the measurements and set a higher $\tau^2$, then the method leads to relatively less adjustment. This is demonstrated in figures 5a and 5b where the $\tau^2$ is set to $10^{-4}$ and $10^{-2}$ respectively. Notice that when $\tau^2 = 10^{-2}$ the uncertainty in the QoIs (particularly $C_m$) is very high that the misfit term in (12) becomes less important and the prior dominates; as a result $C_{P,MAP} \approx \bar{C}_P$. Larger the $\tau^2$ parameter, greater the misfit between the measured QoIs and those obtained by integrating the MAP estimate of $C_P$ (see the legend entries of Figures 5).

As a counter example, consider the case shown in Figure 6 which corresponds to the flight conditions ($Mach = 0.740$, $Re = 2.7M$, $\alpha = 3.19 \ deg.$). For this case the QoI’s integrated from the wind-tunnel curve (0.7049, -0.0875) match very closely with the measurements (0.7061, -0.087) and as a result the predicted MAP estimate for $C_P$ almost overlaps with the wind-tunnel curve (Fig. 6c). Therefore in this case, the choice of the parameters play a relatively minor role.
B. Common Research Model

For the CRM test case, the approach is repeated and results are presented for a select few cases as follows. The hyperparameters are set to $\tau^2 = 10^{-6}$, $\sigma_1^2 = 10^{-2}$, $\sigma_2^2 = 10^{-2}$ and $\ell = 0.01$. Figure 7 shows results for $Mach = 0.87$ and $\alpha = 4.0 \ deg$. where the CFD predicts a relatively stronger shock compared to PSP measurements. Incidentally, the MAP estimate matches more closely with the CFD results compared to PSP. Furthermore, the PSP curve falls outside the 95% confidence region of the MAP in the vicinity of the shock, suggesting that the PSP measurements in this region are less reliable. While a different choice of parameter values will change the size of the confidence bands, the results clearly demonstrate that the CFD results agree more closely with the measurements than PSP for this specific case. A contrasting behavior is observed in Figure 8 where the MAP estimate matches more closely with the PSP results compared to CFD. Finally, in the case of Figure 9, the MAP estimate lies between the CFD and PSP estimates where
the CFD predictions fall outside of the 95% confidence bounds around the MAP estimate, in the vicinity of the shock. These results demonstrate that the proposed approach allows one to quantify the spatially distributed confidence in the fused data by accounting for the uncertainties in the original datasets. We claim that this makes the fused data more useful than the original datasets and one may then use the inferred posterior distributions to construct surrogate models of the fields using model reduction methods [25, 26] for instance.

An alternative way to use the proposed method is to invoke domain knowledge in the prior specification. For instance, if domain knowledge recommends more trust in one amongst the PSP or CFD data, then the \( \theta \) parameter can be treated as a tuning parameter whose value can be directly set as opposed to estimating it. The plots in Figure 10 represent an operating condition (\( Mach = 0.87 \), \( \alpha = 1.5 \) deg.) where a strong shock wave is expected on the wing. While this is predicted by the CFD simulations, the PSP measurements show rather shallow gradients in the coefficient of pressure (Figs. 11d through 11f). Therefore the user might choose to put more prior belief in the CFD predictions by setting \( \theta = 0 \) in (5). In this case, the method tries to adjust the CFD predictions so as match the measured QoI which results in a MAP estimate that looks relatively similar to the actual CFD prediction. This is demonstrated in Figure 11.

It should be noted that the MAP estimate is itself non-unique - i.e. there can be more than one MAP estimate that agrees with the measurements \( \mathbf{z} \) equally well. The estimate depends on the prior specification as illustrated through Figures 10 and 11 re-iterating the ill-posedness issue of the inverse problem mentioned in section I. Therefore, it is critical to specify the prior more judiciously.

The Table 2 summarizes the integrated QoIs (by evaluating forward model) from the MAP, CFD and PSP \( C_P \) distributions and compares them against the measured values. It can be seen that the MAP estimates match the measured
values very closely. Therefore, the methodology overall finds the best estimate that agrees with measurements given the multiple sources of data, regularized by the specified prior belief on the estimate. The methodology is computationally efficient since the posterior distributions are analytically derived as opposed to making numerical approximations and requires only approximately 1 minute of wall-clock time to fuse one pair of dataset \( n \approx 10000 \) on an Intel 4-core i7 processor with 16gb RAM. This establishes the scalability of the proposed approach for large datasets.

One of the key characteristics of the proposed Bayesian approach is that it assumes that more than one source of the dataset is available for a given operating condition and operates only on this dataset. However, typically windtunnel and CFD campaigns are conducted in batches and the available datasets span a range of operating conditions. In the following section, we propose another solution for the same problem that relies on learning flow features from the entire dataset, unlike the proposed Bayesian approach following which, we show comparisons of both methods.
Table 2  Summary of the CRM results in terms of the QoIs ($C_L$ and $C_M$). The data under columns labeled MAP, PSP and CFD contain the QoI’s from evaluating the forward model at the $C_P$ distributions. Bold-face entries indicate best match to measurements

| Mach/Re/$\alpha$ | Measurements | MAP | PSP | CFD |
|------------------|--------------|-----|-----|-----|
| $C_L$            | $C_M$        | $C_L$ | $C_M$ | $C_L$ | $C_M$ | $C_L$ | $C_M$ |
| 0.87/5m/4.0      | 0.6733 -0.0898 | **0.6732** -0.0898 | 0.6024 -0.1111 | 0.6773 -0.0932 |
| 0.87/5m/3.0      | 0.5416 -0.0762 | **0.5416** -0.0767 | 0.5105 -0.1159 | 0.2593 -0.0870 |
| 0.7/5m/1.5       | 0.2938 -0.0429 | **0.2935** -0.0431 | 0.2703 -0.0654 | 0.3646 -0.0959 |
| 0.85/5m/1.5      | 0.3383 -0.0486 | **0.3378** -0.0489 | 0.2758 -0.0482 | 0.4206 -0.1206 |

Fig. 7  $Mach = 0.87$, $Re = 5M$, $\alpha = 4.0$ deg.
Fig. 8  $M = 0.7, \text{Re} = 5M, \alpha = 1.5$

Fig. 9  $\text{Mach} = 0.87, \text{Re} = 5M, \alpha = 3.0 \text{ deg.}$
Fig. 10  \( M = 0.87, Re = 5M, \alpha = 1.5. \) \( \theta \) estimated from (6)

Fig. 11  \( Mach = 0.87, Re = 5M, \alpha = 1.5 \) deg.. \( \theta \) set to 0 to express greater prior belief on the CFD predictions
V. Data Fusion via Proper Orthogonal Decomposition

We now present the proper orthogonal decomposition with constraints (CPOD), which uses the POD [7, 8] to first construct an orthogonal sub-space (also known as 'POD modes') from data (available $C_P$ distributions). Then the unknown fused $C_P$ is approximated as a linear expansion of the POD modes, whose parameters (coordinates) can be estimated via linear least-squares methodology. A schematic of the method is provided in Figure 12. We begin with a brief review of POD before proceeding to outline the proposed CPOD method.

A. Proper Orthogonal Decomposition

POD was originally introduced in the context of turbulent flow modeling by Holmes et al [7], where it was used to characterize the coherent structures in the flow from wind tunnel measurements. POD has a special characteristic of optimality in that it provides the most efficient means to capture the dominant components of a process [27]. Given a state variable $u \in \mathbb{R}^n$ which may be the numerical solution of a PDE on a computational mesh of size $n$ or measurements from a physical experiment (such as Particle Image Velocimetry), the POD expresses $u$ as the linear expansion on a finite number of $k$ orthonormal basis vectors $\phi_i \in \mathbb{R}^n$. That is,

$$u \approx \sum_{i=1}^{k} a_i \phi_i$$

(16)

where, $a_i$ is the $i$th component of $a \in \mathbb{R}^k$ and are the coefficients of the basis expansion. It can be shown that [27, 28] the POD modes in the above equation are the same as the left singular vectors of the snapshot matrix (obtained by stacking $q$ snapshots of $u$), $U = [u_1, \ldots, u_q]$. That is,

$$U = \Phi D \Psi^T$$

(17)

then $\Phi_k$ represents the first $k$ columns of $\Phi \in \mathbb{R}^{n \times q}$, after truncating the last $q - k$ columns based on the relative magnitudes of the cumulative sum of their singular values. The $L_2$ error in approximation of the state variables due to the POD basis expansion is then given as

$$\sum_{j=1}^{q} \|u_j - (\Phi_k \Phi_k^T)u_j\|_2^2 = \sum_{i=k+1}^{q} d_i^2$$

(18)

where $d_i$ is the singular value corresponding to the $i$th column of $\Phi$ and is also the $i$th diagonal element of $D$. We choose $k$ such that $\sum_{i=1}^{k} d_i / \sum_{i=1}^{q} d_i = 0.99$, which essentially means we retain the modes that explain approximately 99% of the variability in the dataset.

Let the (unknown) fused $C_P$ be denoted $\tilde{u}$. Firstly, we make the assumption that $\tilde{u}$ is missing from the dataset but is in the subspace spanned by $\Phi_k$ and hence $\tilde{u} \approx \Phi_k a$, where $a$ is to be estimated. Secondly, we know that $\tilde{u}$ has to...
match the measurements $z$ and hence one may consider estimating $a$ by minimizing $\|H^T \Phi_k a - z\|_2^2$, however this is an underdetermined problem unless $k \leq m$, which is practically an unlikely scenario. To overcome this we combine it with the minimization of $\|\Phi_k a - \tilde{u}\|_2^2$, however $\tilde{u}$ is unknown. Therefore, we introduce an iterative method that implements the CPOD which is described as follows.

### B. CPOD

The unknown $\tilde{u}$ is first provided an initial guess following which $a$ is estimated via solving the minimization problem

$$a^* = \arg\min_a \frac{1}{2} \|\Phi_k a - \tilde{u}\|_2^2$$

subject to $H^T \Phi_k a = z$  

Then, the guess for $\tilde{u}$ is updated as $\tilde{u} \leftarrow \Phi_k a^*$. The updated $\tilde{u}$ is then used to enrich the POD modes, i.e. $\Phi_k \leftarrow \text{thin-svd}([U, \tilde{u}])$ and Eq.(19) is solved again. The procedure is repeated until the cost function ($J$, to be defined in (20)) reaches a steady value, which in this work is measured based on the standard deviation ($\Delta_c$) of $J$ over the previous $c$ iterations falling within some specified threshold $\epsilon_c$, i.e. $\Delta_c = \frac{1}{c} \sum_{i=1}^c J_i^2 - \left(\frac{1}{c} \sum_{i=1}^c J_i \right)^2 \leq \epsilon_c$, where $J_i$ is the value of the cost function at the $i$th iteration. This stopping criterion is used based on the assumption that we are interested in the objective function reaching a steady value rather than strictly zero. Note that this iterative approach involves
re-computing the SVD of the data at every iteration whose computational cost can be mitigated by performing rank-1 updates to SVD \cite{29,30}. We now show how to solve \cite{19}, which is a linearly constrained least-squares problem \cite{31}.

We introduce the penalized cost-function $J$ and re-write the constrained optimization problem \cite{19} as

$$a^*, \lambda^* = \arg\min_{a, \lambda} J(a, \lambda) = \frac{1}{2} (\Phi_k a - \tilde{u})^T (\Phi_k^T a - \tilde{u}) + \lambda^T (H^T \Phi_k a - z) \quad (20)$$

where $\lambda \in \mathbb{R}^{m \times 1}$ is the vector of Lagrange parameters. The partial derivatives of the cost function in \cite{20} with respect to $a$ and $\lambda$ are then set equal to zero

$$\frac{\partial J}{\partial a} = \Phi_k^T (\Phi_k a - \tilde{u}) + \Phi_k^T H \lambda = 0 \quad (21)$$

$$\frac{\partial J}{\partial \lambda} = H^T \Phi_k a - z = 0$$

where the first line of \cite{21} consists of $k$ equations and the second consists of $m$ equations. The full set of equations in \cite{21} represent the Karush-Kuhn-Tucker (KKT) conditions \cite{31,32} for constrained optimization which can be written in matrix form as

$$\begin{bmatrix}
I_k & \Phi_k^T H \\
H^T \Phi_k & 0
\end{bmatrix}
\begin{bmatrix}
a \\
\lambda
\end{bmatrix} =
\begin{bmatrix}
\Phi_k^T \tilde{u} \\
z
\end{bmatrix} \quad (22)$$

where the $I_k$ is a $k \times k$ identity matrix that results from $\Phi_k^T \Phi_k = I_k$ and $0$ is a $m \times m$ matrix of zeros. It can be shown that the coefficient matrix in \cite{22} is invertible provided $H$ is full rank (see Appendix A) and hence the least-squares problem has a unique solution. The solution of \cite{22} involves computation of a matrix inverse during every iteration but is computationally cheap since the matrix has the reduced $(k + m) \times (k + m)$ size instead of the full $n \times n$ size where, $k + m << n$.

The CPOD method is summarized in Algorithm 2 where it treats the available data to be devoid of any uncertainties and hence the $u_i$ are equivalent to the $\mu$ used the in the Bayesian method. Furthermore, the CPOD algorithm relies on a rich basis set obtained from POD and hence the snapshot matrix ($U$) combines all the available data (CFD and
wind-tunnel) in step-1 of Algorithm 2.

Result: Fused vector $\tilde{u}$

Data: Snapshot matrix $U = [u_1, \ldots, u_q]$, 
Fields at target flight condition $u_{CFD}, u_{WT}$,  
QoIs at target flight condition $z$,  
Stopping criterion $c = 5, \epsilon_c = 10^{-6}$

1) Extract POD modes $U = \Phi D\Psi^T$ Determine rank $k$;  
   $$\Phi_k = \Phi(:, 1 : k)$$

2) Guess $\tilde{u}$ as $\theta \times u_{CFD} + (1 - \theta) \times u_{WT}$ where $\theta \subseteq [0, 1]$;

3) while $\Delta c \geq \epsilon_c$ do
   • Compute $a^*$ via solving (19)  
   • Update $\tilde{u} \leftarrow \Phi_k a^*$  
   • Update $\Phi_k$: $[U, \tilde{u}]_{\text{thin-svd}} = \Phi D\Psi^T$  
   • Compute cost function, (20)  
   • Update $\Delta c$
end

Algorithm 2: CPOD - Proper Orthogonal Decomposition with Constraints

C. Discussion

We now compare and contrast the performance of the CPOD against the Bayesian approach. As we shall demonstrate, the CPOD relies on learning the flow features from the entire dataset. Therefore when it is deemed necessary to enrich the POD basis set, additional data is generated via CFD since experimental data are not available on demand. Additionally, since the cost of generating CFD data for the RAE test case is significantly cheaper than the CRM model, we restrict comparison of the methods based on the RAE test case only. Experimental data corresponding to 11 flight conditions are available (see Table 1) for the RAE test case; the number of snapshots, $q$ is therefore the CFD data generated in addition to these 11 cases. We pick two contrasting cases (2 & 11) to demonstrate the comparison in favor of keeping the discussion precise.

1. Effect of Snapshot Size $q$

Since the CPOD relies on searching for the unknown $\tilde{u}$ from a subspace, the richness of the snapshot matrix plays a major role in the performance of the method unsurprisingly; this is demonstrated in Figure 13. Here, a maximin Latin Hypercube designs [17] of size 20, 40 and 80 are generated in the $(Mach, Re, \alpha)$ space to augment the experimental data leading to $q = 31, 51$ and 91 snapshots respectively. Note that the $q = 22$ case contains 11 CFD cases corresponding to exactly the same flight conditions as in Table 1.
With too few snapshots ($q = 22$), the CPOD determines a $\tilde{u}$ that satisfies the forward model within the specified tolerance but results in a $C_P$ curve that looks noisy and physically unrealistic (Figure 13a). This also demonstrates the ill-conditioning of the problem mentioned in section I, i.e., there are more than one $C_P$ curve that satisfies the forward model to match the measured QoIs. Recall that the Bayesian approach discussed earlier overcomes this issue with the prior regularization. With more snapshots, the CPOD subspace is enriched leading to smoother $C_P$ curves. The sensitivity to the snapshot size manifests more prominently only when the discrepancy between the forward model predictions and the measured QoIs is significant, as in Case 2 shown in Figure 13a. For this specific case, $q = 91$ is necessary to ensure sufficiently smooth $C_P$ curves. As a counter example, Figure 13b shows that fewer snapshots are sufficient for this specific case. However, based on the worst-case scenario (Case-2), $q$ is fixed at 91 for the rest of the results.

![Figure 13 Effect of snapshot size $q$ on the smoothness of CPOD predictions.](image)

(a) Case-2: $\text{Mach} = 0.676, \text{Re} = 5.7 \times 10^6, \alpha = -2.18 \text{ deg}$. (b) Case-11: $\text{Mach} = 0.74, \text{Re} = 2.7 \times 10^6, \alpha = 3.19 \text{ deg}$.

An aspect we care about as part of the data fusion framework is uncertainty quantification. Whereas the MAP results are a function of the input uncertainties in the data, the CPOD results do not account for any of the uncertainties and treats the dataset $\{u_i\}$ as outputs of deterministic experiments as mentioned before. However, in the next subsection we show how approximate confidence bounds can be constructed for the CPOD predictions.

2. Confidence bounds for CPOD

One of the obvious limitations of the CPOD is its inability to account for uncertainty in the datasets. However, one can use a frequentist approach to obtain confidence bounds as demonstrated here. Assuming that the unknown fused $C_P$ is the true mean of a multi-variate normally distributed population with unknown covariance function, the confidence intervals are approximated via the Student’s $t$– distribution [33] described as follows. Given $T$ realizations of the $\tilde{u}$ that
results from repeating Algorithm 2 with independent draws of \( \theta \) from a uniform distribution \( \mathcal{U}(0, 1) \), a \( t \)–distribution of \( \nu = T - 1 \) degrees of freedom is defined as the distribution of the location of the sample mean relative to the true mean, divided by the sample standard deviation. This way, the \( t \)–distribution can be used to construct confidence bounds for the true mean. The sample mean (\( \hat{\mu} \)) and covariance (\( \hat{\Sigma} \)) are given by

\[
\hat{\mu} = \frac{1}{T} \sum_{i=1}^{T} \tilde{u}_i
\]

\[
\hat{\Sigma} = \frac{1}{T-1} \sum_{i=1}^{T} [\tilde{u}_i - \hat{\mu}] [\tilde{u}_i - \hat{\mu}]^T
\]

The \( 1 - \beta \) confidence intervals can then be provided as \( \hat{\mu} \pm t_{1-\beta, \nu} \, \text{diag}(\hat{\Sigma}) / \sqrt{T} \) where \( \text{diag}(\hat{\Sigma}) \) are the diagonal elements of \( \hat{\Sigma} \) and \( t_{1-\beta, \nu} \) is the \( t \)–value corresponding to a \( (1 - \beta) \times 100\% \) confidence. The 95\% confidence bounds are plotted for the two RAE cases in Figure 14 with \( T = 1000 \). The bounds are more pronounced in for the CPOD results in 14b compared to 14a particularly around the shock. This is because the CPOD relies mainly on minimizing the QoI data misfit whereas the Bayesian method accounts for uncertainties in the data. It is emphasized that the confidence bounds in CPOD are an artifact of the randomness in the initial guess provided (via the parameter \( \theta \)) and does not account for the the input uncertainties, unlike the Bayesian method. In that sense, the CPOD confidence bounds might not be highly useful in uncertainty quantification although they visualize the effect of the mis-specification of the initial guess to the method. In all the plots in this section showing CPOD predictions, \( T = 1000 \) unless otherwise mentioned.

The CPOD and Bayesian methods present two approaches towards fusing data from multiple fidelity experiments. The common thread between the two approaches is that they both find the best fit between the predictions of the forward model and the measurements of the QoI’s in the least-squares sense. However, they differ in a fundamental way - whereas the CPOD is a data-driven approach that essentially depends on the flow-features learned from the data (in the form of POD modes), the Bayesian approach relies on subjective prior specifications. Another important aspect that distinguish the approaches is that while the Bayesian approach accounts and propagates uncertainties in the available data, the CPOD does not. Despite the differences, the authors have observed that with adequate snapshots the CPOD predictions are very similar to the MAP predictions of the Bayesian method.

3. CPOD vs MAP

The predictions from the Bayesian (MAP) and the CPOD (with \( q = 91 \)) methods are compared for the RAE cases 2 and 10 in Figure 15. Surprisingly, the MAP and the CPOD predictions are very similar despite the fact that the CPOD depends on the entire dataset (\( q = 91 \) in this case) to learn the flow features in the data and provide a physically realistic result whereas the MAP required only one snapshot each from the two fidelity experiments. The difference is more pronounced in cases where the discrepancy between the QoI’s predicted from the forward model and the measurements
is more pronounced, as shown in Figure 15a. Although the CPOD has learned flow-features from data, in this case the snapshots are dominated by CFD data which inherits the bias in the data as well. Therefore it leads to a slightly different prediction compared to the MAP which depends only on the model misfit and prior specification (which depends only on $u_{CFD}$ and $u_{WT}$). Despite the differences, one can appreciate the similarity between the predictions at a very high level.

4. Computational Aspects

The CPOD relies on an appropriate initial guess very similar to the requirement of the prior in the Bayesian method. As mentioned in section 2.1, both methods can lead to physically unrealistic results while still satisfying the forward problem within a specified tolerance. This work relies on the assumption that the unknown (fused) $C_P$ is somewhere in the neighborhood of the available $C_P$’s from the various fidelity experiments. And within such a sufficiently small search space, it is observed that the CPOD and the MAP predictions are quite similar.

One of the significant differences between the two methods (in addition to those mentioned in the previous subsection)
is the computational time complexity. It would appear that the CPOD incurs significantly more computational cost since it operates on a snapshot set in contrast to the Bayesian approach. Additionally, the SVD (and recursive SVD-update) steps scale with the data size. However, the Bayesian approach involves the inversion of matrices of size $n \times n$ which is expensive as $n$ gets bigger. On the other hand, the CPOD method is iterative although it was observed to converge monotonically within 3-5 iterations for all the cases tested in this work; an example of which is shown in Figure 16. It is however worth recalling that the reduced computations in the Bayesian approach are a consequence of the choice of conjugate priors that lead to analytically closed-form expression for the MAP and the confidence bounds. This will not be the case in a more general setup with generic priors and the estimation of the hyperparameters included in the method. A very rough comparison of the computational time-complexity for both methods are provided in Table 3 in terms of the order of magnitude ($O$) of floating point operations. Note that in the comparison, the Bayesian method assumes Gaussian prior and likelihoods and known hyperparameters whereas the CPOD represents one repetition at a fixed $\theta$. Although from a time-complexity point of view, the Bayesian method seems expensive, from a wall-clock time point of view either methods take $O(\text{secs})$ in a standard workstation (Intel i7, 1.9GHz, 4 Core) at the time of writing this article for both the 2D and 3D cases. This demonstrates the feasibility of the present implementation with real large-scale data.
VI. Conclusion

A novel method to fuse noisy, incomplete and biased aerodynamic field information from wind-tunnel measurements and high-fidelity mathematical model predictions is presented. The method applies Bayesian inference to solve for the fused $C_P$ distribution given predictions from computer experiments and wind-tunnel measurements. The approach depends on providing a prior belief on the unknown $C_P$ distribution as a linear combination of the field data. Then a likelihood model is defined that minimizes the least-squares misfit between the output QoIs (computed via the forward model (2)) and the measurements. Combining the prior belief with the likelihood via Bayes’ rule, the fused $C_P$ distribution is estimated as a probability distribution. The parameters of the methodology are the uncertainties associated with the datasets as well as the measurements in addition to a correlation length-scale that determines the smoothness of the prior. One of the main assumptions made in this work is the specification of the prior via the estimation of the parameter $\theta$. Future directions in this work shall add more rigor to the method by acknowledging our ignorance on $\theta$ and estimating it via a Hierarchical Bayes method. The method is successfully demonstrated on the transonic flow past a wing section (RAE2822 airfoil) as well as an entire aircraft (CRM). Overall, the proposed method performs a statistical adjustment to the available field data based on very measurements of QoIs treated as ground truth. The adjusted data can then be used to construct data-driven models towards realizing the DT goal.

As an alternative method to solve the same problem, the CPOD method is introduced which, like the Bayesian method minimizes the misfit between the predicted and measured QoIs, however the search space is constrained by the POD subspace. Comparison of the CPOD and Bayesian methods revealed that they result in surprisingly similar results despite significant differences in their approach. A known limitation of the CPOD is that it currently does not account for the uncertainty in the dataset; future directions in this work aims to devise methods that account for the noise in the SVD step. Another observed limitation of the CPOD is that in cases where experimental measurements of the fields is
limited, additional snapshots are likely to be generated from CFD. This could lead to biased results since the POD basis could inherit the bias present in the CFD data.

Another focus of future work is the hybridization of the Bayesian and CPOD method - i.e. specification of informative priors derived from the POD. While this is a relatively straightforward extension, the authors are investigating the remedies to the known limitations of the proposed methodologies as a first step.

**Acknowledgments**

The material is based upon work supported by Airbus in the frame of the Airbus / Georgia Tech Center for MBSE-enabled Overall Aircraft Design and by the U.S. Department of Energy, Office of Science, under contract number DE-AC02-06CH11357. The insets used in Figures 11 and 12 showing wind-tunnel measurements are adapted from [www.nasa.gov](http://www.nasa.gov).

**Government License:** The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory ("Argonne"). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

**References**

[1] Glaessgen, E., and Stargel, D., “The digital twin paradigm for future NASA and US Air Force vehicles,” 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA, 2012, p. 1818.

[2] Johnson, F. T., Tinoco, E. N., and Yu, N. J., “Thirty years of development and application of CFD at Boeing Commercial Airplanes, Seattle,” *Computers & Fluids*, Vol. 34, No. 10, 2005, pp. 1115–1151.

[3] Keane, A., “Wing optimization using design of experiment, response surface, and data fusion methods,” *Journal of Aircraft*, Vol. 40, No. 4, 2003, pp. 741–750.

[4] Edeling, W., Cinella, P., Dwight, R., and Bijl, H., “Bayesian estimates of parameter variability in the k–ε turbulence model,” *Journal of Computational Physics*, Vol. 258, No. 4, 2014, pp. 1–9.

[5] Emory, M., Larsson, J., and Iaccarino, G., “Modeling of structural uncertainties in Reynolds-averaged Navier-Stokes closures,” *Physics of Fluids*, Vol. 25, No. 11, 2013. doi:10.1063/1.4824659.
[6] Mathelin, L., Hussaini, M. Y., and Zang, T. A., “Stochastic approaches to uncertainty quantification in CFD simulations,” *Numerical Algorithms*, Vol. 38, No. 1-3, 2005, pp. 209–236.

[7] Holmes, Philip., Lumley, John L., Berkooz, Gahl and Rowley, C. W., *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Vol. 36, 1998. doi:10.2514/2.399.

[8] Sirovich, L., “Turbulence and the dynamics of coherent structures. I - Coherent structures. II,” *Quarterly of Applied Mathematics (ISSN 0033-569X)*, Vol. 45, No. July, 1987, p. 561.

[9] Everson, R., and Sirovich, L., “Karhunen–Loeve procedure for gappy data,” *JOSA A*, Vol. 12, No. 8, 1995, pp. 1657–1664.

[10] Bui-Thanh, T., Damodaran, M., and Willcox, K. E., “Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition,” *AIAA Journal*, Vol. 42, No. 8, 2004, pp. 1505–1516.

[11] Ruscher, C. J., Dannenhoffer, J. F., and Glauser, M. N., “Repairing Occluded Data for a Mach 0.6 Jet via Data Fusion,” *AIAA Journal*, Vol. 55, No. 1, 2016, pp. 255–264. doi:10.2514/1.j054785, URL: [https://doi.org/10.2514/1.J054785](https://doi.org/10.2514/1.J054785).

[12] Wen, X., Li, Z., Peng, D., Zhou, W., and Liu, Y., “Missing data recovery using data fusion of incomplete complementary data sets: A particle image velocimetry application,” *Physics of Fluids*, Vol. 31, No. 2, 2019. doi:10.1063/1.5079896.

[13] Dee, D. P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, d. P., et al., “The ERA-Interim reanalysis: Configuration and performance of the data assimilation system,” *Quarterly Journal of the royal meteorological society*, Vol. 137, No. 656, 2011, pp. 553–597.

[14] Kalnay, E., *Atmospheric modeling, data assimilation and predictability*, Cambridge university press, 2003.

[15] Renganathan, A., Harada, K., and Mavris, D. N., “Multifidelity Data Fusion via Bayesian Inference,” *AIAA Aviation 2019 Forum*, 2019, p. 3556.

[16] Calvetti, D., and Somersalo, E., “Inverse problems: From regularization to Bayesian inference,” *Wiley Interdisciplinary Reviews: Computational Statistics*, Vol. 10, No. 3, 2018, p. e1427.

[17] Santner, T. J., Williams, B. J., Notz, W., and Williams, B. J., *The design and analysis of computer experiments*, Vol. 1, Springer, 2003.

[18] Williams, C. K., and Rasmussen, C. E., *Gaussian processes for machine learning*, Vol. 2, MIT press Cambridge, MA, 2006.

[19] “EXPERIMENTAL DATA BASE FOR COMPUTER PROGRAM ASSESSMENT - Report of the Fluid Dynamics Panel Working Group 04,” Tech. rep., AGARD Advisory Report No. 138, 1979. URL: [http://eda-ltd.com.tr/caeeda_doc/AGARD-AR-138.pdf](http://eda-ltd.com.tr/caeeda_doc/AGARD-AR-138.pdf).

[20] Vassberg, J., Dehaan, M., Rivers, M., and Wahls, R., “Development of a common research model for applied CFD validation studies,” *26th AIAA Applied Aerodynamics Conference*, 2008, p. 6919.
[21] McLachlan, B. G., and Bell, J. H., “Pressure-sensitive paint in aerodynamic testing,” Experimental thermal and fluid science, Vol. 10, No. 4, 1995, pp. 470–485.

[22] Morris, M., Donovan, J., Kegelman, J., Schwab, S., Levy, R., and Crites, R., “Aerodynamic applications of pressure sensitive paint,” AIAA journal, Vol. 31, No. 3, 1993, pp. 419–425.

[23] Rivers, M. B., and Dittberner, A., “Experimental investigations of the NASA common research model,” Journal of Aircraft, Vol. 51, No. 4, 2014, pp. 1183–1193.

[24] STAR-CCM+, 2019. URL https://mdx.plm.automation.siemens.com/star-ccm-plus

[25] Renganathan, S. A., Liu, Y., and Mavris, D. N., “Koopman-Based Approach to Nonintrusive Projection-Based Reduced-Order Modeling with Black-Box High-Fidelity Models,” AIAA Journal, Vol. 56, No. 10, 2018, pp. 4087–4111.

[26] Renganathan, S. A., “A Methodology for Non-Intrusive projection-based model reduction of expensive black-box PDE-based systems and application in the many-query context,” Ph.D. thesis, Georgia Institute of Technology, 2018.

[27] Holmes, Philip., Lumley, John L., Berkooz, Gahl and Rowley, C. W., Turbulence, Coherent Structures, Dynamical Systems and Symmetry, Vol. 36, 1998. doi:10.2514/2.399.

[28] Chatterjee, A., “An introduction to the proper orthogonal decomposition,” Current Science, Vol. 78, No. 7, 2000, pp. 808–817.

doi:10.1109/LPT.2009.2020494.

[29] Levey, A., and Lindenbaum, M., “Sequential Karhunen-Loeve basis extraction and its application to images,” IEEE Transactions on Image processing, Vol. 9, No. 8, 2000, pp. 1371–1374.

[30] Brand, M., “Fast low-rank modifications of the thin singular value decomposition,” Linear algebra and its applications, Vol. 415, No. 1, 2006, pp. 20–30.

[31] Boyd, S., and Vandenberghe, L., Convex optimization, Cambridge university press, 2004.

[32] Nocedal, J., and Wright, S., Numerical optimization, Springer Science & Business Media, 2006.

[33] Ross, S., A first course in probability, Pearson, 2014.

[34] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., “Numerical recipes in C++,” The art of scientific computing, Vol. 2, 1992, p. 1002.

[35] Ahrens, J., Geveci, B., and Law, C., “Paraview: An end-user tool for large data visualization,” The visualization handbook, Vol. 717, 2005.
Appendix

A. Proof of existence of solution for (22)

Here we show that a unique solution exists for the Equation (22).

Theorem 1: The KKT matrix

\[
\begin{bmatrix}
I_k & \Phi_k^\top H \\
H^\top \Phi_k & 0
\end{bmatrix}
\]

is invertible if \( H \) is full rank. Consider the following matrix inversion lemma by Press et al. [34]. If \( A = \begin{bmatrix} P & Q \\ R & S \end{bmatrix} \), then \( A^{-1} = \begin{bmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{bmatrix} \)

where,

\[
\begin{align*}
\tilde{P} &= P^{-1} + P^{-1} Q M R P^{-1} \\
\tilde{Q} &= -P^{-1} Q M \\
\tilde{R} &= -M R P^{-1} \\
\tilde{S} &= M 
\end{align*}
\]

and

\[ M = (S - R P^{-1} Q)^{-1} \]

Proof. For the coefficient matrix in (22) to be invertible, it suffices to show that \( M \) exists, since \( P = I_m \) and hence \( P^{-1} \) exists. The \( M \) can be written as

\[
M = \left[ 0 - (H^\top \Phi_k) \times (\Phi_k^\top H) \right]^{-1}
\]

(25)

The \( m \) columns of \( H \) contain the integral operators to compute the forces and moments from the \( C_p \) distributions. More specifically, the first column computes \( C_L \) and the second column computes \( C_M \). If the columns of \( H \) were linearly dependent, then \( C_L \) and \( C_M \) would be scalar multiples of each other which is not the case. Therefore it is concluded that \( H \) has full column rank of \( m \). Since the orthogonal transformation preserves the rank, \( \text{rank}\left( (H^\top \Phi_k) \times (\Phi_k^\top H) \right) \) is equal to \( \text{rank}(H^\top H) \) which is full-rank. Therefore \( M \) exists and hence the inverse to the coefficient matrix in (22) exists.

B. Pre-Processing

The pressure distribution from the PSP measurement contains missing data and in the datasets considered in the present study, they are found predominantly in regions near nose and tail of the fuselage; see Figure 17. These missing regions are then interpolated from the surrounding surface data using the voronoi kernel interpolation available in open-source visualization tool Paraview [35], although other techniques for interpolation may be used. Once the
missing data are imputed, we interpolate them onto a common grid to perform the data fusion. The common grid is created by coarsening the PSP measurement grid to keep the computational costs tractable. The final number of cells and points of the grids are shown in Table 4. The comparison of the computed coefficients of lift and moment for the original and coarse grids are shown in Table 5 where the summary statistic of the percent error across all the test cases is provided; a maximum error of 1.72% error confirms that the final grid is not too coarse. The Figure 18 and Figure 19 also show the contour plots of pressure distribution comparing the original grid and the coarse grid. Overall we observe that the grid coarsening has a negligible impact on the quantities of interest while using only ~10% of the information on the original grids which leads to improved computational efficiency.

|            | Num. of Cells | Num. of Points |
|------------|---------------|----------------|
| CFD        | 595642        | 533443         |
| PSP        | 84288         | 85711          |
| Coarse     | 8688          | 9103           |

**Table 4** Summary of grids geometry

|         | Mean   | Max   | Min   |
|---------|--------|-------|-------|
| CFD CL  | 0.53   | 0.74  | 0.37  |
| CFD CM  | 0.39   | 0.61  | 0.09  |
| PSP CL  | 1.07   | 1.72  | 0.02  |
| PSP CM  | 0.45   | 1.30  | 0.03  |

**Table 5** Absolute percent difference between original and coarse grids

![Fig. 17](image1.png)  
(a) Raw Data  
(b) Invalid Data Removed

**Fig. 17** PSP measurements containing invalid data points
Fig. 18  CFD pressure distribution on upper wing

Fig. 19  PSP pressure distribution on upper wing