Available online at www.eccomasproceedia.org
Eccomas Proceedia EUROGEN (2023) 73-84

ECCOMAS
Proceedia

EUROGEN 2023
15th ECCOMAS Thematic Conference on
Evolutionary and Deterministic Methods for Design, Optimization and Control
N. Gauger, K. Giannakoglou, M. Papadrakakis, J. Periaux (eds.)
Chania, Crete Greece, 1–3 June 2023

FUSING AERODYNAMIC DATA USING MULTI-FIDELITY GAUSSIAN PROCESS REGRESSION

Javier Nieto-Centenero[1,2], Rodrigo Castellanos[1], Alejandro Gorgues[1,3] and Esther Andrés[1]

[1] Theoretical and Computational Aerodynamics Branch, Flight Physics Department
Instituto Nacional de Técnica Aeroespacial (INTA)
Ctra. Ajalvir Km.4 Torrejón de Ardoz 28850, España
e-mail: {jniecen, rcasgar, gorguesva, eandper}@inta.es

[2] Escuela Técnica superior de Ingeniería Aeronáutica y del Espacio (ETSIAE)
Universidad Politécnica de Madrid (UPM)
28223 Madrid, España.

[3] Escuela Politécnica Superior
Universidad de Alcalá de Henares (UAH)
28805 Alcalá de Henares, España.

Abstract. *In aircraft aerodynamic design, it is common to have data from computational fluid dynamics simulations and wind-tunnel tests that provide datasets with different levels of fidelity. It is desirable to combine the strengths of both sources of information to generate models as close as possible to reality. In this paper, two multi-fidelity methods will be combined to model the pressure coefficient over a wing section in the transonic regime, namely Bayesian Gappy POD (BGPOD) and Multi-fidelity Gaussian Process Regression (MFGPR). Without the need for new datasets, the combined model improves the performance compared to the use of BGPOD in terms of both accuracy and uncertainty.*

Keywords: multi-fidelity analysis, Gaussian process, regression, aerodynamics, transonic wing, Gappy POD

Peer-review under responsibility of the organizing committee of EUROGEN 2023.
doi: 10.7712/140123.10191.18890

1 INTRODUCTION

Data from different sources is used to study aircraft aerodynamics. The three main sources of aerodynamic data are the following: Computational Fluid Mechanics (CFD), Wind Tunnel Test (WTT), and Flight Test Data [1]. Although all of these sources provide information on the same aircraft, their characteristics are different, including accuracy in showing the aerodynamic reality of the aircraft. Flight test data is considered to be the most representative reflection of real-world conditions, but it is usually limited in scope due to its high cost and safety issues. Moreover, these tests are performed in the final design phase of the aircraft, therefore, the data would not be useful for a preliminary design phase, although it would be useful for later analysis such as aircraft performance analysis. CFD simulations can be performed from very early design stages and may have different levels of accuracy depending on the used formulation. However, even with the advances in computing capacity of the last decades and those expected in the near future, it is not possible to perform a direct simulation of the Navier-Stokes Equations (DNS) at the industrial level, so the physics of the problem has to be simplified. Such model simplifications, combined with discretization errors and the complexity of simulating near flight envelope boundary conditions, resulting in limited accuracy [2]. Additionally, more accurate models lead to a higher computational cost, so a trade-off is necessary between the simulation's accuracy and the number of flight conditions to be studied to fit the available computational power. Finally, the accuracy of WTT data is intermediate between the CFD and the flight test data.

Aerodynamic data do not only differ in their accuracy to show reality, one of their most significant differences is the amount and dispersion of the available data. CFD gives a complete distribution of local pressure over the entire surface under study, while WTT has several local pressure measurements limited to the number of sensors on the surface [2], and flight tests have an even more limited number of measurements than WTT.

The combination of data from different acquisition methods and accuracy levels is achieved through *multi-fidelity* or *Data Fusion* method. The basic idea behind these techniques is to fuse data from different fidelities, which usually, as they increase in accuracy, decrease in the amount of available data [3], to generate a model as close as possible to the highest-fidelity data but with higher resolution. Within the field of aerodynamics, the most widespread Data-Fusion models are the Gappy Proper Orthogonal Decomposition (GPOD) and models based on Gaussian process regression (GPR). The principle behind the GPOD technique is to combine the Proper Orthogonal Decomposition (POD) modes with a least-squares problem to reconstruct a vector of incomplete data. The first use of GPOD was the reconstruction of human faces [4]. Two ways of using this technique are shown in this paper: the first consists of generating POD modes from a complete database and then reconstructing faces with sparse data from these modes; the second way is to reconstruct POD modes from a database in which all photos have randomly missing data. This method, with its two variants, was used in [5] to reconstruct the fluid field around a NACA0012 airfoil. It was also employed for the fusion of CFD and WTT data in [3], introducing a regularization of the least squares problem and a constrained method using aerodynamic forces. In [6], an extension of the constrained GPOD is presented, and the results are compared with a Bayesian-Data-Fusion framework. In [7] a Bayesian extension of this method is presented, in which the least-squares problem is solved by applying a Gaussian process regression, allowing the estimation of the degree of confidence in the regression. A variation of Gaussian process regression, known as Hierarchical Kriging, Cocriking or Multi-fidelity Gaussian process regression (MFGPR), is a popular modification of this method for Data-Fusion

applications. The method involves building a Gaussian process regression model for each level of fidelity. These regression models are then combined using a Bayesian framework to obtain a more accurate estimate of the output. The combination is done by assuming that the difference between the outputs of two adjacent levels of fidelity can be modeled as a Gaussian process with a certain covariance structure. This hypothesis allows the model to learn how to transfer information from one level of fidelity to another, and to propagate uncertainties from one level to the next. An application of cokriging to generate surrogate models for aerodynamic magnitudes by fusing data from different fidelities can be found in [8, 9, 10, 11]. It is also possible to fuse data from non-hierarchical sources of information, where each source of information receives a degree of confidence from experts in the problem or empirical data. In [12, 13] non-Hierarchical Kriging is implemented in aerodynamic applications.

This paper presents a data-fusion model that takes advantage of the strengths of BGPOD and MFGPR to predict pressure coefficients (C_p) on a wing section operating in the transonic regime. By incorporating MFGPR, the model is able to reduce the error in the C_p prediction and simultaneously narrow the confidence interval. This improvement in accuracy is accomplished without the need of providing supplementary databases.

2 METHODOLOGY

In this section, the database and methodology used are presented. First, a description of the aerodynamic database is provided 2.1. A mathematical introduction of the employed methods follows. Gappy POD is described in 2.2, followed by a description of Gaussian process regression in 2.3 and how it relates to Gappy POD. Finally, a brief review of Multi-fidelity Gaussian process regression and its use for the study case is presented in 2.4.

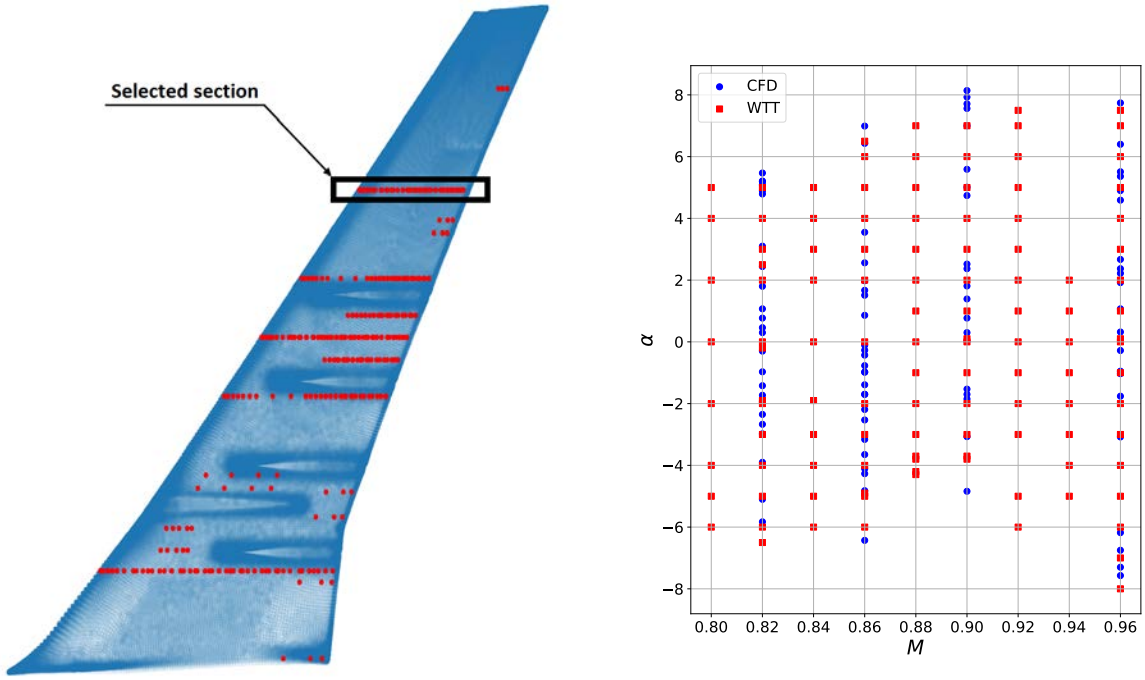
2.1 Database

A database of CFD simulations and WTT of the wing of the XRF1 research aircraft is used. The XRF1 [14] is a research aircraft model provided by Airbus, which is representative of a long-range wide-body aircraft. The work presented here has been carried out within the framework of the Group for Aeronautical Research and Technology in Europe (GARTEUR) for the AD/AG60 [15] research project. CFD was performed by means of Reynolds Average Navier-Stokes (RANS) simulations, made with the TAU solver [16]. The entire aircraft was considered, so the aerodynamics interactions between different subsystems of the aircraft exposed to the air stream were captured. WTT were performed at the European Transonic Wind Tunnel (ETW) facility, and surface pressure data was acquired using pressure taps located in 26 spanwise locations along the wing span, η. The Reynolds for both CFD simulations and WTT was set to $Re = 25 \times 10^6$.

The test case presented in this study employs a subset of the above-mentioned database. One of the airfoils was isolated sufficiently far from the wing pods, located at $\eta = 0.75$, to avoid significant aerodynamic interference. This subset contains a significant population of high-fidelity data collected in the WTT. The data within the selected section consist of 199 (C_p) values obtained from CFD simulations and 59 C_p values obtained from pressure taps. The WTT dataset is then subdivided into training (86.5%) and test (13.5%) sets, the latter being approximately distributed at 20, 40 and 60% of the airfoil chord.

Finally, the CFD simulation database is made up of a total number of 89 flight conditions with Mach numbers $M \in [0.82, 0.96]$ and angles of attack $\alpha \in [-7.5°, 8°]$ and the WTT database is composed by a total number of 114 flight conditions with Mach numbers $M \in [0.8, 0.96]$ and

angles of attack $\alpha \in [-8°, 7.5°]$, as shown in the Figure 1.



(a) CFD grid points (blue) and pressure taps (red)

(b) Flight conditions

Figure 1: Subfigure (a) shows the distribution of CFD grid points and the pressure taps of the WTT. The selected section for this work is also indicated. Subfigure (b) shows the distribution of flight conditions for CFD simulations and WTT, in terms of Mach number (M) and angle of attack (α).

2.2 Gappy POD

POD is a data-driven method that aims to find an approximation of the input data in a low-dimensional space, preserving the essential information of the high-dimensional dataset [17]. For this particular case, the high-dimensional dataset is composed of the vector C_p obtained by CFD simulations, $\mathbf{y}_i \in \mathbb{R}^P$ which is stored in the snapshot matrix $\mathbf{Y} \in \mathbb{R}^{P \times N}$, where P denote the number of grid points and N are the number of flight conditions that are being using as pseudo-time [18]. Singular Value Decomposition (SVD) is applied to the snapshot matrix $\mathbf{Y}$ to obtain the matrix decomposition that follows,

$$\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* = \sum_{i=1}^{d} \sigma_i \mathbf{u}_i \mathbf{v}_i^* \,, \tag{1}$$

where $\mathbf{U} = [\mathbf{u_1}, \mathbf{u_2}, \ldots, \mathbf{u_d}] \in \mathbb{R}^{P \times d}$ and $\mathbf{V} = [\mathbf{v_1}, \mathbf{v_2}, \ldots, \mathbf{v_d}] \in \mathbb{R}^{N \times d}$ are orthogonal semi-unitary matrices such that $\mathbf{U}^*\mathbf{U} = \mathbf{V}^*\mathbf{V} = \mathbf{I}_d$, and $\mathbf{\Sigma} = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d) \in \mathbb{R}^{d \times d}$ and $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d > 0$. For the database proposed in this study, d coincides with the number of simulated flight conditions N, which is the dimension of the low-dimensional basis given by $\mathbf{U}$. The truncation of the number of modes is also possible by choosing a reduced rank $d^* < d$ so that only the d^* most energetic modes are conserved upon reconstruction.

The low-rank approximation provided by POD allows reconstructing the solution $\mathbf{u} \in \mathbb{R}^P$ with r measurements of the $P-$dimensional state [19]. For this particular case, the r measurements are the C_p of the pressure sensors recordings from the WTT, with $r << P$. This sparse vector can be represented by the variable $\tilde{\mathbf{u}}$, being it defined as,

$$\tilde{\mathbf{u}} = \mathbf{P^T u} \, , \tag{2}$$

where $\mathbf{P} \in \mathbb{R}^{P \times r}$ is a mask that takes the unity value at locations where there is a sensor and zero elsewhere. This sparse vector can then be approximated with the standard POD projection:

$$\tilde{\mathbf{u}} \approx \, \mathbf{P^T} \sum_{\mathbf{k=1}}^{\mathbf{d}} \tilde{\mathbf{a}}_\mathbf{k} \psi_\mathbf{k} = \, \mathbf{P^T U_d} \tilde{\mathbf{a}} \, , \tag{3}$$

where $\mathbf{U_d} \in \mathbb{R}^{P \times d}$ is the left-mode matrix of SVD decomposition and $\tilde{\mathbf{a}} \in \mathbb{R}^d$ is the vector of coefficients that minimizes the error in the approximation $\|\tilde{\mathbf{u}} - \mathbf{P^T u}\|$ in the L_2 sense. If $\mathcal{X} = \mathbf{P^T U_d} \in \mathbb{R}^{r \times d}$ has a full column rank, then this minimization has as its solution the ordinary least-squares estimator,

$$\tilde{\mathbf{a}} = (\mathcal{X}^\mathbf{T} \mathcal{X})^{-\mathbf{1}} \mathcal{X}^\mathbf{T} \tilde{\mathbf{u}} \, . \tag{4}$$

The problem of a near-singular moment matrix on $(\mathcal{X}^\mathbf{T} \mathcal{X})$ could be alleviated by adding a small positive constant λ to the diagonal before taking the inverse. Applying this modification to equation (4), the ridge estimator is obtained,

$$\tilde{\mathbf{a}} = (\mathcal{X}^\mathbf{T} \mathcal{X} + \, \lambda \mathbf{I})^{-\mathbf{1}} \mathcal{X}^\mathbf{T} \tilde{\mathbf{u}} \, . \tag{5}$$

With the coefficient vector $\tilde{\mathbf{a}}$ determined, the reconstructed vector $\mathbf{u}$ can be obtained by applying equation (6),

$$\mathbf{u} \approx \mathbf{U_d} \tilde{\mathbf{a}} \, . \tag{6}$$

2.3 Gaussian process regression

An overview of Gaussian process regression theory is herein presented. For a detailed analysis of the outlined mathematics, the interested readers are encouraged to consult [20, 21, 22].

A Gaussian process is a collection of random variables, any finite number of which have joint Gaussian distributions. A Gaussian process is specified by its mean function, $m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$, and covariance function, $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$, then we can write the Gaussian process as $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$. Although it is not required, the mean function usually takes a zero value to simplify the notation. Each element of the training dataset, $\mathbf{y}$, is a sample with Gaussian distribution, representing the true value of the observation, $f(\mathbf{x})$, affected by some independent Gaussian noise, ϵ, with variance, σ_n. Thus, the observations can be interpreted as $\mathbf{y} = f(\mathbf{x}) + \epsilon$. The objective of the regression is to predict $\mathbf{f}_*$ values at new points $\mathbf{x}_*$. The joint distribution of the training values and the function at new points is

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I} & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right) \, , \tag{7}$$

where $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_n}]$ are the observed datapoints, $\mathbf{X}_* = [\mathbf{x_{1*}}, \ldots, \mathbf{x_{n*}}]$ are the new points where to make predictions, and $\mathbf{K}$ are matrices constructed using any function $k(\mathbf{x}, \mathbf{x}')$ that can perform as a covariance function, that is, any function that takes two arguments, such that

$k(\mathbf{x}, \mathbf{x}')$ generates a non-negative definitive covariance matrix $\mathbf{K}$. These functions are known as kernel functions.

By deriving the conditional distribution, we arrive at the predictive equations for the Gaussian process regression as $\bar{\mathbf{f}}_*|\mathbf{X}, \mathbf{y}, \mathbf{X}_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*))$ where

$$\bar{\mathbf{f}}_* = K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right]^{-1} \mathbf{y} \ , \tag{8}$$

$$\text{cov}(\mathbf{f}_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) \left[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathbf{I}\right]^{-1} K(\mathbf{X}, \mathbf{X}_*) \ . \tag{9}$$

The method used for learning the noise variance and, if there were, the kernel hyperparameters, is the maximization of the log marginal likelihood given by

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^\mathsf{T}[K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}]^{-1}\mathbf{y} - \frac{1}{2}\log|K(\mathbf{X}, \mathbf{X}) + \sigma^2 \mathbf{I}| - \frac{n}{2}\log(2\pi) \ . \tag{10}$$

In [7] the Bayesian Gappy POD (BGPOD) extension is presented. This method uses Gaussian process regression to estimate the coefficients $\tilde{\mathbf{a}}$ of the Gappy POD method and then obtain a predictive distribution for all rows of the left-mode POD matrix. That extension is used in this study; however, the kernel function proposed here is different from the one used in the mentioned paper. The mean equation (8) is reminiscent of the Ridge estimator, equation (5). In fact, the mean of a GPR is equal to that given by the ridge regression, if the kernel function is the dot product kernel, $k(\mathbf{x}, \mathbf{x}') = \sigma_0^2 + \mathbf{x} \cdot \mathbf{x}'$, with $\sigma_0^2 = 0$, also called the homogeneous linear kernel. This kernel function is used in this study to have a direct relationship between the regularized GPOD and the BGPOD, but in the latter case providing the confidence intervals in the regression. In addition, the homogeneous linear kernel is useful if the original features are individually informative, so the decision boundary is likely to be representable as a linear combination of the original features [23], as is the case for POD modes. Figure 2 depicts a comparison between solving the GPOD using a GPR with the dot product kernel or the Ridge estimator.

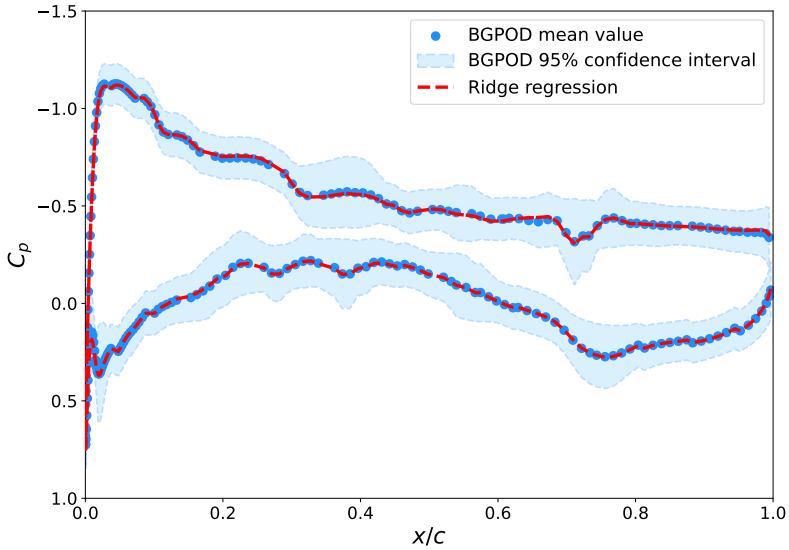


Figure 2: Comparison between BGPOD and GPOD solved with the Ridge estimator.

2.4 Multi-fidelity Gaussian process regression

Gaussian process regression can be extended to construct probabilistic models that allows the combination of variable fidelity information sources [24, 25, 26]. If there are s levels of information that produce an output $y_t(\mathbf{x}_t)$. These data can be organized with increasing fidelity as $D_t = \{\mathbf{x_t}, \mathbf{y_t}\}, t = 1, ..., s$. Assuming the Markov property introduced in [24], $\mathrm{Cov}(f_t(\mathbf{x}), f_{t-1}(\mathbf{x}')|f_{t-1}(\mathbf{x})) = 0, \forall \mathbf{x} \neq \mathbf{x}'$, that means no further information can be acquired about $f_t(\mathbf{x})$ from lower fidelities $f_{t-1}(\mathbf{x}')$ for $\mathbf{x}' \neq \mathbf{x}$, we lead to the autoregressive model,

$$f_t(\boldsymbol{x}) = \rho_t f_{t-1}(\boldsymbol{x}) + \delta_t(\boldsymbol{x}) \ , \tag{11}$$

where $\delta_t(\mathbf{x})$ is a Gaussian process independent of $f_{t-1}(\mathbf{x}), \ldots, f_0(\mathbf{x})$ with mean μ_{δ_t} and covariance function k_{δ_t}, and ρ_t represents a scale factor between $f_t(\mathbf{x})$ and $f_{t-1}(\mathbf{x})$.

A numerically efficient recursive inference scheme can be constructed by adopting the derivation proposed by Le Gratiet & Garnier [25]. With this scheme, the inference problem is essentially decoupled into s standar GPR problems, yielding the multi-fidelity posterior distribution $\bar{\mathbf{f}}_\mathbf{t}|\mathbf{y_t}, \mathbf{X}_t, \mathbf{f}_{*\mathbf{t-1}}, \ t = 1, \ldots, s$, with predictive mean and variance at each level given by

$$\bar{\mathbf{f}}_{*\mathbf{t}} = \rho_t \bar{\mathbf{f}}_{*\mathbf{t-1}} + \mu_{\delta_t} + K\left(\mathbf{X}_*, \mathbf{X_t}\right)\left[K(\mathbf{X_t}, \mathbf{X_t}) + \sigma_{nt}^2 \mathbf{I}\right]^{-1}\left[\mathbf{y_t} - \rho_t \bar{\mathbf{f}}_{*\mathbf{t-1}} - \mu_{\delta_t}\right] \ , \tag{12}$$

$$\mathrm{cov}(\mathbf{f}_{*\mathbf{t}}) = \rho_t^2 \,\mathrm{cov}(\mathbf{f}_{*\mathbf{t-1}}) + K\left(\mathbf{X}_*, \mathbf{X}_*\right) - K\left(\mathbf{X}_*, \mathbf{X_t}\right)\left[K(\mathbf{X_t}, \mathbf{X_t}) + \sigma_{nt}^2 \mathbf{I}\right]^{-1} K\left(\mathbf{X_t}, \mathbf{X}_*\right) \ . \tag{13}$$

The kernel used for high and low fidelity databases is the Radial Basis Fuction (RBF), which is the most widely used kernel [27], and reads as follows

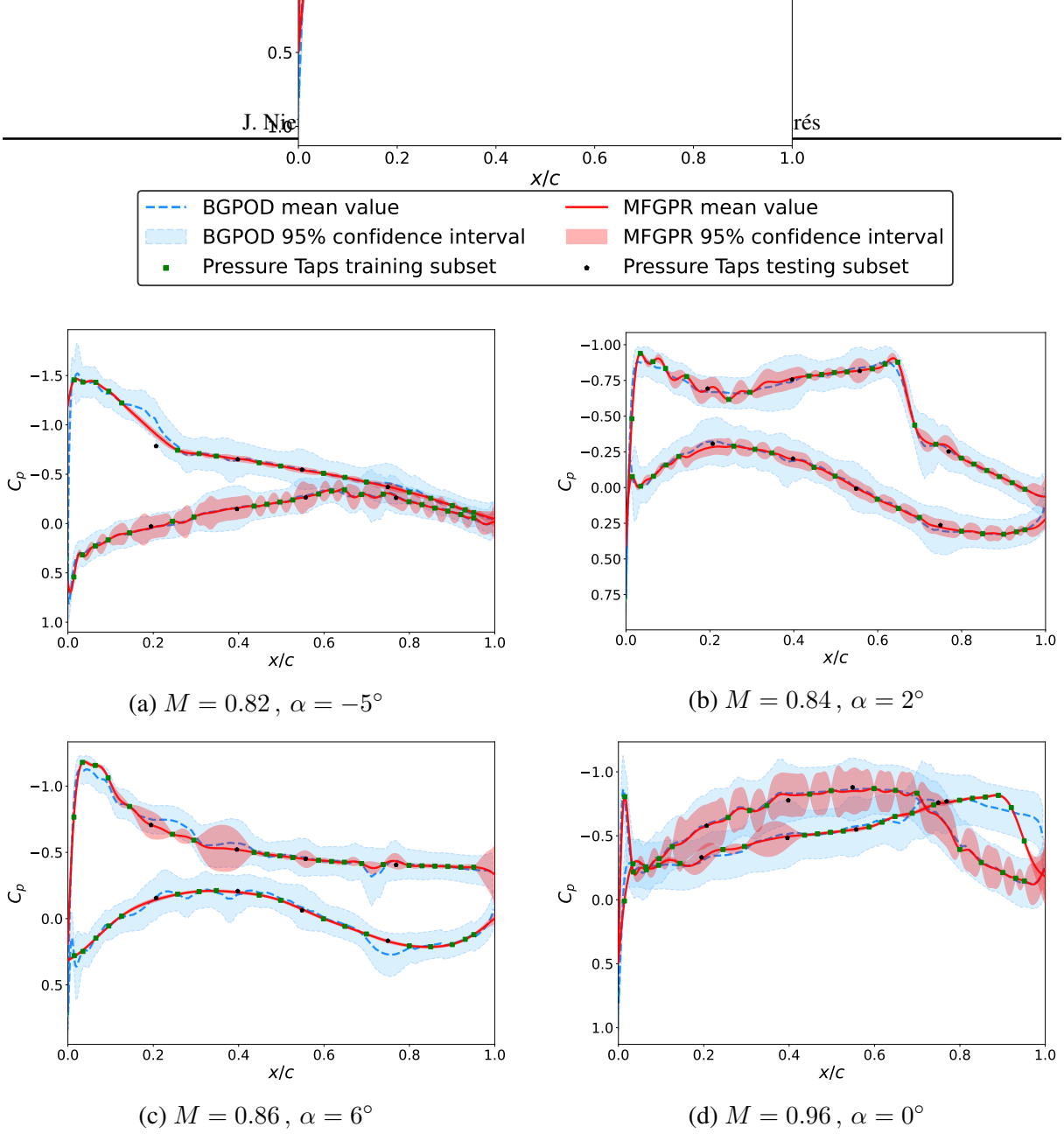
$$k\left(\mathbf{x}, \mathbf{x}'\right) = \sigma_k^2 \exp\left(-\frac{\left(\mathbf{x} - \mathbf{x}'\right)^2}{2l^2}\right), \tag{14}$$

where σ_k^2 , $l \in \mathbb{R}$ are the free parameters of the kernel. The process followed to obtain the hyperparameters of the models is the maximization of the log-likelihood [28]. Generally, the log-likelihood is not convex. Therefore, an exploration is performed starting from different initialization values. Once the exploration is done, we select the parameters from a minimum that does not lead to an overfitted model.

3 RESULTS

Gappy POD and Multi-fidelity Gaussian process regression are the two most widely used multi-fidelity models for aerodynamic applications. In this work, we intend to combine both techniques. For this purpose, the autoregressive model in equation (11), with two fidelities is used. The higher fidelity is the C_p data obtained in WTT. Furthermore, for this fidelity the noise variance will be assumed to be zero, so an interpolation between the data would be performed. The choice for the noise to be zero is given by the unknown measurement tolerance of the pressure taps, which could be introduced directly if known and would lead to a model closer to reality. Sometimes this error variance is taken as a hyperparameter in the optimization of the GPR; however, this methodology requires several repetitions of the observation with the same test conditions [2] in order to be feasible and provide valid results.

For low-fidelity data, the results obtained by applying the BGPOD on the database will be used. However, the resultant Gaussian process is not fed directly into the multi-fidelity model.



(a) $M = 0.82\,,\ \alpha = -5°$

(b) $M = 0.84\,,\ \alpha = 2°$

(c) $M = 0.86\,,\ \alpha = 6°$

(d) $M = 0.96\,,\ \alpha = 0°$

Figure 3: Comparison of the pressure coefficient reconstruction achieved by Bayesian Gappy POD (blue) and Multi-fidelity Gaussian process regression (red).

This is because the BGPOD regression is performed on the POD modes and not on the C_p data, so this Gaussian process will not have a direct relationship with the high-fidelity data. The solution proposed here is to use as low-fidelity data the mean C_p obtained from the BGPOD. In addition, this method provided information about the variance of these values. It can be seen that for our test case, the variance of the BGPOD result is practically constant throughout the whole prediction. With this idea in mind, the variance of the noise for low-fidelity will be selected equal to the mean of the predicted variances. In a case where the variance would be very different by zones, the Heteroscedastic Gaussian process model could be used, where the observed noise can have a different variance at each point [29]; nonetheless, this model is beyond the scope of this study.

Figure 3 depicts the results for four flight conditions in the transonic regime. Figure 3a shows the case $M = 0.82, \alpha = -5°$. As previously discussed, MFGPR intercepts all pressure tap training points since it takes this data as ground truth. In this case, both methods are able to follow the general trend of the C_p distribution. The variance of the multi-fidelity model is lower than that of BGPOD, showing a clear difference between the behavior of the lower surface,

where the variance narrows between the train data, and the upper surface, where the variance grows considerably between these points. It should be noted that there is a C_p value of the pressure taps testing subset that falls outside the confidence interval of the two models. Figure 3b shows the case $M = 0.84, \alpha = 2°$. The BGPOD underestimates the value of C_p in the suction peak area, and slight oscillations occur downstream with lack of physical coherence. With MFGPR these oscillations are notably increased, thus misrepresenting this region. The shock wave is well characterized in both cases. Figure 3c shows the case $M = 0.86, \alpha = 6°$. The MFGPR model is able to correctly follow the C_p trend, making an accurate prediction of the test points, and smoothing the oscillations observed with the BGPOD model. However, it can be observed that the MFGPR model is unable to capture the C_p at the attachment line. Finally, Figure 3d shows the case $M = 0.96, \alpha = 0°$. In this case, it can be seen how the BGPOD cannot accurately capture the shock wave present near the trailing edge of the upper surface. Moreover, the uncertainty margin is much larger than for the cases shown earlier. The mean of the MFGPR model is capable of accurately describing the aerodynamic characteristics of this case. It is also observed that, while for the upper surface there are narrow confidence margins, for the lower surface case the confidence margins between train points are high, and the mean of this model follows the same trend as that of the BGPOD.

In light of the MFGPR results, it is possible to notice the complexity in the choice of hyperparameters for each case, which, even following the same optimization method, lead to results that differ notably in the distribution of the variance or the smoothness. Research of a robust method to select the hyperparameter set is one of the major challenges within Gaussian process regressors.

To observe the benefits of using a multi-fidelity model, a GPR of high-fidelity data has been performed directly (HFGPR). Figure 4 shows the case $M = 0.92, \alpha = 4°$. It can be seen that the 95% confidence interval of the interpolation is very wide in areas where no train data is available and how the predicted mean deviates significantly from the test data. Furthermore, the model does not follow the flow trend at the leading and trailing edges due to insufficient information in these regions.

Finally, a quantitative analysis of the errors in the test dataset is performed. For this purpose, a comparison is made between three different models: the BGPOD, the MFGPR and the HFGPR, evaluating the Root Mean Squared Error (RMSE), to evaluate the error of the test data with respect to the GPR mean, and the Mean Log-Loss error (MLL), which not only takes into account the distance between the test data and the prediction, but also the confidence intervals of the prediction are considered. The lower the MLL, the better the model [20]. In Table 1 it can be seen that the RMSE values by the HFGPR is significantly higher compared to the other two methods, having also an elevated MLL, indicating a poor fit of the model to the test data. On the other hand, the MFGPR has an RMSE 11% lower than that of the BGPOD and the MLL demonstrates that it is a more suitable model for the data studied.

Table 1: Results of the performance metrics applied in the BGPOD, MFGPR and HFGPR study models. The root mean square error, RMSE, and the mean log-loss, MLL, are applied on the test points for all flight conditions.

	BGPOD	**MFGPR**	**HFGPR**
RMSE	0.048	0.043	0.162
MLL	-1.28	-1.42	4.750

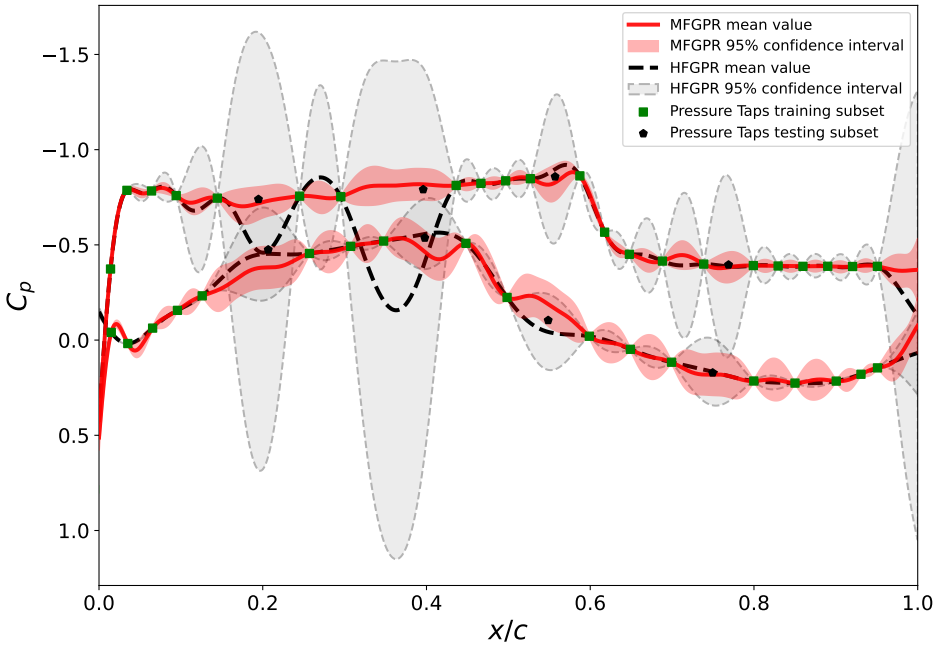


Figure 4: Comparison between MFGPR and HFGPR for the flight condition $M = 0.92, \alpha = 4°$.

4 CONCLUSIONS

This work explores the combination of Bayesian Gappy Proper Orthogonal Decomposition (BGPOD) with Multi-fidelity Gaussian Process Regression (MFGPR) to develop a surrogate model for predicting pressure coefficients over a wing section in the transonic regime. The model fuses data derived from RANS simulations and pressure tap measurements obtained in wind-tunnel experiments, both performed on the XRF1 research aircraft. The results obtained with the proposed model show a reduction of 11% RMSE in test points with respect to the use of BGPOD. MFGPR also decreases the uncertainty of the C_p forecast and provides the ability to obtain the mean and variance of any desired point, not limiting the prediction to CFD grid points, as is the case with BGPOD. Therefore, this improvement in regression has been achieved without incorporating new training data.

Both the choice of the kernels and the hyperparameters tuning present a major problem that requires an in-depth study since the model is strongly dependent on them. Finding a suitable combination of these elements for the problem at hand could alleviate the slight oscillations that appear in the C_p prediction and improve the overall accuracy of the model. It is also desirable to incorporate the real uncertainty of the input data to obtain a model that is more faithful to the nature of the data, which could be introduced directly into the developed model.

ACKNOWLEDGEMENTS

The authors would like to thank Airbus for providing the database for the XRF1 test case.

REFERENCES

[1] Mehdi Anhichem, Sebastian Timme, Jony Castagna, Andrew Peace, and Moira Maina. Multifidelity data fusion applied to aircraft wing pressure distribution. In *AIAA AVIATION 2022 Forum*, page 3526, 2022.

[2] Rubén Conde Arenzana, Andrés F López-Lopera, Sylvain Mouton, Nathalie Bartoli, and Thierry Lefebvre. Multi-fidelity gaussian process model for cfd and wind tunnel data fusion. In *AeroBest 2021*, 2021.

[3] Michael Mifsud, Alexander Vendl, Lars-Uwe Hansen, and Stefan Görtz. Fusing wind-tunnel measurements and cfd data using constrained gappy proper orthogonal decomposition. *Aerospace Science and Technology*, 86:312–326, 2019.

[4] Richard Everson and Lawrence Sirovich. Karhunen–loeve procedure for gappy data. *JOSA A*, 12(8):1657–1664, 1995.

[5] Tan Bui-Thanh, Murali Damodaran, and Karen Willcox. Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition. *AIAA journal*, 42(8):1505–1516, 2004.

[6] S Ashwin Renganathan, Kohei Harada, and Dimitri N Mavris. Aerodynamic data fusion toward the digital twin paradigm. *AIAA Journal*, 58(9):3902–3918, 2020.

[7] Anna Bertram, Philipp Bekemeyer, and Matthias Held. Fusing distributed aerodynamic data using bayesian gappy proper orthogonal decomposition. In *AIAA Aviation 2021 Forum*, page 2602, 2021.

[8] Zhong-Hua Han and Stefan Görtz. Hierarchical kriging model for variable-fidelity surrogate modeling. *AIAA journal*, 50(9):1885–1896, 2012.

[9] Alexander IJ Forrester, Neil W Bressloff, and Andy J Keane. Optimization using surrogate models and partially converged computational fluid dynamics simulations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 462(2071):2177–2204, 2006.

[10] Alexander IJ Forrester, András Sóbester, and Andy J Keane. Multi-fidelity optimization via surrogate modelling. *Proceedings of the royal society a: mathematical, physical and engineering sciences*, 463(2088):3251–3269, 2007.

[11] Yuichi Kuya, Kenji Takeda, Xin Zhang, and Alexander IJ Forrester. Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. *AIAA journal*, 49(2):289–298, 2011.

[12] Rémi Lam, Douglas L Allaire, and Karen E Willcox. Multifidelity optimization using statistical surrogate modeling for non-hierarchical information sources. In *56th AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, page 0143, 2015.

[13] Alex Feldstein, David Lazzara, Norman Princen, and Karen Willcox. Multifidelity data fusion: Application to blended-wing-body multidisciplinary analysis under uncertainty. *AIAA Journal*, 58(2):889–906, 2020.

[14] Norbert Kroll, Mohammad Abu-Zurayk, Diliana Dimitrov, Thomas Franz, Tanja Führer, Thomas Gerhold, Stefan Görtz, Ralf Heinrich, Caslav Ilic, Jonas Jepsen, et al. Dlr project digital-x: towards virtual aircraft design and flight testing based on high-fidelity methods. *CEAS Aeronautical Journal*, 7:3–27, 2016.

[15] GARTEUR AD/AG-60 Machine learning and data-driven approaches for aerodynamic analysis and uncertainty quantification. `https://garteur.org/`. Accessed: 2023-03-08.

[16] Norbert Kroll, Stefan Langer, and Axel Schwöppe. The dlr flow solver tau-status and recent algorithmic developments. In *52nd Aerospace Sciences Meeting*, page 0080, 2014.

[17] Lawrence Sirovich. Turbulence and the dynamics of coherent structures. i. coherent structures. *Quarterly of applied mathematics*, 45(3):561–571, 1987.

[18] Rodrigo Castellanos, Jaime Bowen Varela, Alejandro Gorgues, and Esther Andrés. An assessment of reduced-order and machine learning models for steady transonic flow prediction on wings.

[19] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

[20] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[21] Robert B Gramacy. *Surrogates: Gaussian process modeling, design, and optimization for the applied sciences*. Chapman and Hall/CRC, 2020.

[22] José Melo. Gaussian processes for regression: a tutorial. *Technical Report*, 2012.

[23] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[24] Marc C Kennedy and Anthony O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[25] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.

[26] Paris Perdikaris, Maziar Raissi, Andreas Damianou, Neil D Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

[27] Martin Klapacz. Multifidelity gaussian processes for uncertainty quantification. 2021.

[28] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[29] Miguel Lázaro-Gredilla and Michalis K Titsias. Variational heteroscedastic gaussian process regression. In *ICML*, pages 841–848, 2011.