10.2514/6.2014-2301

AIAA AVIATION Forum
16-20 June 2014, Atlanta, GA
15th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference

Surrogate Models and Mixtures of Experts in Aerodynamic Performance Prediction for Mission Analysis

Rhea P. Liem*
University of Toronto Institute for Aerospace Studies, Toronto, ON, Canada
Joaquim R. R. A. Martins†
University of Michigan, Department of Aerospace Engineering, Ann Arbor, MI, USA

Accurate aircraft fuel burn evaluation requires performing a detailed mission analysis covering the entire mission, from takeoff to landing. This process is computationally expensive, as it requires up to millions of aerodynamic performance evaluations, and thus it is advantageous to use surrogate models as approximations of the actual aerodynamic models. Training surrogate models is challenging due to the high nonlinearity of the aerodynamic performance functions in the transonic regime. Conventional surrogate models, such as radial basis function and kriging, are deemed insufficient to model these functions accurately. To address this issue, we explore several ways to improve the predictive performance of surrogate models. First, we employ an adaptive sampling algorithm in addition to the more traditional space-filling algorithm. Second, we improve the kriging performance by including gradient information in the interpolation (gradient-enhanced kriging), as well as by introducing a known trend in the global model component (kriging with a trend). Lastly, we propose a mixture of experts approach, which is derived based on the divide-and-conquer principle. In this last approach, we use multiple surrogate models as local experts to approximate different parts of the input space, using machine learning techniques to infer about the function profile to automatically partition the input space. These various surrogate models are tested using aerodynamic data for conventional and unconventional aircraft configurations. We then perform a surrogate-based mission analysis using the selected surrogate models. Our results show that the proposed mixture of experts approach can significantly improve the predictive performance when approximating the aerodynamic performance. For example, a mixture of five gradient-enhanced kriging models (with adaptive sampling) achieves 5% approximation error with around 100 samples, whereas the adaptive sampling fails to converge when training a global model. However, when we have a simple function profile, using a global model is more efficient than a mixture of experts, due to the added computational complexity in the latter.

I. Introduction

Fuel efficiency and fuel economy have increasingly become the key drivers in aircraft performance evaluation and design process [1, 2]. Evaluating aircraft fuel burn accurately is not an easy task, considering the complex physics involved in aircraft operation. Moreover, other factors, such as atmospherics conditions and engine performance, contribute to the evaluation as well, making the computation even more complex. Such a detailed computation, when performed in an optimization process (which requires many iterations, sometimes hundreds, prior to reaching optimality), can be computationally intractable. Some common approaches typically involve simplification of physics or the mission profile. The classical Brequet range equation is a popular example of such an approach [3, 4, 5]. Other simplified models include using fuel fraction [5], analytical, and empirical models [6]. These simplifications and assumptions reduce the computational time, albeit at the expense of accuracy.

Recent work has shown that surrogate models can significantly reduce the required computational cost to perform detailed fuel burn computation in an optimization problem setting. Surrogate models, or metamodels, are commonly used as simpler approximations of the physical systems to reduce the cost of computationally intensive analysis and optimization tasks [7, 8, 9]. Surrogate models have previously been shown to assist various optimization procedures in aerospace engineering. Chung and Alonso [10, 11] used a gradient-enhanced kriging method in a supersonic business jet design optimization, Toal and Keane [12] used a cokriging method to perform a multipoint drag minimization, Zimmermann and Görtz [13] developed and used a POD-subspace restricted least squares model for solving the governing fluid flow equations, and Amsallem *et al.* [14] performed offline precomputations to construct fluid reduced order bases (ROB) and structural reduced order models (ROM) database for aeroelastic computations. In the context of mission analysis, Koko [15] used a Lagrangian interpolation as a surrogate to model the aerodynamic forces

*Ph.D. Candidate, AIAA Student Member
†Associate Professor, AIAA Associate Fellow

at different points along the flight mission of interest in a trajectory optimization problem aiming to minimize fuel consumption of morphing wingtip devices. The authors have previously used kriging models to approximate the aerodynamic data required in a detailed mission analysis procedure, to give an accurate estimation of the amount of fuel burned during a mission. This approach significantly reduces the required number of aerodynamic performance evaluations from millions to the number of samples required to build the kriging models, thus enabling the integration of mission analysis in aerostructural optimization cases. Using this surrogate-based mission analysis procedure, a new strategy was derived to formulate multi-point design optimization problems to maximize the aircraft performance over a large number of different missions [16]. This strategy was demonstrated in a fuel burn minimization problem for a long-range wide-body aircraft configuration, where only the cruise portion was modeled in detail. A similar approach was demonstrated in a direct operating cost (DOC) minimization problem for a 100-passenger regional jet configurations [17]. In the latter work, a shorter range mission was considered, in which the cruise portion was no longer the only dominant mission segment. The contribution from other segments to the amount of fuel burned, in particular the climb segments, was as significant and therefore needed to be considered. The surrogate model training was consequently more challenging when the input space was larger, since it needed to model the high drag gradient region outside the cruise regime. Further challenges may arise when we consider other unconventional configurations. For example, a blended wing body (BWB) configuration exhibits more correlation between drag and trim, causing more nonlinearity in drag profile with respect to the tail angle dimension [18].

Before we can perform more complicated surrogate-based mission analyses in optimizations, it is imperative to have reliable surrogate models that can approximate the aircraft performance over the entire flight operating regime (from takeoff to landing) of the various mission profiles considered. The modeling techniques are ideally flexible to be used with different aircraft configurations (conventional and unconventional), for both short and long range missions. Improving the accuracy of surrogate models, however, is not straightforward. Adding more samples as training data is a classical way to improve the surrogate modeling accuracy in sample-based surrogate models such as regression models, kriging, and radial basis function (RBF). However, more training samples means more function evaluations using the actual model, which can be computationally expensive especially when high-fidelity models are used. When new surrogate models need to be constructed at each optimization iteration, the computational cost can become prohibitive. In addition, care must be taken as adding too many samples can lead to *overfitting* in regression models [19], and increase the computational burden in constructing interpolation models such as kriging (as the size of the linear system of equations grows) [20, 21]. Thus, we need to find the right balance between accuracy and efficiency. Also, since the mission analysis is to be used in aerostructural optimizations using gradient-based optimizers (due to the large number of design variables and constraints considered), it is critical to have continuous surrogate models.

In this work, we explore and analyze the performance of various surrogate models in the context of performing surrogate-based mission analysis. Based on our specific requirements, which will be discussed later, we select some techniques that are suitable for our purpose, which narrows down to kriging and RBF models. Some variants of kriging models will also be considered here, including primarily those that allow incorporating some "extra knowledge" to further fine tune the models to follow the actual profile. The first is the gradient-enhanced kriging (GEK) model, which incorporates gradient information at sample points so the surrogate model can have better approximations of the curvature around the sample points, and thus better function approximation overall. GEK is a very well-established technique and has been shown to improve kriging performance; see, [10, 11, 21, 22, 23], for example, for some aerospace applications of GEK. Second, we have the "kriging with a trend" model, where we specify the basis functions for the global model of kriging [24]. Instead of using the commonly used low-order polynomials, we select the basis functions based on the system physics, e.g., by setting a quadratic trend in a certain direction. This second approach has been demonstrated in a previous work by the authors [17]. Third, we propose to use multiple surrogate models in the input space, instead of just a single global model. The main rationale is that we let each local surrogate model to perform well in a smaller subset of the input space, instead of forcing one model to approximate the entire terrain, which might have contrasting profiles across the input space (e.g., when it exhibits high nonlinearity only in a certain part of the input space). The first important question to answer is thus, how do we partition the input space? We consider using machine learning algorithms to "learn" the input space based on the available training data. We will use and compare the performance of these surrogate models in approximating the aerodynamic lift and moment coefficients of two aircraft configurations, conventional and unconventional. A surrogate-based mission analysis will then be performed using the selected models as a demonstration.

We start the remainder of this paper by describing the surrogate-based mission analysis procedure in Section II. In Section III, we first discuss the surrogate modeling classification, to select the techniques that are suitable for our purpose. We then explain the details of the selected techniques, namely kriging and RBF models, and their comparison. Our proposed *mixture of experts* model is presented in Section IV. The description of our case studies are given in Section V. We then discuss our results and findings in Section VI, followed by the conclusion in Section VII.

II. Surrogate-based Mission Analysis

The classical Breguet range equation is commonly used to compute the amount of fuel burn during flight [3, 4, 5]. This widely used range equation was derived and published independently in 1920 by Coffin [25] and later in 1923 by Breguet [26]. This equation has since become a basic model describing the physics of aircraft, encompassing the three dominant disciplines within an aircraft system: engine (by the thrust specific fuel consumption or TSFC), aerodynamics (by the lift to drag ratio, L/D), and structural technologies (by the structural weight). This equation, however, is only applicable under the assumption that TSFC, L/D, and flight speed are constant. One important implication is that the takeoff, climb, and descent segments are not properly modeled by this equation [27]. Simple fuel fractions (the ratio of the aircraft total weight at the end of a flight segment to the weight at the start of the same segment) are typically used to compute the amount of fuel burned in flight segments other than cruise. See, for example, Roskam [5] for values of suggested fuel-fractions corresponding to several mission phases for various aircraft types. Lee and Chatterji [28] presented the approximation functions for total fuel burn in climb, cruise, and descent phases. To compute fuel burn during climb, they applied a climb fuel increment factor, which was defined as the additional fuel required to climb the same distance as it was for cruise, normalized with respect to the takeoff weight [29]. Henderson *et al.* [30] presented an object-oriented aircraft conceptual design toolbox, `pyACDT`, which analyzed a given mission profile to estimate the mission fuel burn and point performance parameters. The Breguet range equation was used to calculate the cruise range. This toolbox uses a potential flow panel method for its aerodynamic module. The Program for Aircraft Synthesis Studies (PASS), created by Desktop Aeronautics, Inc., is a conceptual design tool which evaluates all aspects of mission performance [31]. This software package can incorporate several analyses, including linear aerodynamic models for lift and inviscid drag, sonic boom prediction for supersonic cases, weight and center of gravity estimation, and full mission analysis. These rapid analyses are coupled with optimization tools (gradient or non-gradient based) to perform aircraft design optimizations.

The fuel burn computations mentioned above are done with simplificatons of the aircraft performance and mission profile, resulting in inaccurate prediction of the total aircraft fuel burn. For example, the constant L/D, TSFC, and flight speed assumed in the Breguet range equation do not reflect the actual aircraft operation, as their values vary across the flight operating points in the mission profile. Moreover, most fuel burn computation focuses more on the cruise portion, which is critical for long range missions, but not necessarily so for shorter range missions. For shorter range missions, the climb segments will contribute significantly to the total fuel consumption as well. For a more accurate fuel burn computation, performing a detailed mission analysis that include all phases in the mission profile is thus necessary. Instead of using the Brequet range equation, the range equation now needs to be evaluated via a numerical integration procedure. However, performing a detailed mission analysis is computationally expensive due to the many performance evaluations required in the procedure. The computational issue is further exacerbated when we use the mission analysis in optimizations, which are typically done iteratively, or uncertainty quantifications (e.g., using the Monte Carlo method), which require multiple function evaluations.

We now describe the mission analysis procedure to compute the fuel weight W_{fuel}, range R, and mission time t by numerically integrating a given mission profile. This mission analysis procedure has also been used in the previous work by the authors [17]. As inputs we have the mission profile parameters (such as altitude and Mach number for cruise segments; flight speed, initial and final altitudes for climb and descent segments), the initial takeoff weight, and the final zero fuel weight (ZFW) for each mission. The weight, mission segment range and time are then solved iteratively using an *all-at-once* approach. Using this approach, a set of residual equations, $\mathcal{R}$, is set up using the endpoint weights of each segment as states. The residual is set to zero to determine the characteristics for the entire mission profile. Each segment can then be analyzed independently, based only on the current states of the system. To set up the residual equations, we need to match the endpoint weights of two adjacent segments: $W_{f_j} - W_{i_{j+1}} = 0$, where $j = 1, ..., N_{\text{seg}}$ denotes the segment index; W_i and W_f denote the segment's initial and final weight, respectively. Similarly, at the boundaries, $W_{i_1} = W_{\text{TO}}$ and $W_{f_{N\text{seg}}} = W_{\text{ZF}}$, where W_{TO} and W_{ZF} refer to the takeoff and zero fuel weights. We then use a Newton–Krylov algorithm to solve the nonlinear system. This forces the weights of the various segments to be consistent with each other, providing a valid and continuous mission profile.

The amount of fuel burned during startup, taxi, takeoff, and landing is computed using the fuel fraction method, where $W_f = (1 - \zeta) W_i$, with ζ referring to the fuel fraction value. The numerical integration to compute the fuel burn for the climb, cruise, and descent segments is derived from the range equation. With the TSFC (c_T) defined as the weight of fuel burned per unit time per unit thrust (N/s), we can compute the rate of reduction of aircraft weight as $dW/dt = -c_T T$, where W and T denote aircraft weight and thrust, respectively. Using this relation and the generic integral equation for range, $R = \int_{t_i}^{t_f} V dt$, the numerical integrations for range are given below. The subscripts i and f in the integration limits correspond to the initial and final values, respectively. For the cruise segment, the integration

is done with respect to weight,

$$R = \int_{W_i}^{W_f} -\frac{V}{c_T T} dW, \tag{1}$$

whereas for the climb and descent segments, the range equation is integrated over the change in altitude,

$$R = \int_{h_i}^{h_f} \frac{V \cos\gamma}{RC} dh, \tag{2}$$

where h and γ denote the altitude and the flight path angle. The rate of climb, RC, is derived from the equation of motion, $T_{\text{av}} \cos(\phi_T + \alpha) - D - W \sin\gamma = (W/g)(dV/dt)$, and that $RC = V \sin\gamma$. The symbol T_{av} denotes the available thrust, D denotes drag, and g is the gravitational acceleration. The thrust inclination angle is denoted by ϕ_T (typicaly assumed to be zero [32]), and α refers to the angle of attack. With small angle approximations, this equation yields

$$RC = \frac{(T_{\text{av}} - D)\, V}{W\left(1 + \frac{V}{g}\frac{dV}{dh}\right)}. \tag{3}$$

We have the information of flight speed and altitude for each segment interval from the mission specification. TSFC is a property of the aircraft engine, which is assumed constant in this work. We then need to compute T to evaluate (1) and (2), which we can find once we know D. We can evaluate drag upon determining α and tail rotation angle, η, which satisfy the lift (e.g., level flight, $L = W$, for cruise) and trim ($C_M = 0$) constraints simultaneously. These two angles can be found by solving a Newton search algorithm. This procedure computes the mission range given the fuel weight ($W_{\text{TO}} - W_{\text{ZF}}$). When the mission range is specified, we perform a secant algorithm to find the corresponding fuel weight, W_{fuel}. Following this procedure, the required number of aerodynamic performance evaluations would be equal to the product of the number of missions, number of secant iterations, number of iterations to solve the residual equations, number of integration intervals, and number of Newton iterations to solve for the angles. This analysis would require millions of aerodynamic solutions, which would be computationally prohibitive. Surrogate models are thus built to approximate the aerodynamic force and moment coefficients (C_L, C_D, and C_M) to be used in the mission analysis computation. When a sample-based surrogate modeling technique is used, the required number of aerodynamic performance evaluation calls is reduced to the number of samples used to build the surrogates, making the procedure computationally tractable.

The surrogate-based mission analysis procedure described here allows us to perform mission optimizations, where we set some parameters (e.g., cruise Mach number and altitude) as design variables; aerostructural optimizations (e.g., to minimize fuel burn or DOC) with an accurate fuel burn computation; and coupled mission and aerostructural optimizations. To obtain meaningful results from these optimizations, however, we first need to have reliable and accurate surrogate models.

III. Surrogate Modeling Techniques

Surrogate models use mathematical models to provide simpler approximations of physical systems, to reduce the computational expenses of analyses and optimizations [7, 8, 9]. Essentially, surrogate models are used as low-cost substitutes for exact evaluations in the computational task [33]. These approximation models are also known as *metamodels* [7], or *models of models* [20, 34]. To select the suitable surrogate models for our surrogate-based mission analysis procedure, we first discuss the available surrogate modeling techniques.

Eldred *et al.* [35] classified the surrogate models into three categories: data-fits, reduced-order models, and hierarchical models. The derivation of data-fit models typically involves interpolation or regression of data generated by solving the full physical system at a set of sample points, which can be generated through various sampling techniques [36]. Some popular examples of models belonging to this category include polynomial regression, kriging [37, 38, 39], projection pursuit regression [40], and RBF [41]. Reduced-order models approximate the relationships between system inputs and outputs, while reducing the order of the original system. The approximation is obtained by projecting the original model onto a basis that spans a space of lower dimension. An overview of model reduction methods is provided by Antoulas [42]. The reduced space basis can be computed using a number of different methods, including Krylov-subspace methods [43], approximate balanced truncation [44, 45], and proper orthogonal decomposition (POD) [46, 47, 48, 49]. Hierarchical models are also called multifidelity or variable-fidelity models [50, 51]. The derivation of low-fidelity models is typically problem-dependent, such as by using the same high-fidelity models but with a higher residual tolerance [40], a coarser grid [52, 53], or simpler engineering models that simplify the physics [54, 55].

The reduced-order and hierarchical surrogate models can be classified as *physics-base* approaches, since they exploit and simplify the governing equations [56]. These models are thus considered as *intrusive* methods. The data-fit surrogate models, on the other hand, belong to the *black-box* approach category, where the derivations are only based on the inputs and outputs of the high-fidelity models, without necessarily knowing the underlying governing equations (*non-intrusive* methods). Black-box models typically approximate a function at a point in the N_d-dimensional input space $\mathbf{x}_0 \in \mathbb{R}^{N_d}$ based on the available N_s sample information, including the sample locations $\mathbf{x}_s \in \mathbb{R}^{N_s \times N_d}$ and values $\mathbf{y}_s \in \mathbb{R}^{N_s}$,

$$y\left(\mathbf{x}_0\right) \approx \hat{y}\left(\mathbf{x}_0, \mathbf{x}_s, \mathbf{y}_s, \boldsymbol{\alpha}\right). \tag{4}$$

The symbol $\boldsymbol{\alpha}$ denotes a vector of model parameters, i.e., the undetermined coefficients that are typically derived based on the available training sample set.

Black-box surrogate models can be further categorized into local, multipoint, and global models based on the data points used in the model construction [57]. Local surrogate models approximate the actual functions around a single data point, such as Taylor approximation and intervening variables. These models are only valid in the vicinity of that particular point. Multipoint models use more than one data point, typically two, in the model constructions. Some examples of this model category include two-point exponential approximation (TPEA) and two-point adaptive nonlinear approximation (TANA). Global models, on the other hand, produce approximation models that are valid over the entire input space of interest, thus having a broader applicability than the former two models. Readers should be familiar with some examples of global surrogate models, such as polynomial regression, multivariate adaptive regression splines (MARS), support vector machine (SVM), artificial neural network (ANN), kriging, and RBF.

Based on the derivation technique, the black-box models can be categorized into *regression* and *interpolation*. Regression models are derived in a least-squares sense, which do not necessarily reproduce the exact function values at training sample locations $\mathbf{x}_s$,

$$y\left(\mathbf{x}\right) = \hat{y}\left(\mathbf{x}\right) + \epsilon \quad \Rightarrow \quad y\left(\mathbf{x}_s\right) \neq \hat{y}\left(\mathbf{x}_s\right) \tag{5}$$

The symbol ϵ represents the random independent and identically distributed (i.i.d.) error component in the data. One popular example of regression models is the polynomial response surface,

$$\hat{y}\left(\mathbf{x}\right) = \mathbf{f}^T\left(\mathbf{x}\right)\boldsymbol{\beta}, \tag{6}$$

where $\mathbf{f}\left(\mathbf{x}\right) = \left[f_1\left(\mathbf{x}\right), f_2\left(\mathbf{x}\right), \ldots, f_{N_f}\left(\mathbf{x}\right)\right]^T$ is a vector of N_f basis functions and $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_{N_f}\right]$ is a vector of the undetermined coefficients. A one-dimensional quadratic polynomial regression, for example, has the following set of basis function $\mathbf{f}\left(x\right) = \left[1, x, x^2\right]$. Since regression models are derived in a least-squares sense, they are more suitable to approximate functions with inherent random error, such as measurement data. When the training data are from deterministic computer experiments, the premise that the error between regression models and data are i.i.d. is false [58]. The term *computer experiments* refers to numerical codes that mimic some relevant physical phenomena. A computer experiment is *deterministic* when repeated experiments with the same input settings return exactly the same outputs [59]. In other words, no measurement (random) error component is involved. To approximate computer experiment data, interpolation models are thus more appropriate since they can reproduce the function values at sample locations exactly,

$$\hat{y}\left(\mathbf{x}_s\right) = y\left(\mathbf{x}_s\right). \tag{7}$$

Kriging and RBF models belong to the interpolation model category, whereas other global black-box surrogate models (MARS, SVM, ANN) are regression models. Constructing and using black-box surrogate models involve the following components,

1. Data generation (sampling): these data are interchangeably called *sample data*, *observational data* (in statistics community), and *training data* (in machine learning community). These data contain the sample locations, $\mathbf{x}_s$, and the corresponding function values y_s.

$$\mathcal{S} = \left\{\mathbf{x}_{s_i}, y_{s_i}\right\}_{i=1,2,\ldots,N_s} \tag{8}$$

2. Model structure selection and parameter estimation: these attributes are specific to the selected surrogate modeling technique. For example, we need to select the correlation function and estimate its parameters when using kriging models.

3. Model assessment: this procedure typically aims to evaluate the goodness of fit for regression models (due to the i.i.d. error component as shown in (5)) and the approximation accuracy at untested data.

A surrogate model can be *parametric* or *nonparametric*, depending on how the approximation functions express the input-output relations [57]. A parametric model no longer uses the training data once the model parameters or undetermined coefficients are derived. The polynomial regression described above is one popular example of parametric models. While models belonging to this category are relatively simple, they have limited utility when the actual input-output relationship is complex. A nonparametric model, on the other hand, still uses the training sample data in making prediction at new points, even after the set of undetermined coefficients has been estimated from the data. SVM, ANN, and RBF models belong to this category.

A survey of various *sampling plans*, also referred to as a *design of experiments*, can be found in [60, 61]. When building surrogate models on unknown landscapes, a sampling plan that is uniform, irregular, and space-filling is favorable [58]. The random Monte Carlo simulation (MCS) method is a very popular choice in industry, mainly due to its simplicity [61]. Another popular choice is the latin hypercube sampling (LHS) plan [62], as its projections onto each variable axis are uniform. Since there are not any specific guidelines to determine the "appropriate" sampling size *a priori*, sequential and adaptive sampling plans have become more popular recently. The new points (*infill points*) are selected based on some *infill criteria* to improve the model's predictive capability. There are mainly two categories, namely *exploitation* and *exploration* [58]. The exploitation criteria are used mostly in surrogate-based optimization (i.e., when surrogate models are used to approximate the objective function), to help finding the optimum point. Some examples include the minimizing the predictor approach and the trust-region method. The exploration criteria aim to "fill the gaps" between existing sample points to ensure that the samples are evenly distributed spatially. This category consists of sequential space-filling sampling plans such as Sobol' [63] and Halton sampling sequence [64], as well as an adaptive approach that locates infill points with the highest estimated error (e.g., using the kriging variance or MSE as a metric). In general, maximizing variance when adding samples tends to maximize the intersite distances (*D-optimality*) [65].

We need to validate the models before using them as surrogates in computationally-intensive analyses and optimizations, lest they will render the results invalid and meaningless. Meckesheimer *et al.* [34] presented an overview of the cross-validation method. In the p-fold cross validation approach, the sample set is first divided into p subsets. Then, we reconstruct the metamodel p times, by omitting one of the subsets each time, to compute the approximation errors. When each subset contains only one sample point, this procedure is called the *leave-one-out* cross-validation [66]. However, the cross-validation approach tends to be more biased towards over-represented regions. Due to this limitation, a more reliable model validation approach that employs additional test points to compute the approximation errors is preferred [61]. One of the most commonly used error measure is the root mean square error (RMSE),

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}, \tag{9}$$

where m denotes the number of validation (test) points. The normalized RMSE is also often used, especially when the function value varies a lot in the input space of interests. In this error measure, each error component, $(y_i - \hat{y}_i)$, is normalized with respect to its actual value, y_i, before computing RMSE, to give us the relative approximation error.

In this work, we want to use surrogate models to approximate the aerodynamic force and moment coefficients with data obtained from solving aerodynamic models, which are essentially deterministic computer experiments. The surrogate models need to be globally accurate, i.e., to cover the entire input space of interest. With these two considerations, we consider only global and interpolative models, which narrows down our options to kriging and RBF models. These two models are described in more details below.

A. Radial Basis Function Model

RBF is a nonparametric black-box surrogate model which emulates complicated design landscapes using a weighted sum of simple functions,

$$\hat{y}\left(\mathbf{x}_0, \mathbf{x}_s, \mathbf{y}_s, \boldsymbol{\alpha}\right) = \Psi_0^T \mathbf{w} = \sum_{i=1}^{N_c} w_i \psi\left(\|\mathbf{x}_0 - \mathbf{c}_i\|\right) \tag{10}$$

The function $\psi\left(\|\cdot\|\right)$ is the kernel function, centered at $\mathbf{c}_i$, $i = 1, \ldots, N_c$. The notation $\|\cdot\|$ denotes a Euclidean distance, d. Typically, the training sample points are used as the centers, thus $\mathbf{c} = \mathbf{x}_s$ and $N_c = N_s$. The vector of undetermined coefficients, $\mathbf{w}$, is determined by solving the following system of linear equations,

$$\Psi \mathbf{w} = \mathbf{y}_s. \tag{11}$$

The notation Ψ refers to the *gram matrix*, where $\Psi_{ij} = \psi\left(\left\|\mathbf{x}_{s_i} - \mathbf{x}_{s_j}\right\|\right)$, i.e., the kernel function evaluated at the Euclidean distance between the i^{th} and j^{th} samples. RBF belongs to the class of generalized linear models, as it is

linear in terms of the weights $\mathbf{w}$. The main difference between RBF and the standard polynomial regression response surface is in the choice of basis functions, where RBF models use kernel functions as the basis functions. Some commonly used kernel functions are listed in Table 1 below.

Kernel function	$\phi(d)$	Typical θ
Thin-plate splines	$d^2 \log d$	–
Inverse multi-quadric	$1/\sqrt{(1+\theta d^2)}$	1
Inverse multi-quadratic	$1/\left(1+\theta d^2\right)$	1
Cubic	d^3	–
Square-exponential (Gaussian)	$\exp\left[-\theta d^2\right]$	1

Table 1: Kernel functions for RBF models.

Following the conventions for the surrogate model expression (4), the model parameters $\boldsymbol{\alpha}$ include the basis function weights $\mathbf{w}$ and the hyperparameters for the basis functions, θ. In this RBF formulation (10), the dependence on the sample values, $\mathbf{y}_s$, is implicit in $\mathbf{w}$.

B. Kriging Model

The kriging surrogate model was initially developed in the field of geostatistics by Danie G. Krige (after whom the method is named) in 1951 [67]. The term "kriging" was first coined by Matheron in 1963 [68], who was also the first to formulate kriging mathematically. When first derived in the geostatistics field, kriging was used to model continuous and uniquely defined functions relating numbers (e.g., measurement data) to a domain of geographic coordinates (in one-, two-, or three-dimensional domains) [69]. The foundation of using kriging models in the design and analysis of computer experiments (DACE) was first developed by Sacks *et al.* [39], where points in the input space are analogous to the spatial (geographical) coordinates.

In kriging models, we assume that the deterministic response $y(\mathbf{x})$ is a realization of a stochastic process $Y(\mathbf{x})$ [38, 39],

$$Y(\mathbf{x}) = \sum_{k=1}^{N_f} f_k(\mathbf{x})\,\beta_k + Z(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\,\boldsymbol{\beta} + Z(\mathbf{x}). \tag{12}$$

The first term is a generalized linear model that determines the trend of the kriging model, which looks similar to that of the standard regression model (5). The notations used in the global model are as previously defined. The critical difference between the two models lies in the stochastic component. For the kriging model, instead of the i.i.d. assumption, the stochastic component $Z(\mathbf{x})$ is treated as the realization of a stationary Gaussian random function with zero expected value, $\mathrm{E}[Z(\mathbf{x})] = 0$, and covariance

$$\mathrm{Cov}\left[Z(\mathbf{x}_i), Z(\mathbf{x}_j)\right] = \sigma^2 R(\mathbf{x}_i, \mathbf{x}_j), \tag{13}$$

where $R(\cdot)$ denotes the correlation function with $R(0) = 1$. Therefore, kriging models give exact prediction at sample points, with increasing error variance as we go further from these sample points. In other words, in kriging models the data are assumed to be exact but the function is a realization of a Gaussian process [70]. This second term is called the *localized deviation* [60], *bias*, or the *systematic departure* from the linear model [38]. A *stationary* correlation function is typically assumed in kriging models, where the correlation between any two points in the input space, $y(\mathbf{x}_i)$ and $y(\mathbf{x}_j)$, depends only on the difference vector $\Delta\mathbf{x} = \mathbf{x}_i - \mathbf{x}_j$, thus $R(\mathbf{x}_i, \mathbf{x}_j) = R(\mathbf{x}_i - \mathbf{x}_j)$. Under this stationarity assumption, the prior distribution for $\mathbf{y}_s$ at the sample set $\mathcal{S}$ within the input space does not change if $\mathcal{S}$ is shifted within the space. The derivation of kriging models follows the theory of regionalized variables, which is a special statistical theory that explicitly considers spatial properties, and uses random variables to model spatial functions [69]. This theory is very general since it neglects the physical nature of the phenomenon under study [71].

For higher-dimensional problems, the correlation function in a kriging model typically satisfies the *product correlation rule*, where the correlation function can be expressed as a product of stationary, one-dimensional correlations,

$$R_{ij}(\boldsymbol{\theta}, d) = \prod_{k=1}^{N_d} R\left(\theta^{(k)}, d_{ij}^{(k)}\right). \tag{14}$$

The vector of correlation parameters is denoted as $\boldsymbol{\theta} = \left\{\theta^{(k)}\right\}$, $k = 1, \ldots N_d$. The notation $d_{ij}^{(k)}$ is the distance between two points in the k^{th} dimension, $\left|x_i^{(k)} - x_j^{(k)}\right|$. These correlation parameters (kriging hyperparameters) are

also referred to as *length scales* or *distance weights*, and are typically found via the maximum likelihood estimation (MLE) approach. Large θ values correspond to weak spatial correlation, whereas small values correspond to strong spatial correlation [72]. When each variable has a distinct physical meaning, it makes sense to use an anisotropic correlation function, i.e., having different $\theta^{(k)}$ values in different dimensions. In that case, we have more flexibility in the modeling, but at the expense of a more complex MLE [65, 73].

The correlation functions can be classified into those that exhibit a linear behavior near the origin ($R(\theta, d) \propto d$ for small d), and those that exhibit a parabolic behavior ($R(\theta, d) \propto d^2$) [73]. The latter is more suitable when the underlying function is continuously differentiable. The commonly used correlation functions for these two categories are tabulated below. For simplicity, the superscripts and subscripts are dropped from the correlation function expressions. The exponential and hole-effect models are more typically used in hydrologic applications, and the latter is only

Linear		Parabolic	
Exponential	$R(\theta, d) = \exp(-\theta d)$	Gaussian	$R(\theta, d) = \exp\left(-\theta d^2\right)$
Linear	$R(\theta, d) = \max[0, 1 - \theta d]$	Cubic spline	$R(\theta, d) = \begin{cases} 1 - 6(\theta d)^2 + 6(\theta d)^3 & \theta d \le 0.5 \\ 2(1 - \theta d)^3 & 0.5 < \theta d < 1 \\ 0 & \theta d \ge 1 \end{cases}$
Spherical	$R(\theta, d) = 1 - 1.5\xi + 0.5\xi^3$ where $\xi = \min[1, \theta d]$		
Hole-effect	$R(\theta, d) = \sigma^2 (1 - \theta d) \exp(-\theta d)$		

Table 2: Correlation function classification for kriging models

suitable for one-dimensional processes [74]. When deemed necessary, we can model the random measurement error in data by adding the nugget-effect model in the correlation function,

$$R(\theta, d) = C_0 \delta(d) = \begin{cases} 0 & d > 0 \\ C_0 & d = 0 \end{cases} \tag{15}$$

where C_0 is the *nugget variance*. In addition to modeling random error, this model can also represent *microvariability*, i.e., variability at a scale smaller than the separation distance between the closest measurement points. The term "nugget" comes from mining, where the concentration of a mineral or the ore grade varies in a practically discontinuous fashion due to the presence of nuggets at the sampling points [74].

When the global model is assumed known, kriging model produces the *best linear predictor* (BLP). When the known global model is a constant, we have a *simple kriging* [75]. On the other hand, when the global model is unknown and thus needs to be derived, the model is referred to as the *best linear unbiased predictor* (BLUP). When a constant global model is assumed, a BLUP model is called *ordinary kriging*, whereas when a set of basis functions is used (typically low-order polynomials, e.g., linear or quadratic), it is referred to as *universal kriging* [75] or *kriging with a trend* [24]. Ordinary kriging models are more popular and commonly used, as the *a priori* knowledge of the trends in the data is typically unknown [58]. While BLP is a non-parametric model, BLUP is a semi-parametric model. The linear regression in the global model component forms the parametric part, whereas the stochastic component forms the non-parametric part [76].

The kriging derivation as a BLUP model, which is based on the mean squared error (MSE) minimization, and its Bayesian interpretation are presented next. Kriging as a BLP model can be derived in a similar manner, by ommitting the unbiasedness constraint.

1. *Kriging Derivation: MSE Minimization*

Kriging approximation method belongs to the class of generalized linear model [57], where we can express its prediction at an arbitrary evaluation point $\mathbf{x}_0$ as a linear combination of the sample data $\mathbf{y}_s$, with weights $\mathbf{c}(\mathbf{x}_0)$,

$$\hat{y}(\mathbf{x}_0) = \mathbf{c}^T(\mathbf{x}_0)\, \mathbf{y}_s. \tag{16}$$

For conciseness, we will drop $\mathbf{x}_s$, $\mathbf{y}_s$, $\boldsymbol{\alpha}$ from $\hat{y}(\cdot)$ (4) in the subsequent discussion. To derive the *optimum* $\mathbf{c}(\mathbf{x}_0)$, we minimize the MSE of the approximation,

$$\text{MSE}[\hat{y}(\mathbf{x}_0)] = \text{E}\left[\left(\mathbf{c}^T(\mathbf{x}_0)\, \mathbf{y}_s - Y(\mathbf{x}_0)\right)^2\right], \tag{17}$$

subject to the unbiasedness constraint, $\text{E}\left[\mathbf{c}^T(\mathbf{x}_0)\, \mathbf{y}_s\right] = \text{E}[Y(\mathbf{x}_0)]$, where $\text{E}[\cdot]$ denotes an expected value. The unbiasedness constraint gives the following relation, $\mathbf{F}_s^T \mathbf{c}(\mathbf{x}_0) = \mathbf{f}(\mathbf{x}_0)$, where $\mathbf{F}_s \in \mathbb{R}^{N_s \times N_f}$ is the basis function

matrix corresponding to sample data, with $(\mathbf{F}_s)_{ik} = f_k(\mathbf{x}_{s_i})$. By expanding the MSE expression (17), and using the unbiasedness constraint, covariance function (13), and the basic relations for the variance and covariance functions, we obtain the following expression,

$$\begin{aligned} \text{MSE}\left[\hat{y}(\mathbf{x}_0)\right] &= \text{E}\left[\left(\mathbf{c}^T(\mathbf{x}_0)\right)^2\right] - 2\text{E}\left[\mathbf{c}^T(\mathbf{x}_0)\, Y(\mathbf{x}_0)\right] + \text{E}\left[Y^2(\mathbf{x}_0)\right] \\ &= \sigma^2\left[1 + \mathbf{c}^T(\mathbf{x}_0)\,\mathbf{R}_s\mathbf{c}(\mathbf{x}_0) - 2\mathbf{c}^T(\mathbf{x}_0)\,\mathbf{r}(\mathbf{x}_s,\mathbf{x}_0)\right]. \end{aligned} \tag{18}$$

$\mathbf{R}_s$ is the correlation matrix of samples, where $(\mathbf{R}_s)_{ij} = R\left(\mathbf{x}_{s_j} - \mathbf{x}_{s_i}\right)$ and $\mathbf{r}(\mathbf{x}_s,\mathbf{x}_0)$ is the correlation vector between the evaluation point and the samples.

We now transform the MSE minimization problem formulated above into an unconstrained one using a Lagrange multiplier $\boldsymbol{\lambda}$ as shown below,

$$\underset{\mathbf{c}(\mathbf{x}_0)}{\text{minimize}} \quad f(\mathbf{c}(\mathbf{x}_0)) = \sigma^2\left[1 + \mathbf{c}^T(\mathbf{x}_0)\,\mathbf{R}_s\mathbf{c}(\mathbf{x}_0) - 2\mathbf{c}^T(\mathbf{x}_0)\,\mathbf{r}(\mathbf{x}_s,\mathbf{x}_0)\right] + \boldsymbol{\lambda}^T\left(\mathbf{F}_s^T\mathbf{c}(\mathbf{x}_0) - \mathbf{f}(\mathbf{x}_0)\right) \tag{19}$$

The minimum is achieved when the optimality condition, $df(\mathbf{c}(\mathbf{x}_0))/d\mathbf{c}(\mathbf{x}_0) = 0$, is satisfied. We then have the following system of equation,

$$\begin{bmatrix} \mathbf{R}_s & \mathbf{F}_s \\ \mathbf{F}_s^T & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{c}(\mathbf{x}_0) \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} \mathbf{r}(\mathbf{x}_s,\mathbf{x}_0) \\ \mathbf{f}(\mathbf{x}_0) \end{bmatrix}. \tag{20}$$

Solving this equation, we obtain the expression for kriging model,

$$\hat{y}(\mathbf{x}_0) = \mathbf{f}^T(\mathbf{x}_0)\,\hat{\boldsymbol{\beta}} + \mathbf{r}^T(\mathbf{x}_s,\mathbf{x}_0)\,\mathbf{R}_s^{-1}\left(\mathbf{y}_s - \mathbf{F}_s\hat{\boldsymbol{\beta}}\right), \tag{21}$$

where the generalized least squares estimate of the undetermined coefficients for the global model can be expressed as,

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}_s^T\mathbf{R}_s\mathbf{F}_s\right)^{-1}\left(\mathbf{F}_s^T\mathbf{R}_s^{-1}\mathbf{y}_s\right). \tag{22}$$

The MSE of $\hat{y}(\mathbf{x}_0)$ is obtained by substituting (20) to (18),

$$\text{MSE}\left[\hat{y}(\mathbf{x}_0)\right] = \sigma^2\left[1 - \begin{bmatrix}\mathbf{r}(\mathbf{x}_s,\mathbf{x}_0) & \mathbf{f}(\mathbf{x}_0)\end{bmatrix}\begin{bmatrix}\mathbf{R}_s & \mathbf{F}_s \\ \mathbf{F}_s^T & \mathbf{0}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{r}(\mathbf{x}_s,\mathbf{x}_0) \\ \mathbf{f}(\mathbf{x}_0)\end{bmatrix}\right]. \tag{23}$$

This MSE, or the kriging variance, represents the model uncertainty in predicting the function value at $\mathbf{x}_0$. This quantity is helpful in assessing the model, and can be used in an adaptive sampling approach, as will be demonstrated in this work.

2. *Bayesian Interpretation of Kriging*

The Bayesian interpretation of kriging as the BLUP is presented here. While in general the kriging derivation via the MSE minimization approach and Bayesian approach do not yield the same expression for the estimators, they are identical when the prior dstribution of $Z(\mathbf{x})$ is Gaussian and the undetermined coefficients for the global model $\boldsymbol{\beta}$ has a diffuse, or noninformative, prior, as we will show here. For more details on the Bayesian interpretation of kriging models, readers are referred to some previous works [38, 65, 77, 78, 79, 80].

The general kriging equation shown in (12) is now treated as the Bayesian prior on the true response functions. Let us assume that the prior distribution of $Z(\cdot)$ is Gaussian with expected value zero and covariance function as previously described (13). The prior mean value of this random function $Y(\mathbf{x})$ is thus $\mathbf{f}^T(\mathbf{x})\,\boldsymbol{\beta}$. Let us also assume that $\boldsymbol{\beta}$ has a prior Gaussian distribution, $\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{b}, \tau^2\boldsymbol{\Sigma}\right)$, where $\mathbf{b}$, τ^2, and $\boldsymbol{\Sigma}$ denote the corresponding prior mean, variance, and correlation matrix, respectively. We can then do a Bayesian update using the information contained in the training sample set and compute the posterior probability, $P(\boldsymbol{\beta}\,|\mathbf{y}_s) \sim \mathcal{N}\left(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}\right)$, which is also a Gaussian as we use a conjugate prior. Here we apply the *Bayes' rule*, $P(\boldsymbol{\beta}, \mathbf{y}_s) = P(\boldsymbol{\beta}\,|\mathbf{y}_s)\,P(\mathbf{y}_s)$, where the joint probability can be expressed as

$$P(\boldsymbol{\beta}, \mathbf{y}_s) \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_s \\ \mathbf{b}\end{bmatrix}, \begin{bmatrix}\sigma^2\mathbf{R}_s & \mathbf{0} \\ \mathbf{0} & \tau^2\boldsymbol{\Sigma}\end{bmatrix}\right). \tag{24}$$

In the above equation, $\boldsymbol{\mu}_s$ is the expected value of the samples, which is $\mathbf{F}_s\boldsymbol{\beta}$ for the kriging model (12). Solving for the posterior distribution, we obtain the posterior mean $\tilde{\boldsymbol{\beta}}$ and covariance $\tilde{\boldsymbol{\Sigma}}$ for the undetermined coefficients of the global model,

$$\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}\left[\mathbf{F}_s^T\sigma^{-2}\mathbf{R}_s^{-1}\mathbf{y}_s + \tau^{-2}\boldsymbol{\Sigma}^{-1}\mathbf{b}\right], \quad \tilde{\boldsymbol{\Sigma}} = \left[\mathbf{F}_s^T\sigma^{-2}\mathbf{R}_s^{-1}\mathbf{F}_s + \tau^{-2}\boldsymbol{\Sigma}^{-1}\right]^{-1}. \tag{25}$$

Similarly, the posterior distribution of the random function $Y(\mathbf{x})$ at an evaluation point $\mathbf{x}_0$ can also be derived following the same Bayesian update approach. We can then use the posterior mean, $\mathrm{E}[Y(\mathbf{x}_0)|\mathbf{y}_s]$, as the expression for a kriging predictor,

$$\hat{y}(\mathbf{x}_0) = \mathbf{r}^T(\mathbf{x}_s, \mathbf{x}_0)\mathbf{R}_s^{-1}\mathbf{y}_s + \mathbf{a}^T\tilde{\boldsymbol{\beta}}, \tag{26}$$

where $\mathbf{a} = \mathbf{f}(\mathbf{x}_0) - \mathbf{F}_s^T\mathbf{R}_s^{-1}\mathbf{r}(\mathbf{x}_s, \mathbf{x}_0)$, and the variance is given as

$$\mathrm{Var}[Y(\mathbf{x}_0)|\mathbf{y}_s] = \sigma^2\left(1 - \mathbf{r}^T(\mathbf{x}_s, \mathbf{x}_0)\mathbf{R}_s^{-1}\mathbf{r}(\mathbf{x}_s, \mathbf{x}_0)\right) + \mathbf{a}^T\tilde{\boldsymbol{\Sigma}}\mathbf{a}. \tag{27}$$

The predictive (posterior) distribution is no longer stationary. In general, the variance is greater as the point $\mathbf{x}_0$ is further away from the training samples. When we have a diffuse prior, i.e., $\tau^2 \to \infty$,

$$\tilde{\boldsymbol{\beta}} \to \hat{\boldsymbol{\beta}} \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}} \to \left[\mathbf{F}_s^T\mathbf{V}_s\mathbf{F}_s\right]^{-1}, \tag{28}$$

where $\hat{\boldsymbol{\beta}}$ is the global model coefficients derived through the MSE minimization procedure (22), and $\mathbf{V}_s$ denotes the covariance matrix of the samples, $\sigma^2\mathbf{R}_s$. The kriging predictor, $\hat{y}(\mathbf{x}_0)$, now agrees with the one previously derived based on the MSE minimization approach (21), and so does the kriging variance or MSE (23). The derivation of the BLP model can be done similarly as the one presented here, but without the unbiasedness constraint. The BLP model has the same expression for the predictor, $\hat{y}(\mathbf{x}_0)$, while the variance is reduced to $\mathrm{MSE}[\hat{y}(\mathbf{x}_0)] = \sigma^2\left[1 - \mathbf{r}^T(\mathbf{x}_s, \mathbf{x}_0)\mathbf{R}_s^{-1}\mathbf{r}(\mathbf{x}_s, \mathbf{x}_0)\right]$.

In the derivation presented above, the model assumes that the prior variance σ^2, the family and parameters of the correlation function, $R(\cdot)$, are known. Typically, the designers will determine the correlation function, and apply the *empirical Bayes* approach to find the parameters to be most *consistent* with the observed data [38, 65], in particular by employing the MLE approach. Consistent with our assumption that $Z(\cdot)$ is Gaussian, the likelihood function can be expressed as,

$$L\left(\boldsymbol{\beta}, \sigma^2, R\middle|\,\mathbf{y}_s\right) = -\frac{1}{2}\left[N_s\log 2\pi + N_s\log\sigma^2 + \log|\mathbf{R}_s| + \frac{1}{\sigma^2}\left(\mathbf{y}_s - \mathbf{F}_s\boldsymbol{\beta}\right)^T\mathbf{R}_s^{-1}\left(\mathbf{y}_\mathrm{S} - \mathbf{F}_s\boldsymbol{\beta}\right)\right] \tag{29}$$

Setting $\partial L\left(\boldsymbol{\beta}, \sigma^2, R\middle|\,\mathbf{y}_s\right)/\partial\boldsymbol{\beta} = 0$, we recover $\hat{\boldsymbol{\beta}}$ at the stationary point, $\boldsymbol{\beta}_\mathrm{MLE} = \hat{\boldsymbol{\beta}}$, where the subscript MLE denotes the corresponding MLE solution. Setting $\partial L\left(\boldsymbol{\beta}, \sigma^2, R\middle|\,\mathbf{y}_s\right)/\partial\sigma^2 = 0$ yields the maximum likelihood solution for the prior variance,

$$\sigma^2_\mathrm{MLE} = \frac{1}{N_s}\left(\mathbf{y}_s - \mathbf{F}_s\hat{\boldsymbol{\beta}}\right)^T\mathbf{R}_s^{-1}\left(\mathbf{y}_s - \mathbf{F}_s\hat{\boldsymbol{\beta}}\right) \tag{30}$$

Some correlation functions have tunable parameters $\boldsymbol{\theta}$ that still need to be determined. Since there is no closed-form solution for these optimum parameters, we solve for $\boldsymbol{\theta}$ by performing a constrained iterative search. In this work, we employ the Hooke-Jeeves pattern search method [81].

3. *Gradient-enhanced Kriging*

A gradient-enhanced kriging model (GEK) interpolates gradient information, in addition to the function value, at each sample location, thus achieving a first-order-consistency requirement, in addition to the zeroth-order-consistency achieved by gradient-free kriging [21]. Depending on how the gradient information is used, there are two types of GEK, namely the indirect GEK and direct GEK. The former uses the gradient information to generate new samples around the available samples via a Taylor series expansion around those samples,

$$y(\mathbf{x}_{N_s+ik}) = y(\mathbf{x}_i) + \frac{\partial y(\mathbf{x}_i)}{\partial x^{(k)}}\Delta x^{(k)}. \tag{31}$$

In the direct GEK approach, the gradients are now directly included in the formulation as additional observations or sample data, as shown below,

$$\mathbf{y}_s = \left[y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_{N_s}), \frac{\partial y(\mathbf{x_1})}{\partial x^{(1)}}, \frac{\partial y(\mathbf{x_1})}{\partial x^{(2)}}, \ldots, \frac{\partial y(\mathbf{x_1})}{\partial x^{(N_d)}}, \frac{\partial y(\mathbf{x_2})}{\partial x^{(1)}}, \ldots, \frac{\partial y(\mathbf{x}_{N_s})}{\partial x^{(N_d)}}\right] \tag{32}$$

This method requires augmenting the correlation matrix with its derivative terms, which considerably increases the order of the correlation matrix to $N_s(N_d+1)$ from kriging's N_s. Consequently, the computational cost to build and use GEK models is higher than that of the original kriging models. See, for example, [21, 82, 83], for more details on the formulations and uses of GEK.

C. Kriging Compared to RBF

There is no clear consensus as to which of the kriging and RBF surrogate models has a better predictive performance. Wang and Shan [61] claimed that RBF is a compromise between kriging models and polynomial regressions, as it can interpolate the sample points (generally more accurate than polynomial regressions) and at the same time easier to construct than kriging models. Forrester and Keane [58] argued that kriging is the least assuming method, which provides a greater flexibility in the modeling. The flexibility comes mainly due to the parameters in the covariance function; however, it comes at the expense of the estimation of hyperparameters [76]. Jin *et al.* [20], on the other hand, concluded that RBF has the best performance overall in terms of accuracy, robustness (the most robust model is the one that is the least problem-dependent), efficiency (the amount of computational effort required for the surrogate model construction), transparency (the capability to provide information on model sensitivity to input variables and the inter-variable interactions), and conceptual sensitivity (ease of implementation). The comparison was performed with 13 analytical problems and one vehicle handling problem, with varying non-linearity, scale (dimensionality), and smoothness. With the varying opinions regarding the two models, it is safe to conclude that their predictive performance is essentially problem-dependent.

We can also analyze and compare the two models mathematically. Consider an ordinary kriging model (21) with a constant global model, $\hat{\beta}$. Rearranging the equation, we obtain

$$\hat{y}\left(\mathbf{x}_0\right) = \underbrace{\hat{\beta}}_{w_0} + \underbrace{\mathbf{r}^T\left(\mathbf{x}_s, \mathbf{x}_0\right)}_{\Psi_0^T} \underbrace{\left[\left(\mathbf{I} - \mathbf{1}\frac{\mathbf{1}^T\mathbf{R}_s^{-1}}{\mathbf{1}^T\mathbf{R}_s^{-1}\mathbf{1}}\right)\mathbf{R}_s^{-1}\mathbf{y}_s\right]}_{\mathbf{w}}, \tag{33}$$

where $\mathbf{1}$ is a vector of ones, as the basis function vector. Comparing the above expression to (10), we can see that an ordinary kriging model is reduced to RBF with an offset (w_0) [76]. The corresponding kriging model, however, has a greater flexibility as we can estimate the hyperparameters, $\boldsymbol{\theta}$, via MLE. The RBF hyperparameters, on the other hand, are fixed by users. Moreover, kriging models typically use anisotropic correlation functions (i.e., those that satisfy the product correlation rule), with different length scales in different dimensions, further increasing the model flexibility.

IV. Mixture of Experts

Black-box surrogate model training seeks to find the "best" model parameters by applying the empirical Bayes approach, which relies on the available data (observations), instead of the physical nature of the phenomenon. For example, in kriging model training we obtain the global model coefficients ($\hat{\boldsymbol{\beta}}$), variance (σ^2), and the correlation parameters ($\boldsymbol{\theta}$). This global fitting can potentially cause a limited modeling flexibility and the resulting model might be inadequate when there is hetereogeneity in the function profile, i.e., when the function complexity is input-dependent. For instance, the stationary covariance structure in the Gaussian process could be restrictive, as the model does not account for some areas of the input space having more activities than others [58]. This is not a serious problem in surrogate-based optimization, but it can pose a problem when building a globally accurate model, which is our current goal. This consideration motivates using multiple models instead of one global model [84].

There has been a growing interest in using multiple surrogates, either by combination or selection, instead of a single model in isolation [70]. Combining models in some way has been shown to improve the approximation performance of the surrogates [19]. For example, in *committees* we take the average of predictions from different trained models. Viana *et al.* [70] proposed using cross-validation error in both model selection and combination. When multiple surrogates are present, we can either select one with the lowest cross-validation error, or to use the cross-validation errors to create a weighted surrogate by minimizing the integrated square error. *Decision tree models* use a sequence of binary selections to select one model as the predictor. In this setting, different models are responsible for making prediction in different regions of input space, thus the selection is a function of input variables. When a probabilistic framework is used for combining models (instead of the hard split in the decision tree models), we have a *mixture of experts*. Different from the model combination and selection approaches mentioned above, in *Bayesian model averaging* we assume that there is only one model responsible for generating the whole data set, with a probability distribution across models reflecting the uncertainty as to which model that is. As the size of the data increases, the posterior probability will focus on just one model, and the rest will be discarded.

In this work, we propose a mixture of experts procedure based on the *divide-and-conquer* approach. The basic idea is to have several surrogate models to be responsible for different parts of the input space, to enable modeling the heterogenous complexity in the function profile. In the proposed approach, the prediction at an evaluation point $\mathbf{x}_0$ is modeled as a linear superposition, or weighted combination, of the *local experts*, $\hat{y}_k\left(\mathbf{x}_0\right)$, $k = 1, \ldots, K$, where K denotes the total number of local experts in the mixture. The weight or *mixing proportion*, denoted as $\pi_k\left(\mathbf{x}_0\right)$, is a

function of the evaluation point location in the input space. This mixture of experts can then be expressed as

$$\hat{y}\left(\mathbf{x}_0\right) = \sum_{k=1}^{K} \pi_k\left(\mathbf{x}_0\right) \hat{y}_k\left(\mathbf{x}_0\right), \quad \text{with} \quad 0 \leq \pi_k\left(\mathbf{x}_0\right) \leq 1 \quad \text{and} \quad \sum_k \pi_k\left(\mathbf{x}_0\right) = 1. \tag{34}$$

This mixture of experts can be considered as a hierarchical model, by introducing an unobservable latent indicator variable $z_k\left(\mathbf{x}_0\right)$ [84]. The discrete latent variables "assign" data points to specific components (local experts) in the mixture. The K-dimensional binary random variable $z_k\left(\mathbf{x}_0\right)$ is a 1-of-K encoding where $z_k\left(\mathbf{x}_0\right) \in \{0,1\}$ and $\sum_k z_k\left(\mathbf{x}_0\right) = 1$. The surrogate model prediction can then be modeled as

$$\hat{y}\left(\mathbf{x}_0\right) | \, z_k = 1 \sim \hat{y}_k\left(\mathbf{x}_0\right), \tag{35}$$

or in other words, the k^{th} local model is active when $z_k = 1$. The mixing coefficient will then be a function of the distribution of z_k, i.e., $\pi_k\left(\mathbf{x}_0\right)$ is a function of $p\left(z_k\left(\mathbf{x}_0\right) = 1\right)$.

Each expert $\hat{y}_k\left(\mathbf{x}_0\right)$ is constructed based on samples taken in the input subregion it is responsible for. With less complexity to be modeled within each local region, fewer samples are required to construct the local surrogate models, compared to when a single global model is used to approximate the entire range of the input space. With this "distributed approach," the computational cost required to build and use the surrogate models will consequently be reduced. For example, the correlation matrix for the ordinary kriging model is $\mathcal{O}\left(N_s^2\right)$ in size and its inversion is $\mathcal{O}\left(N_s^3\right)$ in cost. Even when the total number of samples used are the same for the global model and the mixture of experts model, the total computational expense will be less with the divide-and-conquer approach. When the job is distributed to local experts, we can disregard the correlation between samples that belong to different subregions. Moreover, each local expert is free to select the best model parameters to better reflect the characteristics of the underlying function in the input subregion it is responsible for (e.g., by having different length scales, $\boldsymbol{\theta}$, for each local kriging model). In short, this divide-and-conquer approach allows us to distribute a complex task into multiple simpler tasks.

Given the overview of the proposed mixture of experts approach, we now need to derive the following two important elements: (1) the procedure to divide the input space to appropriate subregions for the local experts, and (2) the expression for the mixing proportion, $\pi_k\left(\mathbf{x}_0\right)$. These procedures are described next.

A. Input Space Partitioning

Depending on whether the input space is already split when the local models are trained, Masoudnia and Ebrahimpour [85] classified the mixture of experts approaches into two categories below:

1. Mixture of implicitly localized experts (MILE): this approach reflects the more conventional mixture of experts as proposed by Jacobs *et al.* [86]. In this approach, the input space partitioning is decided by the gating network stochastically *during* training. Some popular works on MILE in the literature include mixtures of Gaussian processes [87], infinite mixtures of Gaussian process experts [88], and conditional mixtures of linear regressions [89, 90, 91]

2. Mixture of explicitly localized experts (MELE): in this approach, the input space partitioning is done explicitly by clustering *before* training the local experts. With this approach, designers have more flexibility to choose the clustering methods for their problems.

In general, MILE is more complicated and costly to train than MELE, owing to the interdependence between the input space partitioning and local expert training. In this work, we will consider only the MELE approach.

With some prior knowledge on the function profile, the designers can rely on engineering judgment to partition the input space spatially. This is the simplest approach, but lacks quantitative rigor. Nguyen-Tuong *et al.* [92] used a distance-based measure in partitioning the training data, assuming the same kernel width for all local kriging models. Another alternative is to partition the input space based on a certain criterion which better reflects the function profile, such as the function value or the gradient, which we adopt in our proposed procedure. In this approach, however, a set of training data ($\mathcal{T}$) are required for the clustering procedure. Depending on the information contained in the training data, there are two types of learning algorithms:

1. Supervised clustering algorithm: in this learning algorithm, the training data are labeled, i.e., we know which cluster each of them belongs to. The training algorithm derives the general expression to cluster unknown (new) data, such as in the least-squares classification, Fisher's linear classifier, logistic regression, and Gaussian classifier [19].

2. Unsupervised clustering algorithm: the training data for this category are unlabeled. The goal of this algorithm is to find the *hidden structure* in the data, or to discover groups of similar examples within the data. An expression is then derived to cluster unknown data based on the learned pattern. Some algorithms belonging to this category include the classical K-means clustering [93, 94, 95], and the Gaussian mixture models [96].

In the proposed approach, we need both the supervised and unsupervised clustering algorithms, for reasons that will become clear shortly. The unsupervised learning algorithm is required when we first cluster the training data based on the selected clustering criterion, since we have no prior knowledge on how they are split. Thus at this stage, the partitioning will be at the y-space, i.e., the training data are partitioned based on the function values or gradient information, without regard to their locations in the input space. When evaluating an arbitrary test point $\mathbf{x}_0$, however, we first need to compute the corresponding mixing proportion, $\pi_k(\mathbf{x}_0)$, which can be computed based on the input space (the $\mathbf{x}$-space) partitioning. For this purpose, we can perform a supervised clustering algorithm, since the training data are now already labeled upon completing the unsupervised clustering algorithm in the first stage.

For the unsupervised clustering algorithm we employ the *Gaussian mixture models*, by assuming that the underlying distribution of the data is a superposition formed by taking a linear combination of multiple Gaussian distributions [97, 98],

$$p(\mathbf{x}) = \sum_{k=1}^{K} \phi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{36}$$

where ϕ_k denotes the prior probability that a point belongs to the k^{th} cluster, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the corresponding mean and covariance. The parameters ϕ_k, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$ are solved iteratively via the *expectation-maximization* algorithm, or *EM* algorithm [99, 100]. By labeling each sample with its corresponding cluster assignment, we can separate the training data set into $\mathcal{T} \rightarrow \{\mathcal{T}_k\}, k = 1, \ldots, K$. The *regularized Gaussian classifier* is used as the supervised learning algorithm. This classifier belongs to a generative approach, where we model the class conditional densities $p(\mathbf{x}|z_k = 1)$ (Gaussian in this case) together with the prior probabilities for the clusters $p(z_k = 1)$ [19]. The parameters for both probabilities can be obtained by the maximum likelihood solutions.

B. Mixing Proportion Derivation

As mentioned, the mixing proportion is a function of the cluster posterior probability. Once the parameters in the regularized Gaussian classifier are derived at the supervised learning algorithm step, this posterior probability can then be computed as,

$$p(z_k = 1|\mathbf{x}) = \frac{p(\mathbf{x}|z_k = 1)\, p(z_k = 1)}{\sum_j p(\mathbf{x}|z_j = 1)\, p(z_j = 1)} \tag{37}$$

When there are only two clusters ($K = 2$), this cluster posterior probability becomes a *sigmoid function*, an S-shaped curved with values ranging from 0 to 1,

$$p(z_k = 1|\mathbf{x}) = \frac{1}{1 + \exp(-a)} = \sigma(a), \quad \text{where} \quad a = \ln \frac{p(z_k = 1|\mathbf{x})}{1 - p(z_k = 1|\mathbf{x})} \tag{38}$$

The *cluster boundary* is defined at the point where $p(z_k = 1|\mathbf{x}) = 0.5$. We can modify the sigmoid function by introducing weight (ω) and bias (λ), $\sigma(\omega a + \lambda)$. Altering λ shifts the cluster boundary, whereas altering ω changes the slope of the S-shaped curve around the cluster boundary. Since we want to maintain the cluster boundary position, we set λ to the default value 0. Increasing ω drives the sigmoid function to be closer to a step function, or $p(z_k = 1|\mathbf{x}) = \{0, 1\}$ as $\omega \rightarrow \infty$. For cases where $K > 2$, we use a softmax function,

$$p(z_k = 1|\mathbf{x}) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}, \quad \text{where} \quad a_k = \ln\left[p(\mathbf{x}|z_k = 1)\, p(z_k = 1)\right]. \tag{39}$$

The previous discussion on the effects of adding ω and λ to the sigmoid function also applies to this softmax function. In our proposed approach, we use this "modified" softmax function as the mixing proportion,

$$\pi_k(\mathbf{x}_0) = \frac{\exp(\omega a_k(\mathbf{x}_0))}{\sum_j \exp(\omega a_j(\mathbf{x}_0))} \tag{40}$$

We will vary ω and discuss how it affects the predictive performance of the mixture of experts, which will be presented in Section VI.

C. Mixture of Experts Procedure

With the input space partitioning described and the mixing proportion derived above, the procedure for the proposed mixture of experts approach is presented below,

1. Implement the Gaussian mixture model as the unsupervised learning algorithm to cluster the training data. The designers need to decide on the clustering criterion and the number of clusters prior to performing this step. The training data set for clustering, $\mathcal{T} = \{\mathbf{x}_n, y_n\}_n$, is now partitioned into K clusters, $\mathcal{T}_k = \{\mathbf{x}_n, y_n\}_{n \in \mathcal{C}_k}$, $k = 1, \ldots, K$, where $\mathcal{C}_k$ denotes the set of clustering training data indices that correspond to the k^{th} cluster.

2. Map the clustering of training data to the clustering in the input space ($\mathbf{x}$-space) by implementing the regularized Gaussian classifier as the supervised learning algorithm.

3. Build a separate local surrogate model within each cluster, $\hat{y}_k(\mathbf{x})$, $k = 1, \ldots, K$. First, we need to determine the samples to be used to build each local expert. One option is to use all points or a subset of points within each $\mathcal{T}_k$ set, or to perform an adaptive sampling algorithm starting with a subset of $\mathcal{T}_k$.

4. Compute the cluster posterior probability, i.e., the probability that $\mathbf{x}_0$ belongs to the k^{th} cluster, using (37).

5. Compute the corresponding mixing proportion, $\pi_k(\mathbf{x}_0)$, using (40).

6. Compute the mixture of experts estimation, $\hat{y}(\mathbf{x}_0)$, following (34), using the local experts and mixing proportions obtained in steps 3 and 5, respectively.

In this paper, we demonstrate this proposed mixture of experts procedure in approximating the aerodynamic force and moment coefficients of the selected aircraft configurations. The performance is then compared to those of some global surrogate models.

V. Problem Description

In this section we describe the two aircraft configurations considered in this study and the aerodynamic solver used to generate the aerodynamic data. We will then provide more details on the surrogate models that we will demonstrate and compare, as well as the selected sampling and model validation procedures.

A. Aircraft Configurations

In this paper, we test the different surrogate modeling techniques using the aerodynamic data corresponding to both the conventional and unconventional aircraft configurations. For the conventional configuration, we use the wing-tail configuration of the common research model (CRM) [101]. This aircraft exhibits design features typical of a transonic, wide-body, long-range aircraft, with overall dimensions similar to those of the Boeing 777-200ER. For the unconventional configuration, we consider a BWB configuration with the sizing parameters follow those presented by Lyu and Martins [18]. Figure 1 shows the layouts for both aircraft configurations, and the grid to be used in the aerodynamic solver, which is described next.

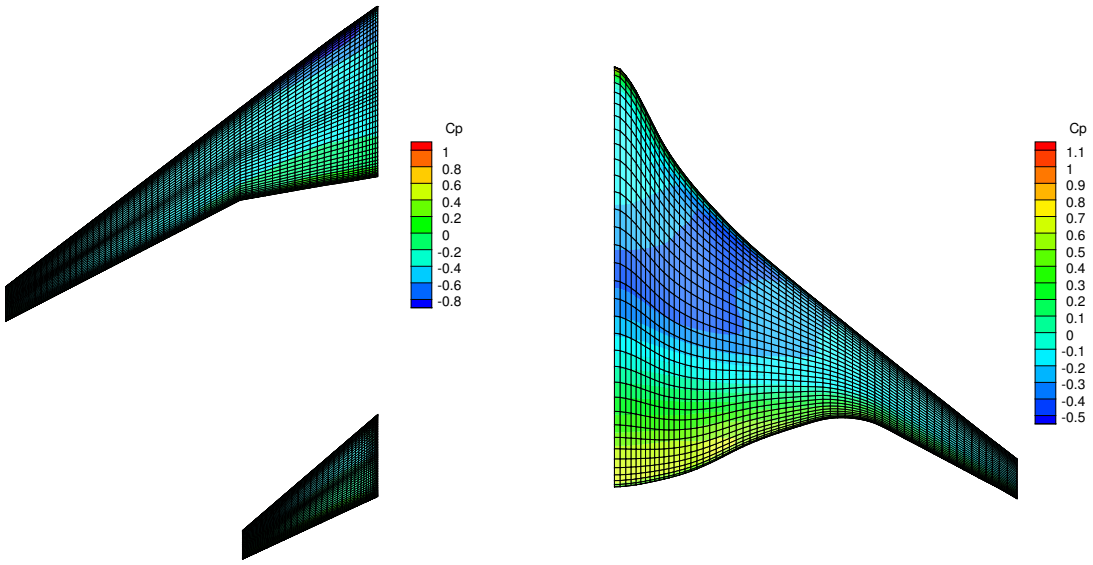
B. Aerodynamic Solver

A medium-fidelity aerodynamic solver, `TriPan`, is used to generate the "true" aerodynamic force and moment coefficient data in this work. This solver, which is developed by Kennedy and Martins [102, 103], uses the panel method. `TriPan` calculates the aerodynamic forces and moments of inviscid, incompressible, external lifting flows on unstructured grid using surface pressure integration, with constant source and double singularity elements. The surface of the body is discretized with quadrilateral and triangular panels; these panels are shown on Figure 1.

C. Surrogate Models

In this paper, we aim to compare the performance of different suitable surrogate modeling techniques in approximating C_L, C_D, C_M (lift, drag, and pitching moment coefficients). The surrogate models are constructed in a four-dimensional space with input variables: Mach number (M), angle of attack (α), flight altitude (h), and tail rotation angle (η). Due to the varying magnitudes of the input variables (in particular between the flight altitude and other input variables), the input variables are scaled to be between 0 and 1 prior to constructing and using the surrogate models.

For this purpose, we consider both global models and mixture of experts models. As previously mentioned in Section III, we consider global black-box (data-fit) surrogate models that are interpolative, i.e., kriging and RBF models. For the RBF models, three kernel functions are used, namely the thin-plate splines, cubic, and square-exponential (Gaussian). For the kriging models, since we know that the aerodynamic force and moment coefficients



(a) CRM (conventional) **(b)** BWB (unconventional)

Figure 1: Aircraft configurations considered in this study, showing the grid used in the aerodynamic solver.

are continuously differentiable, we consider only the correlation functions that exhibit parabolic behavior, namely the Gaussian and cubic spline functions. Direct GEK and universal kriging models are also tested.

In a "kriging with a trend" model (universal kriging), the global model is modeled by an analytical expression, which takes different values in space, to be the trend component [24]. Using this approach, instead of being restricted to use low-order polynomials as the basis functions, we select the basis functions that reflect the physics of the system, to assist the prediction. Since we know that drag coefficient profiles are expected to have a steep gradient in the high Mach (M) and high angle of attack (α) region, we can set the trend to be

$$\psi\left(M,\alpha\right)=\begin{cases} 1/\left(1-M^2\right) & \text{if} \quad \alpha\leq 1.0 \\ \alpha^2/\left(1-M^2\right) & \text{if} \quad \alpha\geq 1.0 \end{cases}. \tag{41}$$

The constant numerator for $\alpha \leq 1.0$ is used to remove the quadratic profile (with respect to α) in the low α region, to be consistent with the C_D profile obtained from the aerodynamic solver. The basis function vector, $\mathbf{f}\left(\mathbf{x}\right)$, and the coefficient vector, $\boldsymbol{\beta}$, are thus expressed as follows,

$$\mathbf{f}\left(\mathbf{x}\right)=\left[1,\psi\left(M,\alpha\right)\right] \tag{42}$$

$$\boldsymbol{\beta}=\left[\beta_0,\beta_1\right]^T. \tag{43}$$

Thus at an evaluation point $\mathbf{x}_0=[M_0,\alpha_0,h_0,\eta_0]$, the kriging equation can be expressed as

$$\hat{y}\left(\mathbf{x}_0\right)=\beta_0+\beta_1\psi\left(M_0,\alpha_0\right)+\mathbf{r}\left(\mathbf{x}_s,\mathbf{x}_0\right)^T\mathbf{R}^{-1}\left[\mathbf{y}_s-\beta_0-\beta_1\psi\left(M_0,\alpha_0\right)\right]. \tag{44}$$

The basis function coefficients, β_0 and β_1, are obtained using (22). These specified basis functions are shown to significantly improve the accuracy of kriging models in the high drag gradient area. The additional computational cost in the kriging construction and use with the nonconstant global model is very minimal, in fact it is barely noticeable.

For the mixture of experts models, we follow the procedure presented in Section IV. Ordinary kriging and GEK models with adaptive sampling are used as the local surrogates. The numbers of clustering training data are 100 for the two-dimensional test cases, and 500 for the four-dimensional test cases.

The Halton sampling sequence, which is a space-filling low-discrepancy method [64], is used to generate training samples to construct the surrogate models, as well as to generate the clustering training data for constructing the mixture of experts. The *discrepancy* in this case is the departure of the sampling points from a uniform distribution, thus ensuring an even distribution of samples over the input space. Moreover, Halton sample generation is done in an incremental fashion, meaning when we increase the size of training samples (N_s), we reuse the points from the smaller

sample set. With this incremental sampling, we can compare the surrogate modeling performance with different sizes of sample set more fairly, compared to other sampling method like LHS, which generates a new set of samples for each sample set size.

For the ordinary kriging and GEK models, we also use the adaptive sampling procedure following the "exploitation" infill criterion. At each iteration, we select a point with the maximum index of dispersion, or *variance-to-mean ratio* (VMR), $D = \sigma^2/\mu$, of the kriging prediction as the next sample. Using the maximum VMR instead of the more commonly used maximum variance criterion takes into account the varying magnitudes of kriging predictions at different parts of the input space. To start this procedure, a few initial points are required; in this work, we use 15 Halton points for the global models, and the first 15 points in each clustering data set $\mathcal{T}_k$ for the mixture of experts approach. The adaptive sampling procedure is terminated when the convergence criterion is achieved (maximum VMR $<$ tolerance), or when the specified maximum number of samples (sampling budget) is reached. For simplicity, we select the next sample out of a set of 10 000 candidate points, which are distributed uniformly in the input space. Note that no actual function evaluations are required to compute the VMR values at those points, since the σ^2 and μ of the kriging prediction come out naturally from the kriging derivation and can be expressed analytically (see Section III. B). The actual function evaluation is only required at the selected sample location, to update the sample set $\mathcal{S}$. The reason why we do so is because running a proper optimization (we tried with the particle swarm optimization method or PSO) results in a painfully slower convergence. Thus to achieve the same level of accuracy, it will require significantly more samples than when the sample is selected from a discrete set of candidate points.

To validate the surrogate models, we generate 10 000 truth set data with the aerodynamic solver. These data are used to compute the normalized RMS error, with which we assess and compare the accuracy of surrogate models tested.

VI. Results

In this section, we present and compare the results of using different surrogate modeling techniques, including the proposed mixture of experts approach, to approximate the aerodynamic force and moment coefficients of BWB and CRM configurations. Table 3 summarizes the various surrogate models, with their corresponding model structures and sampling techniques, used in this study. For simplicity, "ordinary kriging" will just be referred to as "kriging" in the subsequent result presentation and discussion.

Model type	**Kernel/correlation function**	**Sampling**
Global models		
Kriging	Cubic (C)	Halton
	Gaussian (G)	Halton
	Gaussian (G)	Adaptive (maximum VMR)
Universal kriging	Gaussian (G)	Halton
GEK	Cubic (C)	Halton
	Gaussian (G)	Halton
	Gaussian (G)	Adaptive (maximum VMR)
RBF	Cubic (C)	Halton
	Gaussian (G)	Halton
	Thin-plate splines (TPS)	Halton
Mixture of experts		
Kriging	Gaussian (G)	Adaptive (maximum VMR)
GEK	Gaussian (G)	Adaptive (maximum VMR)

Table 3: List of surrogate models used in this study and their corresponding model structures and sampling techniques.

We need to construct the surrogate models in a four-dimensional space to perform the surrogate-based mission analysis. The value ranges for these input variables are tabulated in Table 4.

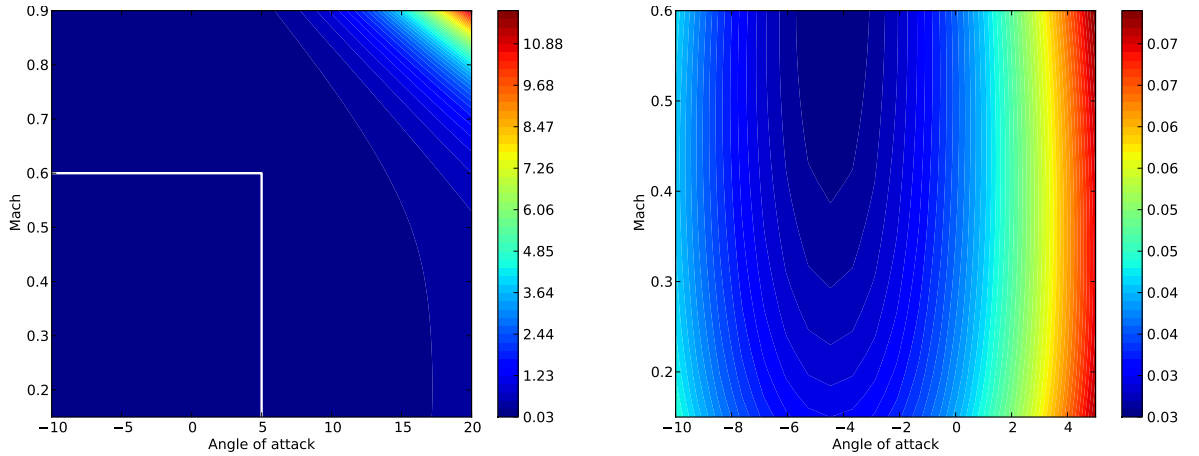
For illustration purposes, we first demonstrate the methods with a two-dimensional case, using data corresponding to the BWB configuration. Next, the results corresponding to the four-dimensional cases are presented for both the BWB and CRM configurations. We then perform the surrogate-based mission analysis using the selected surrogate models for the CRM case.

Input variable	Lower bound	Upper bound
Mach number (M)	0.15	0.90
Angle of attack (α)	$-10.0°$	$20.0°$
Altitude (h)	0 ft	50 000 ft
Tail angle (η)	$-20.0°$	$20.0°$

Table 4: Value ranges for the surrogate models' input variables.

A. Two-dimensional Case with BWB Configuration

For the two-dimensional case, we fix the flight altitude to 38 500 ft and the tail angle to $-7.0°$. For this case, we only discuss the surrogate model performance to approximate C_D, since it exhibits more complex profile than C_L and C_M, especially in the transonic region. Figure 2 shows the C_D contour generated from truth set data. The entire input space is shown in Figure 2a, where we can observe the high drag gradient at the top right corner (high M, high α region). Due to the large value range shown in this contour plot, the lower left corner seems flat. However, when we zoom into the region highlighted by the white rectangle, we see a quadratic profile, as shown in Figure 2b.



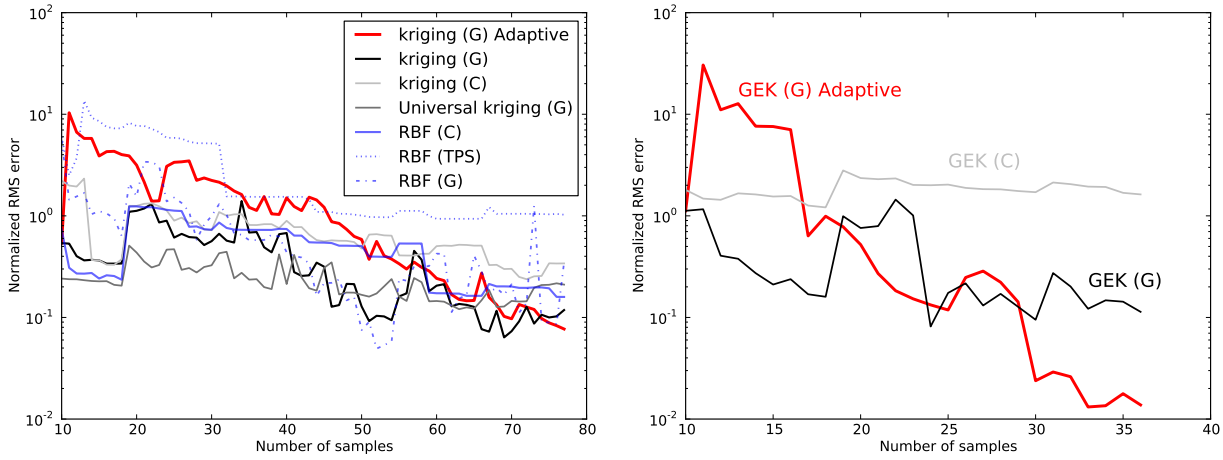
(a) Actual C_D contour showing the high drag gradient

(b) C_D contour subset showing a much lower C_D value range

Figure 2: Drag coefficient contour exhibits different profiles in different input space regions. The white rectangle shown in the left hand side figure highlights the subset shown in the right hand side figure.

Now we look at the performance of global models in approximating this C_D contour. We first build kriging and GEK models with adaptive sampling, and then use the same sample size as the maximum number of Halton samples considered. The convergence criterion used for the adaptive sampling in this case is maximum VMR $\leq 10^{-5}$. The convergence is achieved at $N_s = 77$ for the kriging model, whereas the GEK model requires 36 samples. The normalized RMS errors for different surrogate models that use no gradient information are shown in Figure 3a, and those for GEK models are shown in Figure 3b. From both figures we can observe that the error trend is more monotonically decreasing when adaptive sampling is used. Although using more Halton samples in general decreases the approximation error, the convergence trend is more erratic than when we use adaptive sampling. Among the three RBF models, the one with the thin-plate spline kernel function has the worst performance. Similar performances are observed between kriging and RBF models where the same function is used as the correlation and kernel function (Gaussian or cubic). Universal kriging, using the basis functions given in Section V. C, shows the best performance when fewer samples are used, but is caught up by kriging models (with Gaussian correlation function) as more samples are added. This result shows that adding a known trend to the kriging model does improve the predictive performance, especially when we have a small sample budget. For the GEK models, using a cubic correlation function results in a poor predictive performance. In fact, its performance is worse than when no gradient information is used. GEK models require computing the second derivatives of the correlation function to assemble the extended correlation matrix (to include the correlation between function values and gradients, as well as between gradients). While the second derivatives of a cubic correlation function is continuous, it is only piecewise linear and thus not smooth. The

Gaussian correlation function, on the other hand, has a smooth second derivative.



(a) No gradient information

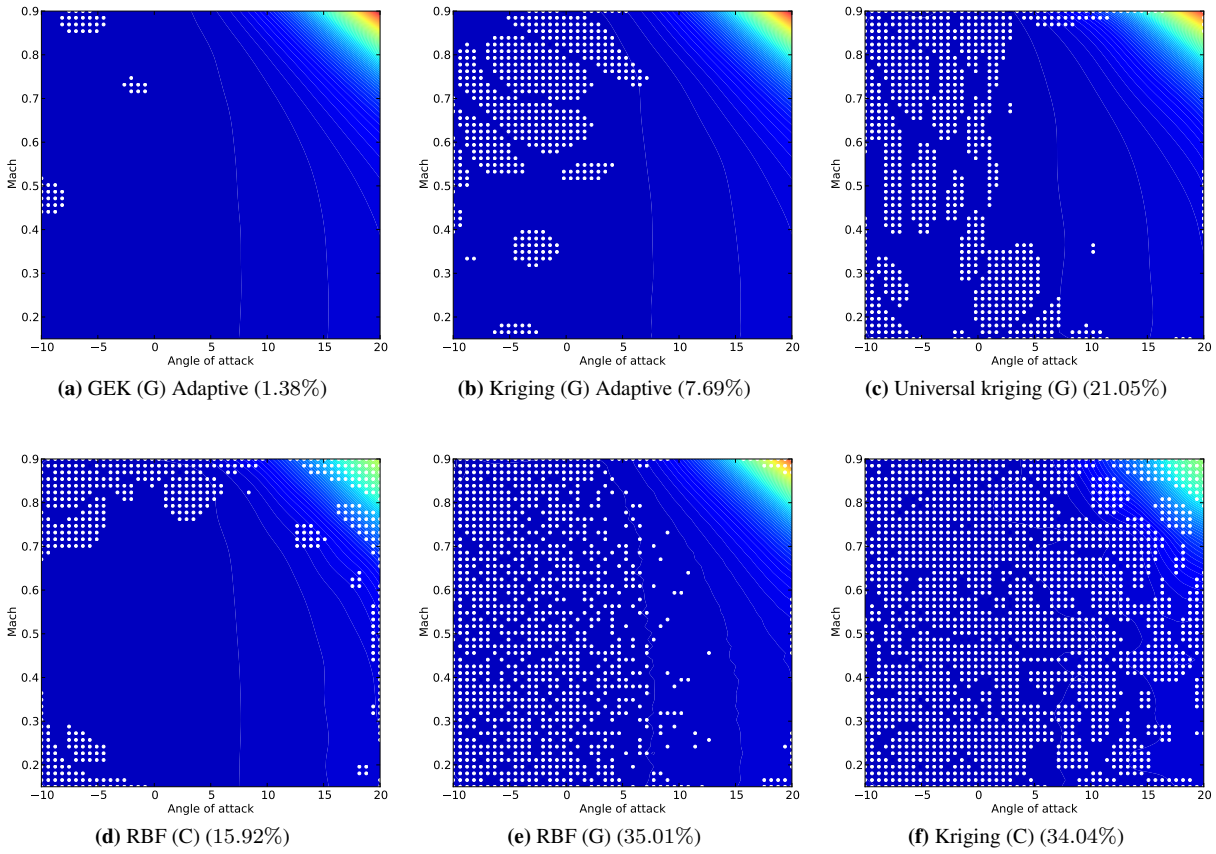
(b) With gradient information

Figure 3: The normalized RMS error plots for the two-dimensional case (BWB) show decreasing trends when adaptive sampling is used, resulting in more accurate surrogate models. Halton sampling is used unless stated otherwise.

The approximated C_D contours of six global models are shown in Figure 4, using the maximum number of samples shown in Figure 3. We indicate the test points with $> 5\%$ approximation errors with white dots, so we can visualize regions within the input space where each surrogate model performs poorly. The GEK model with Gaussian correlation function and adaptive sampling (Figure 4a) shows the best performance, both in terms of the normalized RMS error and the error distribution. The gradient information seems to help fitting the different function characteristics in different input space region significantly. Kriging (Figure 4b) and universal kriging (Figure 4c) models can both follow the trend in the high drag gradient region pretty accurately, but the performance in the low α region is still rather poor, especially for the universal kriging model. A rather similar error distribution is observed when an RBF model with Gaussian kernel function is used (Figure 4e), though the overall normalized RMS error is significantly higher. Choosing a different kernel function affects the RBF model performance, as seen in Figure 4d where a cubic kernel function is used. This model shows an overall good performance, except in the regions that are close to the input space boundary. Kriging with cubic correlation function (Figure 4f) shows a poor predictive performance in the entire input space region, with its normalized RMS error slightly better than the RBF model with Gaussian kernel function. Having reviewed the global model performance, we will now look into the mixture of experts results.

Before generating the mixtures of experts, we first need to determine the mixing proportions $\pi_k(\mathbf{x})$. As mentioned in Section IV. B, we use the modified cluster posterior probability as $\pi_k(\mathbf{x})$ (40), where we need to specify the weight ω. In Figure 5, we show the effect of changing ω on $\pi_k(\mathbf{x})$ (top row), and on the resulting C_D approximation contours (bottom row). For the $\pi_k(\mathbf{x})$ plots, we use different colors to indicate the different clusters. The color intensity within each cluster represents the $\pi_k(\mathbf{x})$ value, where $0 \leq \pi_k(\mathbf{x}) \leq 1$. The lightest color corresponds to $\pi_k(\mathbf{x}) = 0$, whereas the darkest corresponds to $\pi_k(\mathbf{x}) = 1$. We show three ω values, 1.0 (the default value for the original posterior probability), 2.0, and 3.0. For this demonstration, we partition the input space based on the gradient $(\partial C_D/\partial M)$ criterion, and use GEK models as the local experts. As we can observe from these plots, the cluster boundary gets more clearly defined as ω is increased, which increases the sigmoid function slope. When $\omega = 1$, the region in the input space where both local experts "share responsibility" is larger. Consequently, each local expert needs to approximate the function value *beyond* its local area. Since kriging (including GEK) models are not good at extrapolation, this poor predictive performance will be reflected in the overall approximation accuracy, as shown in Figure 5d. Increasing ω decreases the areas outside the local region that each expert needs to predict, resulting in better predictive performance as can be seen from both the error distribution plots and the overall normalized RMS errors. Further increasing ω to be above 3.0 does not affect the predictive performance of the mixture of experts model, as shown in the error convergence plot displayed in Figure 6.

In this two-dimensional study, we consider two clustering criteria, namely the function value (C_D) and the gradient $(\partial C_D/\partial M)$. Both kriging and GEK models are considered as the local experts, and the samples are drawn adaptively. The results are summarized in Figure 7, showing the total number of samples N_s and the overall normalized RMS errors for all cases considered here. When kriging models are used as the local experts, mixture of experts offers quite



(a) GEK (G) Adaptive (1.38%) **(b)** Kriging (G) Adaptive (7.69%) **(c)** Universal kriging (G) (21.05%)

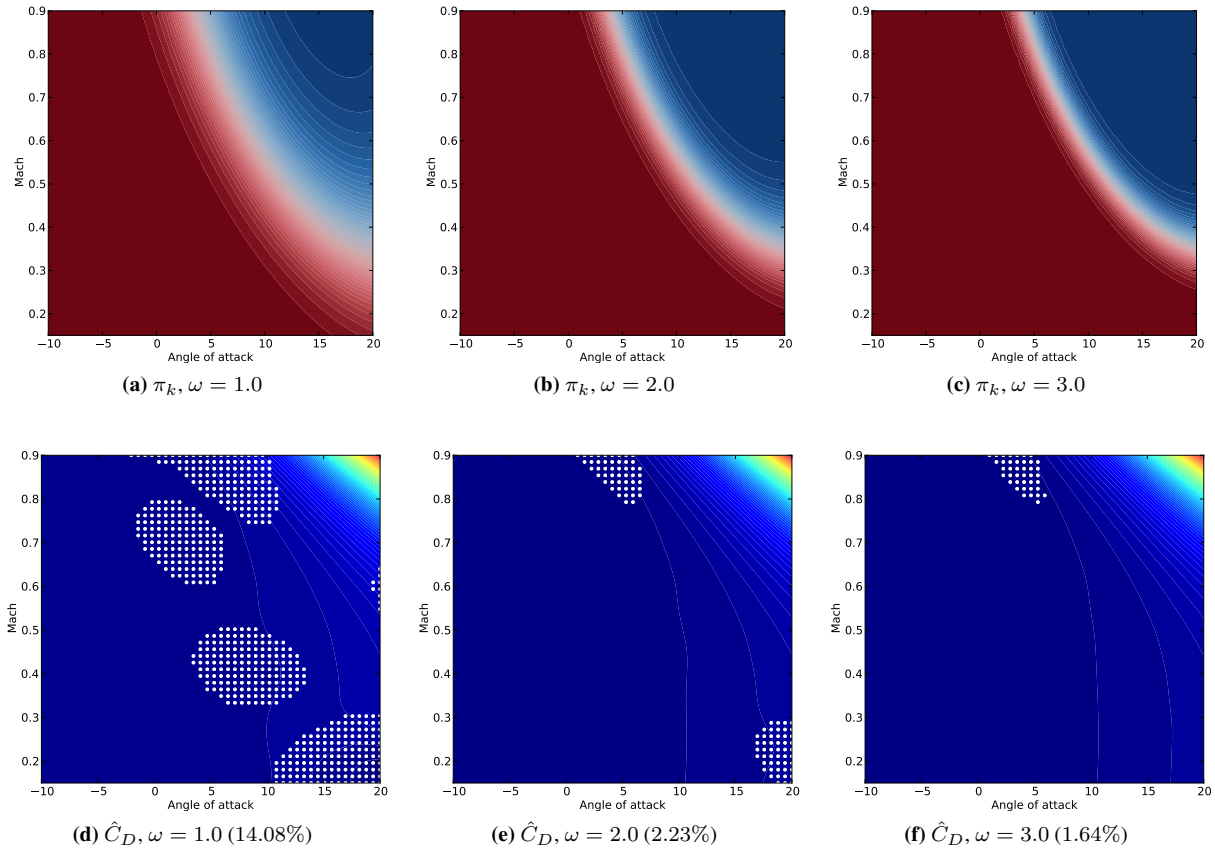
(d) RBF (C) (15.92%) **(e)** RBF (G) (35.01%) **(f)** Kriging (C) (34.04%)

Figure 4: The approximated C_D contours for the BWB configuration in a two-dimensional space from different global surrogate models, with the normalized RMS errors shown inside the brackets. The test points with $> 5\%$ approximation errors are indicated by the white dots.

a significant improvement over the global model, some are achieved with fewer samples. While mixtures of two GEK models perform slightly worse than the global model, using three and four local models result in better predictive performance (with $< 1\%$ overall approximation error) with only slight additions of samples. Comparing between the two clustering criteria, the overall approximation errors are of the same order for each number of clusters considered. However, fewer samples are required when we use $\partial C_D/\partial M$ as the clustering criterion, showing that this gradient value is a better indicator for the heterogeneity in the function profile.

The top row of Figure 8 shows the partitioning of input space (shown as the mixing proportion contour plots) with 2, 3, and 4 clusters when using $\partial C_D/\partial M$ as the clustering criterion. The C_D approximation contours with kriging and GEK models as the local experts are also shown, with the distribution of test points with $> 5\%$ approximation errors shown as white dots. These plots show that the mixtures of GEK models offer a notably better performance than the kriging counterpart.

It is interesting to look at the optimum length scales (kriging hyperparameters, $\boldsymbol{\theta}$) for the various local kriging and GEK models as obtained via the MLE procedure. Table 5 shows the different optimum $\boldsymbol{\theta}$ obtained for each local expert. Each square-bracket corresponds to the $\boldsymbol{\theta}$ of one local expert. The first number refers to the correlation parameter in the M dimension, whereas the second number explains the correlation in the α dimension. A smaller number indicates a stronger correlation. This "correlation" can be interpreted as how much the knowledge of function value at one point helps to deduce the function value at another point. Therefore, a simple linear function has a strong correlation, whereas a highly nonlinear function (e.g., a function which exhibits pronounced oscillations) has a weak correlation. Figure 9 displays the optimum $\boldsymbol{\theta}$ in the different input space partitions corresponding to the mixture of experts with kriging models. For the low M, low α region (blue), we see a stronger correlation in the M dimension than in the α dimension. This outcome is not surprising, as we could see in Figure 2b that C_D values do not vary much in Mach (stronger correlation, lower θ), whereas it exhibits a quadratic profile in α (weaker correlation, higher θ). In the middle region (red), we find weaker correlation in both M and α dimensions, with almost equal length scales. In



(a) $\pi_k, \omega = 1.0$ **(b)** $\pi_k, \omega = 2.0$ **(c)** $\pi_k, \omega = 3.0$

(d) $\hat{C}_D, \omega = 1.0\ (14.08\%)$ **(e)** $\hat{C}_D, \omega = 2.0\ (2.23\%)$ **(f)** $\hat{C}_D, \omega = 3.0\ (1.64\%)$

Figure 5: The effect of changing ω in computing $\pi_k\left(\mathbf{x}\right)$ (40). For the $\pi_k\left(\mathbf{x}\right)$ plots (top row), different colors correspond to different clusters. The highest color intensity within each cluster corresponds to $\pi_k\left(\mathbf{x}\right) = 1$ (maximum value). The overall normalized RMS errors are shown inside the brackets.

the high M, high α region (purple), the correlations are weak but it is stronger in the α dimension. The optimum length scales in the latter partition are the closest to the ones obtained when we use a single global kriging model (see Table 5), suggesting that this is the most dominant profile when fitting a global surrogate model. These observations suggest that partitioning the input space lets each local expert to model the dependence between function value and inputs separately, thus results in a better approximation model overall.

B. Four-dimensional Case with BWB Configuration

Approximating the C_D profile using surrogate models in the four-dimensional space is significantly more complicated than in the two-dimensional space previously discussed. The BWB configuration, in particular, exhibits more correlation between drag and trim, thus the drag profile becomes more nonlinear in the tail angle dimension. We will see how this complex profile imposes challenges in fitting surrogate models that accurately predict the C_D profile in the entire input space.

The adaptive sampling procedures performed for the global kriging and GEK models converge very slowly. We thus set the maximum N_s to be 600 for kriging and 200 for GEK model. The convergence (or the lack thereof) of the maximum VMR, which is the criterion used for the adaptive sampling procedure, and the normalized RMS errors are shown in Figures 10 and 11 for the kriging and GEK model, respectively. Drawing 600 samples adaptively for the kriging model takes almost 20 hours, yet the approximation accuracy is still really poor—the normalized RMS error is 70%. The adaptive sampling procedure for the GEK model takes approximately 31 hours to complete. The resulting approximation accuracy is only 50%, which is really poor. In both cases (kriging and GEK), the maximum VMR converges erratically, though the kriging model starts showing a smoother convergence at $N_s > 450$. However, looking at the convergence slope, it does not seem that the approximation accuracy will greatly improve with adding more samples.

The other global surrogate models are tested with up to 200 Halton samples, since we do not have converged

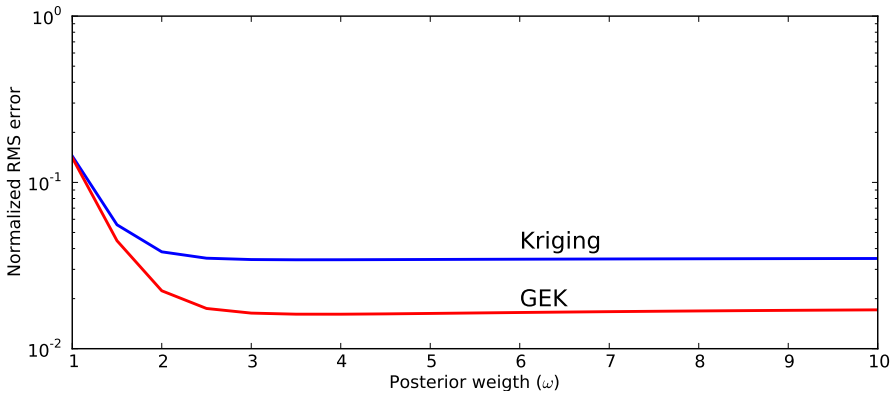


Figure 6: Normalized RMS error converges upon increasing ω in $\pi_k(\mathbf{x})$.

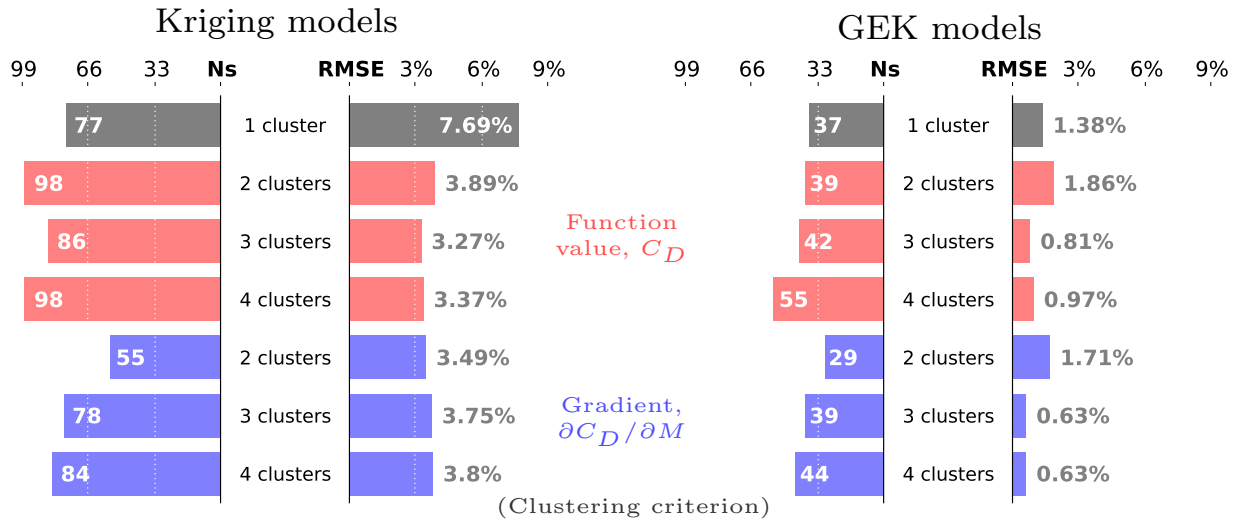


Figure 7: Mixture of experts result summary with two clustering criteria to approximate C_D contour in a two-dimensional space for the BWB configuration.

numbers of samples with the adaptive sampling procedure. The error convergence plots are shown in Figure 12. For the models with no gradient information, the "best" performance is achieved with universal kriging, kriging with cubic correlation function, and RBF model with cubic kernel function, all with Halton samples. Even so, their approximation errors are still around 80%. When gradient information is used, GEK with adaptive sampling shows the "best" performance among the three models. GEK with cubic correlation function has the worst performance, as also observed in the two-dimensional case. Clearly, using any of these surrogate models in any analyses or optimizations will not give us any meaningful results.

Due to the very poor predictive performance of all the global surrogate models considered in this study, we now look into using mixtures of experts. The training samples to build each local expert are selected through the adaptive sampling procedure. For this problem, the convergence is achieved when the maximum VMR $< 10^{-3}$, and the maximum number of samples is set to 40 for each local expert. This adaptive sampling procedure starts with the first 15 clustering training data assigned to the local region, $\mathcal{T}_k$. Similarly to the two-dimensional case, using the gradient, $\partial C_D/\partial M$, as the clustering criterion yields better performance overall with fewer samples than when C_D value is used. Thus, we only show the results from the former clustering criterion here, which are summarized in Figure 13. Here we try partitioning the input space to up to seven clusters. Using kriging models as the local experts result in an average normalized RMS error of approximately 12%, whereas using GEK models further improves it to around 6%. The total number of samples increases as we increase the number of clusters. The approximation error decreases slightly as we add more clusters (up to five clusters), but then it increases. The best performance for

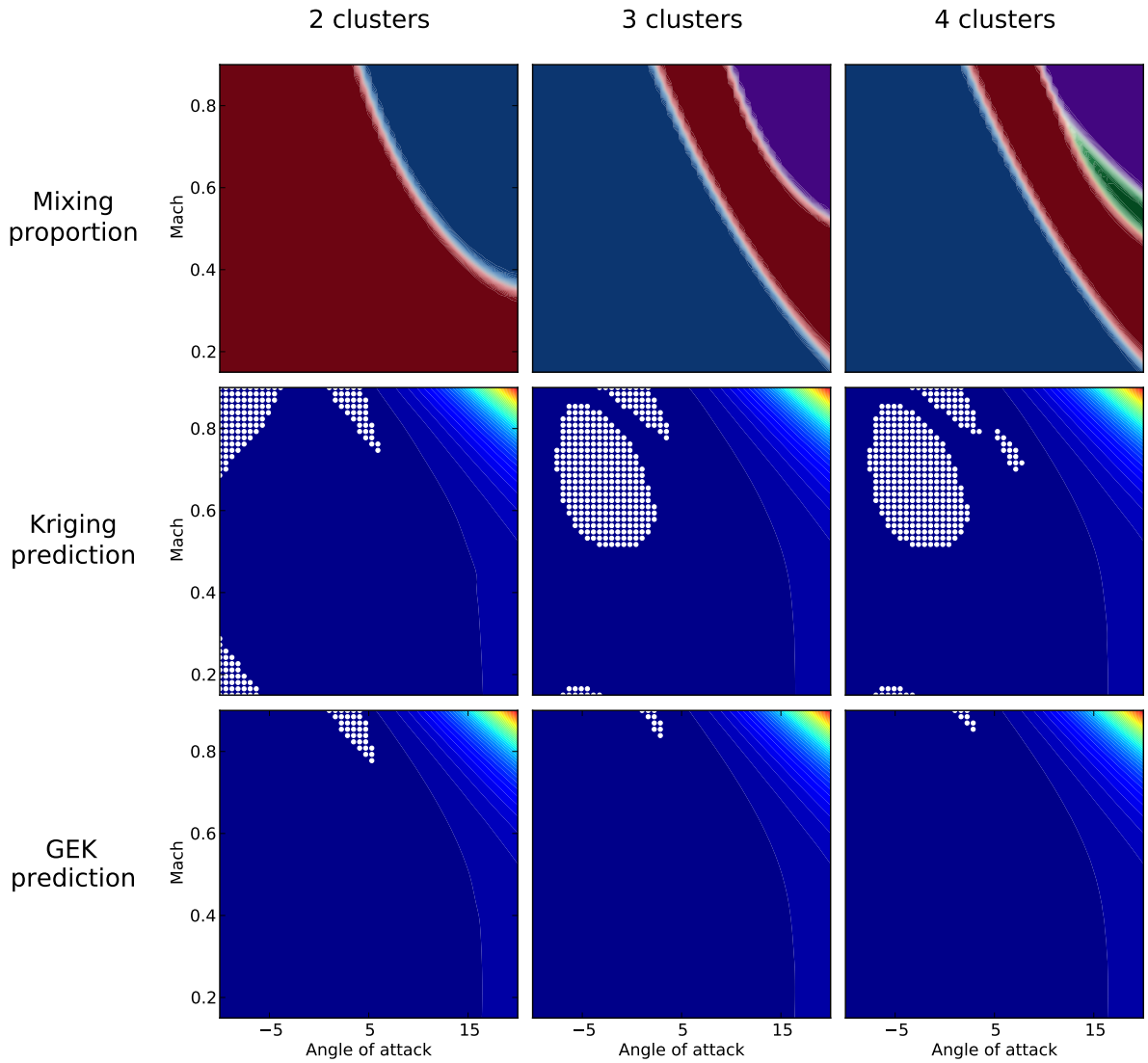


Figure 8: Mixing proportion and C_D approximation contours from the mixture of experts, using $\partial C_D / \partial M$ value as the clustering criterion for the BWB configuration.

both local model types is achieved with 5 clusters, with 10.86% approximation error for kriging and 5.33% for GEK. These results show that applying the divide-and-conquer approach in approximating a complex function profile does improve the predictive performance *significantly*. In this case, each local region gradually resembles a thin strip as we increase the number of clusters when partitioning the input space, which becomes harder for the surrogate to model. Figure 14 visualizes the input partitioning with 6 clusters, by plotting the mixing proportion contours (with different colors representing different local regions) in a two-dimensional slice at $h = 35\,000$ ft and $\eta = 0°$. The thin strips are clearly visible in this plot. Consequently, we see a slight increase in the overall approximation error with 6 and 7 clusters. Compared to the global model training with adaptive sampling procedure that takes more than 20 hours without achieving convergence, the training time for the mixtures of experts (including clustering, adaptive sampling, and constructing the local experts) take less than 3 minutes when using kriging models as the local experts, and less than 8 minutes when using GEK models. This further supports our earlier argument that the adopted distributed approach can help reducing the computational cost to build and use the surrogate models.

Number of clusters	Length scales ($\boldsymbol{\theta}$)
Local experts: kriging models	
1	$[7.61, 1.34]$
2	$[0.08, 1.16]$, $[6.79, 1.81]$
3	$[3.79, 3.56]$, $[0.53, 1.60]$, $[7.58, 3.06]$
4	$[4.89, 4.60]$, $[0.53, 1.60]$, $[6.63, 3.01]$, $[2.84, 3.27]$
Local experts: GEK models	
1	$[10.38, 2.0]$
2	$[0.84, 2.27]$, $[9.34, 1.33]$
3	$[1.25, 0.50]$, $[0.84, 2.27]$, $[9.40, 2.23]$
4	$[1.00, 0.50]$, $[0.84, 2.27]$, $[6.33, 2.54]$, $[9.32, 4.59]$

Table 5: Local kriging and GEK models have different optimum model parameters (length scales $\boldsymbol{\theta} = [\theta_M, \theta_\alpha]$) in the partitioned input space, suggesting that the divide-and-conquer approach is better in modeling the different characteristics in the function profile.

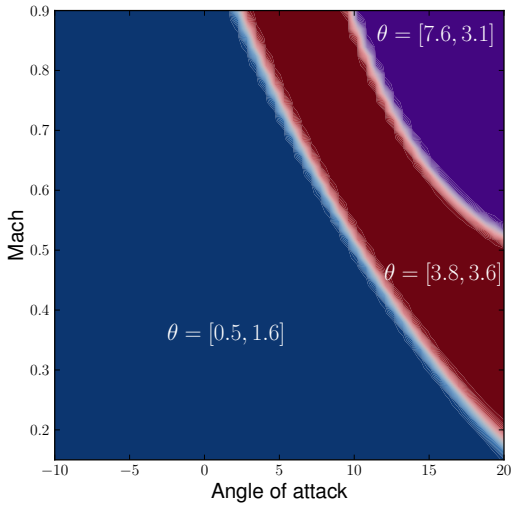


Figure 9: Different local kriging models have notable differences in the optimum model parameters (length scales $\boldsymbol{\theta} = [\theta_M, \theta_\alpha]$).

Figure 15 shows the convergence plots for the maximum VMR and normalized RMS error for each local expert in a mixture of experts with five clusters. Each local expert is built using a GEK model. The convergence displayed in this plot shows a stark difference from those of the global models (Figures 10 and 11). Here, the adaptive sampling procedure within each local expert converges nicely until the convergence criterion is achieved, which translates to a smooth convergence of the normalized RMS error. From these results, we can see that the adopted divide-and-conquer approach overcomes the challenges of modeling a complex terrain by partitioning the input space into smaller subregions which prove to be much easier to tackle.

Building surrogate models for C_L and C_M are much easier than C_D, owing to their much simpler function profiles. Using our aerodynamic solver, C_L and C_M values are independent of the flight altitude, thus their gradients in the altitude dimension are zero. These zero gradients impose difficulties when fitting a GEK model, thus we restrict the following discussion to surrogate models with no gradient information, which will prove to be sufficient in approximating C_L and C_M.

Unlike C_D, performing the adaptive sampling procedure in building a global model results in a nice convergence, for both the C_L and C_M kriging models, as shown in Figure 16. For the C_L kriging model, the maximum VMR decreases to below 10^{-4} with 49 samples. The resulting surrogate model gives an overall approximation error of 1.49%. With the same convergence criterion for the adaptive sampling procedure, the C_M kriging model requires 62

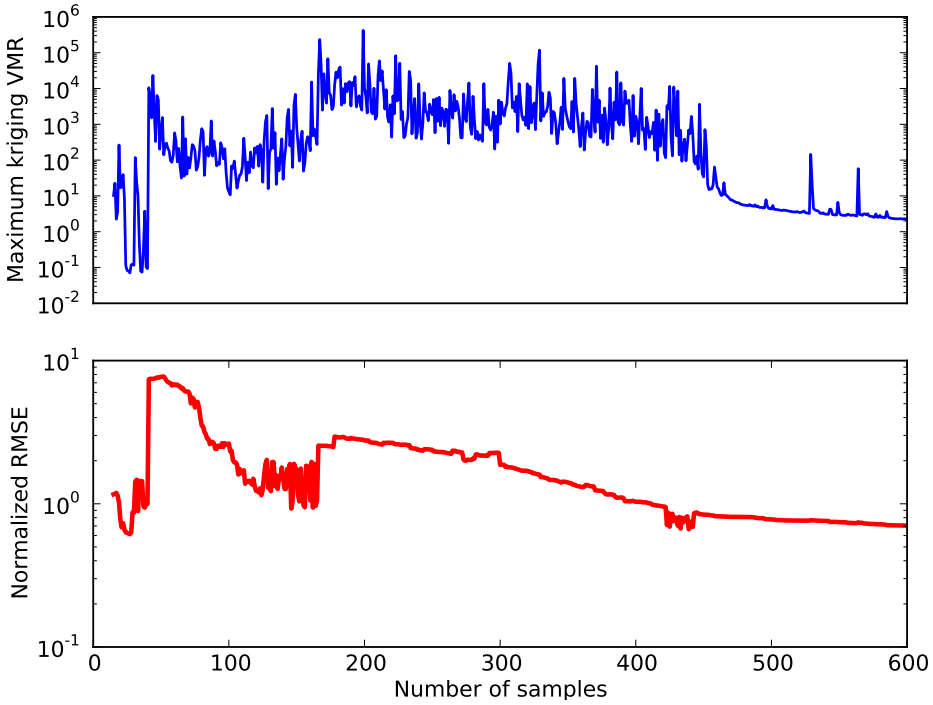


Figure 10: The slow convergence of using a global kriging model (with adaptive sampling) to approximate the complex C_D profile of the BWB configuration in a four-dimensional space. The convergence criterion for the adaptive sampling is not achieved.

samples and achieves an overall approximation error of 1.11%. Using these "converged" numbers of samples, we now run other global models using Halton samples, as shown in Figure 17. In both models, kriging with adaptive sampling offers the best performance. Since C_L and C_M profiles do not exhibit any strong quadratic trends in the M and α dimensions, the universal kriging (which is set to have the same basis functions as the ones for C_D) does not have any advantage over the ordinary kriging models. The kriging models with cubic correlation function show rather poor performance and convergence at $N_s < 35$, but catch up with other kriging models afterwards. The approximation accuracy of the three RBF models converge really slowly, and thus at the selected N_s their approximation errors are still high.

For modeling simple profiles such as C_L and C_M, mixtures of experts do not offer much advantage. In fact, going from using two clusters to three clusters does not show much difference in the input space partitioning. The mixtures of experts result summary for these two function profiles are shown in Figure 18, using C_L and C_M values as the clustering criterion, respectively. The adaptive sampling procedure is performed for each local expert (kriging model). From these results we can see that adding more clusters requires more total samples to build the surrogate models, with no improvement in the approximation accuracy. Therefore, when we deal with simple function profiles, global surrogate models are sufficient and the computational complexity associated with implementing the mixtures of experts is not necessary and should be avoided.

C. Four-dimensional Case with CRM Configuration

The conventional CRM configuration has a simpler C_D profile than that of the BWB in the four-dimensional input space considered here, since drag is not as strongly coupled to trim as it is in BWB. Using the same $10\,000$ input variables to evaluate the truth set data, we find the minimum and maximum C_D values to be 0.006 and 1.174 for the CRM configuration, whereas for the BWB configuration the value range is significantly larger, with a minimum of 0.009 and a maximum of 31.605. Both maximum drag values correspond to the corner cases: $M = 0.9$, $\alpha = 20.0°$, $h = 50\,000$ ft, $\eta = 20.0°$ for CRM and $M = 0.9$, $\alpha = 20.0°$, $h = 50\,000$ ft, $\eta = -20.0°$ for BWB. Consequently, the drag gradient corresponding to the CRM is much lower than that of BWB. We will see the effect of this simpler profile in the predictive performance of surrogate models to approximate C_D of CRM.

In this case, surrogate models with adaptive sampling performs notably better than those with Halton sampling. We will then only discuss about the former in this section. Figures 19 and 20 display the convergence of maximum

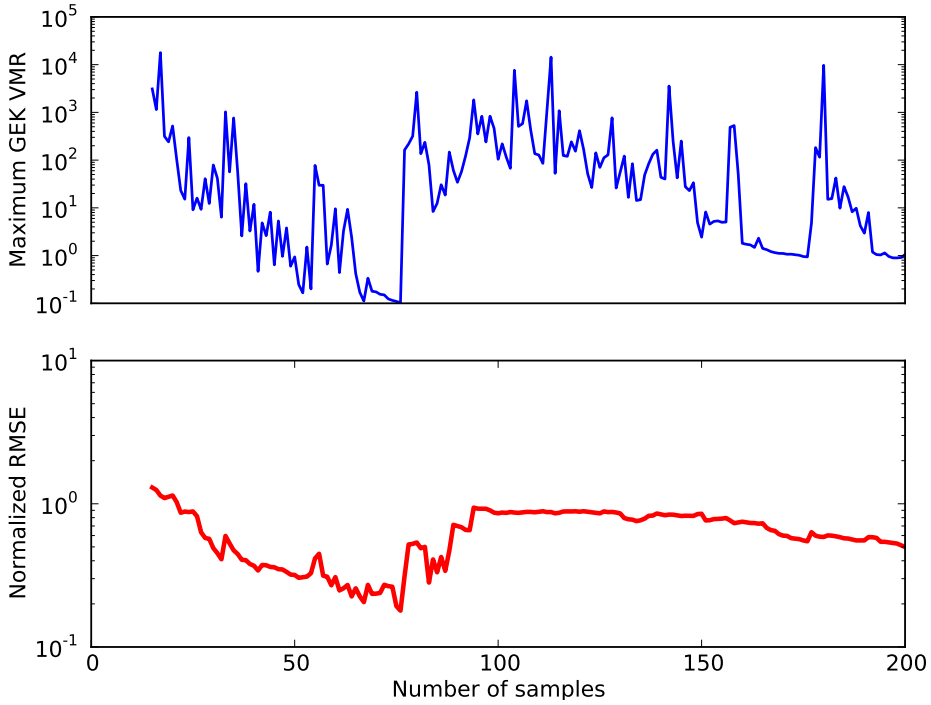
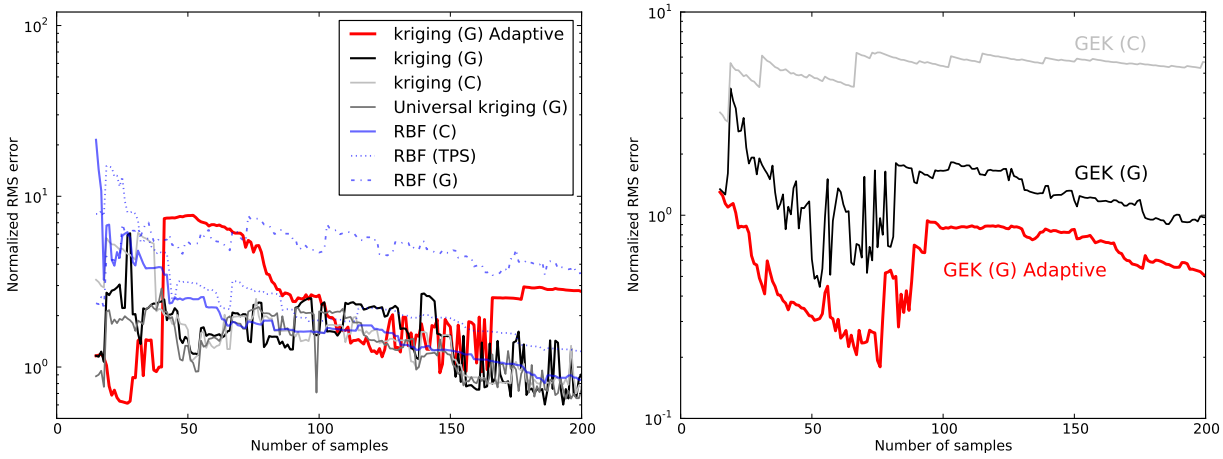


Figure 11: The slow convergence of using a global GEK model (with adaptive sampling) to approximate the complex C_D profile of the BWB configuration in a four-dimensional space. The convergence criterion for the adaptive sampling is not achieved.

VMR and normalized RMS error for both kriging and GEK models. We set the maximum number of samples to be 600 for kriging and 200 for GEK. As the plots show us, the maximum numbers of samples are reached before the adaptive sampling procedures converge (maximum VMR $< 10^{-3}$). At termination, the normalized RMS errors are 14.06% and 15.69% for kriging and GEK, respectively. Another thing we can observe from the convergence plots are the sudden spike of maximum VMR at around 240 samples for kriging and 85 for GEK, which are also reflected in the increasing normalized RMS errors. This phenomenon is common when performing an adaptive sampling procedure based on the maximum variance criterion. The procedure tries to "converge" a certain kriging shape by adding more samples. However, it will reach a point where adding a sample changes the shape it needs to converge to, thus the spike occurs. The kriging shape after the spike is typically more complicated than the one before. In other words, there is a certain profile characteristic that is only captured by the model with "enough" samples. We have seen this phenomenon happens in many cases, even in simpler analytical functions, and they typically converge in the end. Having multiple spikes in the convergence plot is not uncommon either. However, when the function profile is too complex, the procedure converges too slowly, just like the case we observe here.

We now implement the mixture of experts approach to approximate this C_D profile. Both kriging and GEK models are considered as the local experts. The clustering criterion is the $\partial C_D/\partial M$ values, which is previously used in the BWB case. An adaptive sampling procedure is performed for each local expert, where convergence is achieved when maximum VMR $< 10^{-3}$. The results are summarized in Figure 21. With kriging as the local experts, we do not see much improvement in terms of the approximation accuracy as compared to the global kriging model, with approximation errors of around 16%. However, the adaptive sampling procedures converges for all local experts, with a maximum total N_s of 128 (with six clusters). Using GEK models as the local experts, on the other hand, shows a notable improvement in terms of the approximation accuracy, achieving overall normalized RMS errors of approximately 4%. The total N_s increases with more clusters. A good compromise between the approximation accuracy and N_s is achieved with four clusters, which requires 72 samples in total and yields a 3.75% approximation error.

Similarly to the BWB case, simple global kriging models with adaptive sampling offer good predictive performance to approximate C_L and C_M of the CRM configuration. The convergence plots of the maximum VMR and normalized RMS error are shown in Figure 22. We achieve a normalized RMS error of 2.39% with 53 samples in the C_L approximation, and 3.36% error with 80 samples in the C_M approximation.



(a) No gradient information **(b)** With gradient information

Figure 12: The normalized RMS error plots show poor predictive performance by all the global surrogate models considered in approximating the C_D profile of the BWB configuration in a four-dimensional space. Halton sampling is used unless stated otherwise.

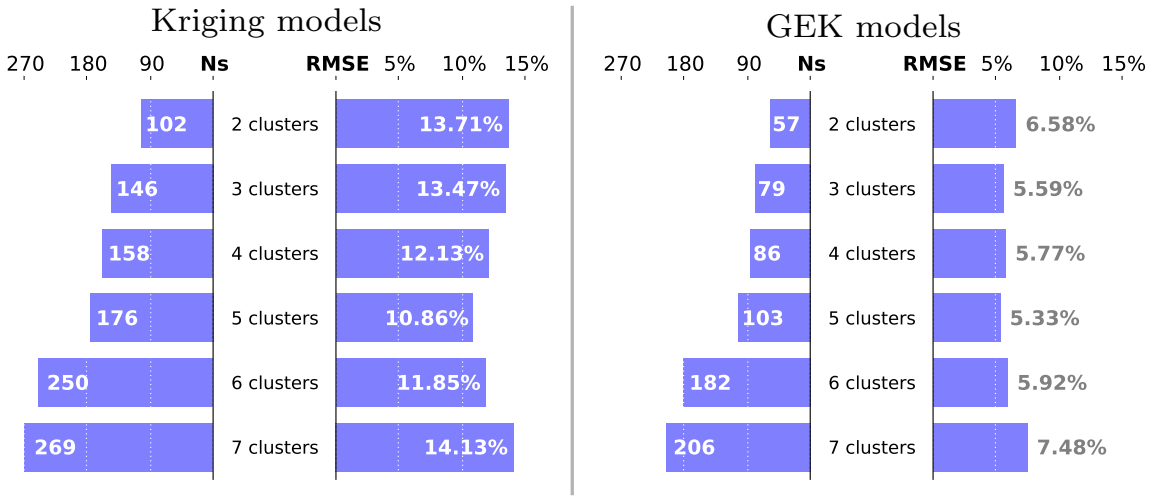


Figure 13: Mixture of experts result summary with $\partial C_D/\partial M$ clustering criterion to approximate the C_D profile of the BWB configuration.

D. Surrogate-based Mission Analysis with CRM Configuration

We now demonstrate using the derived surrogate models on the surrogate-based mission analysis procedure described in Section II. The conventional CRM configuration is used in this demonstration, where global kriging models are used to approximate C_L and C_M, whereas a mixture of experts with four GEK models is selected for the C_D approximation.

Here we consider a sample mission profile as described in Table 6, which is illustrated in Figure (not to scale) 23. In this simple case, we only assume one main profile, without any loiter and reserve profiles. The entire cruise portion of the flight, from the initial cruise (segment 8), through the step climb (segment 9), to the final cruise (segment 10), is done at a constant Mach number. From 10 000 ft, the climb is done at a constant knots indicated airspeed (KIAS) (segment 6), until it intercepts the desired cruise Mach number, at which point the climb is done at a constant Mach number (segment 7). The descent is also done in a similar fashion, with a constant Mach descent (segment 11) followed by a constant KIAS descent (segment 12). This procedure is implemented to reflect common operational procedures. The fuel fraction values, ζ, used in this work are listed in Table 7, following those suggested by Roskam [5], Raymer [104],

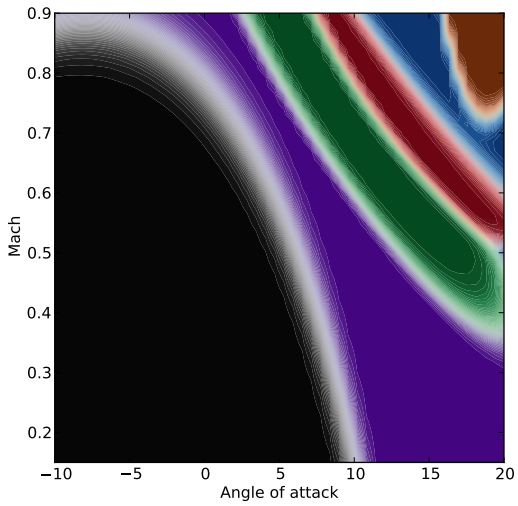


Figure 14: The input partitioning results in thiner strips as we increase the number of clusters, as shown here on a two-dimensional slice of the four-dimensional space, at $h = 35\,000$ ft and $\eta = 0^\circ$. As each local subregion gets thinner, training a local expert gets harder, which might increase the overall approximation errors as we add more clusters.

and Sadraey [105]. For this mission we specify the payload to be 20 tonne and the flight range to be 3 000 nmi. The numerical integration is performed with four intervals per segment.

	Segment	**Altitude** ft	**Speed** Mach/kt	**Range** nmi	**Time** min	**Fuel Burn** (kg)
1	Startup	-	-	-	-	2187.03
2	Taxi	-	-	-	-	2165.16
3	Takeoff	-	-	-	-	1071.75
4	Climb	1 500 → 10 000	150 KIAS → 250 KIAS	27.98	8.41	1991.40
5	Cruise	10 000 → 10 000	250 KIAS → 310 KIAS	4.90	0.94	221.66
6	Climb	10 000 → 23 731	310 KIAS	39.00	6.10	1390.84
7	Climb	23 731 → 30 000	M0.72	13.39	1.87	493.59
8	Cruise	30 000	M0.72	1396.12	197.10	24504.12
9	Climb	30 000 → 32 000	M0.72	4.24	0.60	141.83
10	Cruise	32 000	M0.72	1396.12	198.83	22594.45
11	Descent	32 000 → 23 731	M0.72	15.55	2.19	254.53
12	Descent	23 731 → 10 000	310 KIAS	27.75	4.13	510.54
13	Cruise	10 000	310 KIAS → 250 KIAS	42.34	7.70	759.84
14	Descent	10 000 → 1 500	250 KIAS → 150 KIAS	32.61	8.66	664.77
15	Landing	-	-	-	-	1597.52
Mission Total				3 000.0	436.54	60 549

Table 6: Mission profile parameters and the mission analysis results (range, time, and fuel burn for each segment).

The drag coefficients approximated by kriging models are only inviscid drag coefficients. To obtain the total drag coefficient, we add a constant viscous drag coefficient of 0.0136. This viscous drag is pre-computed based on a flat-plate turbulent skin friction estimate with form factor corrections. In this simple case study, we assume a constant TSFC (0.53 lb/(lbf · h)), instead of using an engine model. We use a weight and balance model with four components, namely the mission payload (20 tonne), fixed weight (100.9 tonne), wings (37.2 tonne), and fuel weight, which depends on the mission analysis. These component weights and moments gives an estimate of the entire aircraft's weight, as well as the nominal, forward, and aft center of gravity (CG) locations. During the mission analysis, the weight and CG locations of these components can be individually updated, giving a more accurate picture of the aircraft's weight and balance as fuel is decremented in the integration. The mission analysis results, including the range, time, and the amount of fuel burned for each segment, are also listed in Table 6.

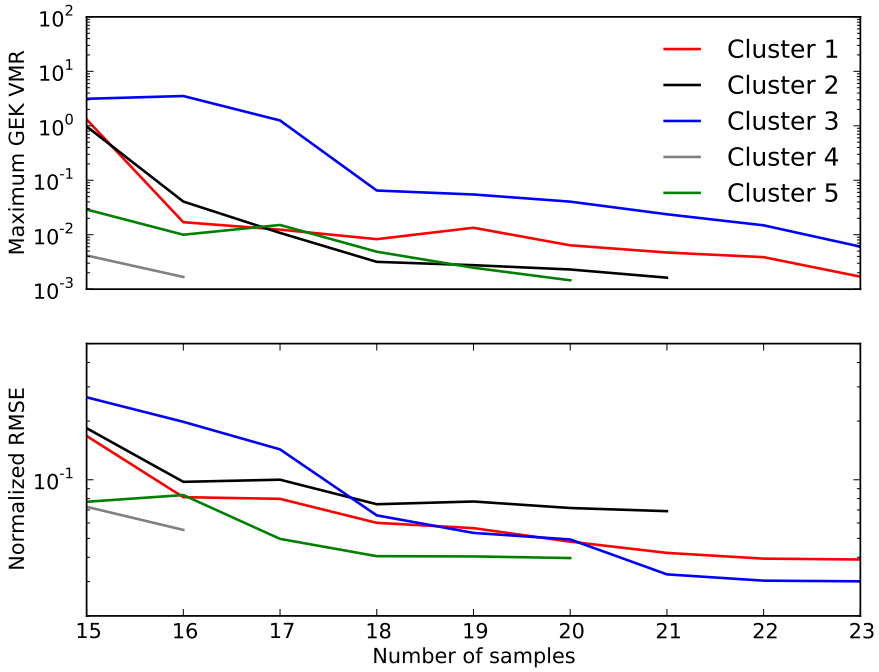


Figure 15: The convergence plots for the maximum VMR and normalized RMS error for each local expert (GEK), with the input space partitioned into 5 clusters, when approximating the C_D profile of the BWB configuration in a four-dimensional space.

Segment	Startup	Taxi	Takeoff	Landing
Fuel fraction, ζ	0.01	0.01	0.005	0.01

Table 7: Fuel fraction values.

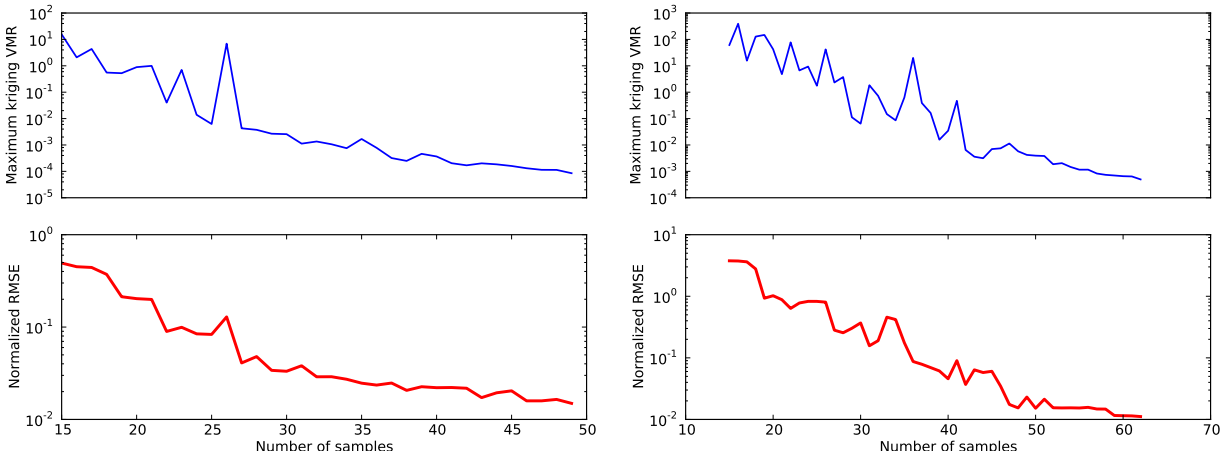
Solving this mission requires a total of 6.6 millions of function evaluations. This number comes from the product of the number of secant iteration to obtain the desired range, number of residual equations in the mission analysis, number of numerical integration intervals, and the number of Newton search algorithm (stabilized with a backtracking line search algorithm) performed at each interval to find α and η. This mission analysis would be computationally intractable should we use an aerodynamic solver for each function evaluation. Table 8 shows a significant computational gain when we use surrogate models instead of the actual aerodynamic performance evalution ($\mathcal{O}\left(10^4\right)$ faster when we use a mixture of experts and $\mathcal{O}\left(10^6\right)$ faster with a global kriging/GEK model). The computational expense of using

`TriPan` solver	Mixture of experts	Kriging/GEK
35 s	0.001 s	2×10^{-5} s

Table 8: Computational time for one function evaluation using one processor.

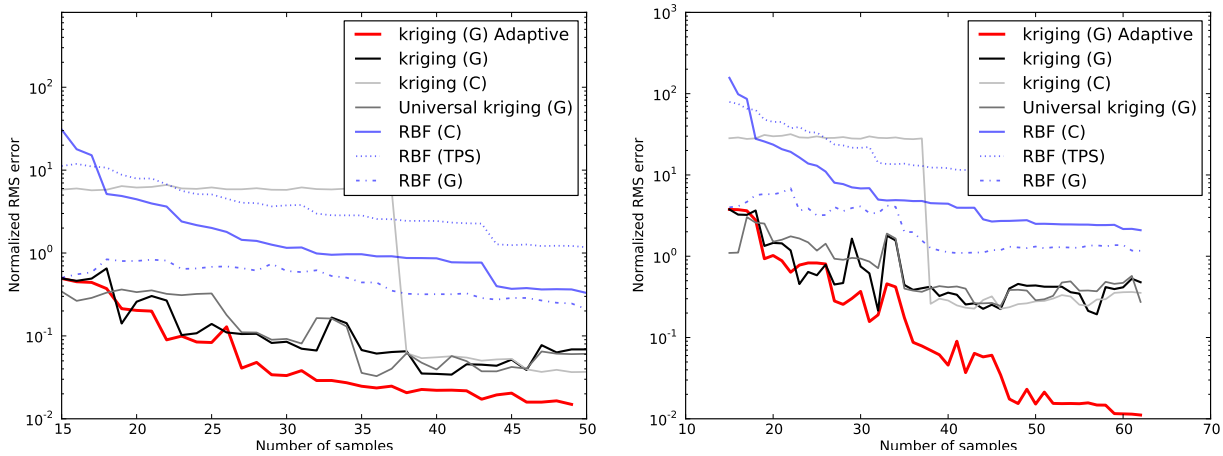
an aerodynamic solver would be exacerbated when we consider multiple missions, use higher-fidelity models, e.g., by solving the Euler or the Reynolds-averaged Navier Stokes (RANS) equations, or a coupled aerostructural model. Performing an optimization with such a high computational cost would be very challenging, if not impossible.

In Figure 24, we visualize the distribution of points in the four-dimensional surrogate model input space that are evaluated during the mission analysis. In this scatterplot matrix, the four-dimensional space is deconstructed into its various two-dimensional projections. Instead of showing all the 6.6 millions points, here we only show the evaluation points after the Newton search algorithms converge (i.e., only the points where C_D's are evaluated), which amount to around 40 thousands points. For the future work, we can use such information to improve the sample



(a) $\hat{C}_L$ **(b)** $\hat{C}_M$

Figure 16: The convergence plots for the maximum VMR and normalized RMS error for kriging model with adaptive sampling to approximate C_L and C_M profiles of the BWB configuration in a four-dimensional space.



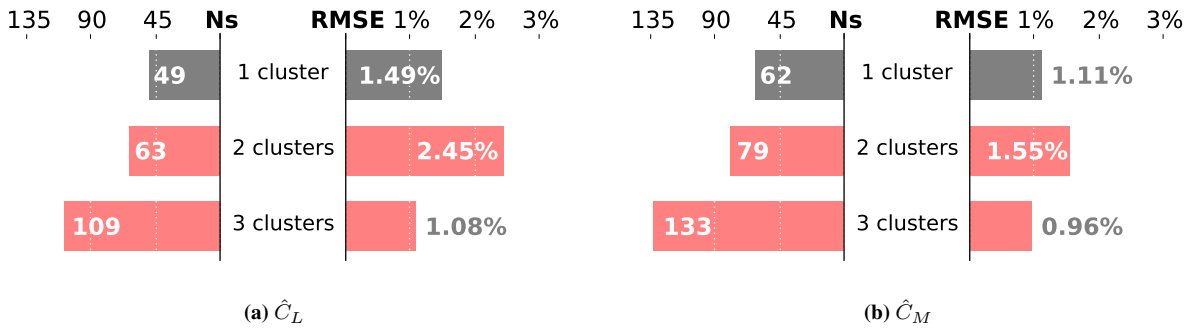
(a) $\hat{C}_L$ **(b)** $\hat{C}_M$

Figure 17: The convergence plots for the normalized RMS errors for global models (with no gradient information) to approximate C_L and C_M profiles of the BWB configuration in a four-dimensional space.

selection in the surrogate model training, to concentrate the samples around the areas where they are used in the mission analysis. We need to take note, however, that the evaluation point distribution would cover larger areas if we include the points evaluated during the Newton search algorithm, or when we perform mission optimizations where some mission parameters (e.g., cruise Mach number and altitude) are now design variables and varied throughout the optimization procedure. With a better sampling strategy, we believe that the predictive performance of the surrogate models can be improved. Ideally, we want to have the same set of sample locations for all C_L, C_D, and C_M surrogate models, to reduce the number of actual function evaluations required to build the models.

VII. Conclusion

The predictive performance of various surrogate models in approximating aerodynamic force and moment coefficients for the conventional (CRM) and unconventional (BWB) configurations have been presented and compared. These surrogate models are to be used in a surrogate-based mission analysis to compute fuel burn of a flight mission in detail, which can be employed in mission and aerostructural optimizations. For this purpose, we need globally accurate surrogate models to approximate C_L, C_D, and C_M in a four-dimensional space of Mach number, angle of



(a) $\hat{C}_L$ **(b)** $\hat{C}_M$

Figure 18: Result summary for using mixtures of experts to approximate the C_L and C_M profiles of the BWB configuration in a four dimensional space.

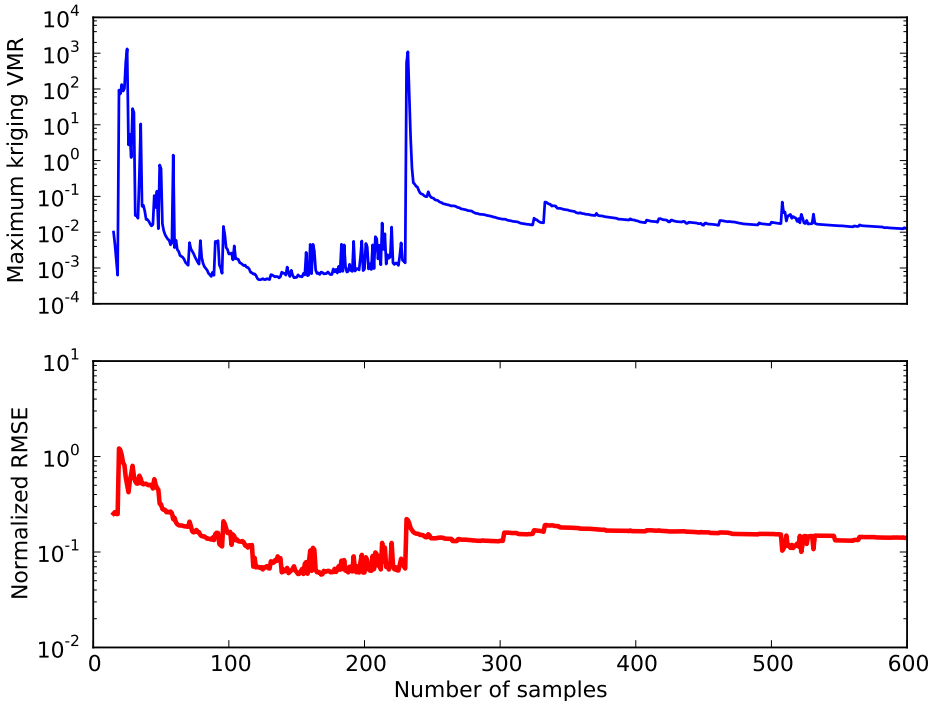


Figure 19: Using a global kriging model with adaptive sampling to approximate the C_D profile of the CRM configuration in a four-dimensional space results in 14.06% approximation error, though the convergence criterion for the adaptive sampling procedure is still not achieved.

attack, flight altitude, and tail rotation angle. In addition to using the well-established kriging and RBF techniques to build the global models, we propose a mixture of experts approach, which adopts the divide-and-conquer principle. In this approach, different local experts are responsible for modeling the different subregions in the input space. Their predictions are then combined, weighted by input-dependent mixing proportions, to yield the final prediction. To generate the training data used in constructing the surrogate models, an adaptive sampling procedure is also performed, in addition to the fixed Halton sequence sampling. To validate the models, we generate truth set data and compute the normalized RMS errors.

Our results suggest that the surrogate model performance is problem-dependent. Even with the same surrogate modeling technique, different model structures and parameters perform differently when applied to different problems. For example, the Gaussian correlation function performs better than the cubic one when modeling C_D in a two-dimensional space (for both kriging and RBF models), whereas the reverse is true when the input space is four-dimensional. In Section III. C, we show that a kriging model can be reduced to an RBF model when we simplify the global model and model hyperparameters. In our two-dimensional case study, we observe that kriging and RBF

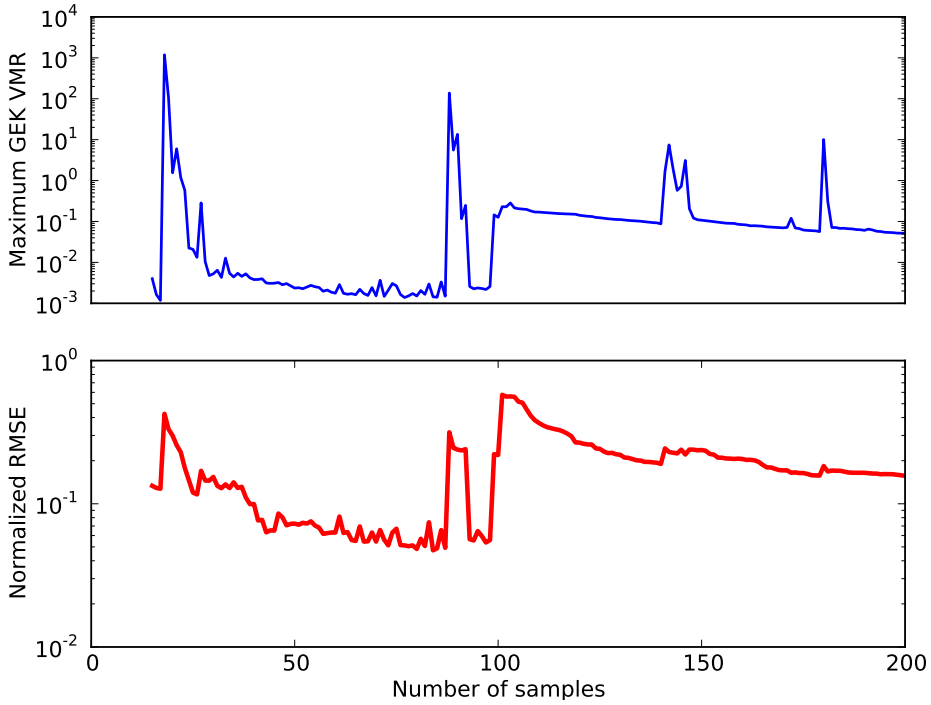


Figure 20: Using a global GEK model with adaptive sampling to approximate the C_D profile of the CRM configuration in a four-dimensional space results in 15.69% approximation error, though the convergence criterion for the adaptive sampling procedure is still not achieved.

models exhibit similar performance when the same kernel/correlation function is used (cubic or Gaussian). This result suggests that for simple problems, RBF might be the better choice since it is simpler to construct than kriging, while yielding similar performance. Adding a known trend in the kriging model (universal kriging) has been shown to improve modeling efficiency, i.e., it requires fewer samples to achieve the same level of accuracy as the ordinary kriging. Results from GEK models show that interpolating the gradients in addition to function values at sample points does improve the predictive performance significantly. However, it comes at a higher computational cost, as we require the gradient computation. An efficient adjoint method [106, 107] needs to be used when computing gradients of high-fidelity models. The size of the linear system of equations to be solved is also inevitably larger. The adaptive sampling procedure helps improving the accuracy of surrogate models; however, the convergence could be very slow in some cases, in particular when modeling complex profiles. Therefore, with smaller sample budget, a simple space-filling sampling technique seems to be a better option.

While the global surrogate models perform well in the two-dimensional case and in modeling C_L and C_M in the four-dimensional cases, they prove to be insufficient to model the complex profile of C_D in a four-dimensional space, in particular with the unconventional BWB configuration. Significant improvements are observed when we use the proposed mixture of experts approach. The divide-and-conquer approach overcomes the challenges of modeling a complex terrain by partitioning the input space into smaller subregions, each with a simpler profile to model. For the four-dimensional case with BWB configuration, this approach achieves a 5.33% approximation error with a mixture of 5 GEK models (103 samples), and 10.86% when kriging models are used as the local experts (176 samples). On the other hand, the adaptive sampling procedures for the global kriging and GEK models fail to converge yielding 70% and 50% approximation errors at termination. Moreover, the distributed approach in the mixture of experts notably helps reducing the computational cost to build and use the surrogate models. The training times for the global kriging and GEK models (with adaptive sampling) are 20 and 30 hours (and yet they still fail to converge); these numbers are reduced to 3 and 8 minutes when using the mixtures of experts. In these case studies, each local expert in the mixtures finds different optimum model parameters, which shows that by partitioning the input space, each local expert models the dependence between function value and inputs separately, yielding a better approximation overall. While in our case studies the same model type is used for all local experts, the proposed mixture of experts approach offers the flexibility of using different models types, e.g., using RBF models in simpler subregions and GEK models to model more complex terrains. The derived mixing proportions can still be used in such a case. The advantages of using the

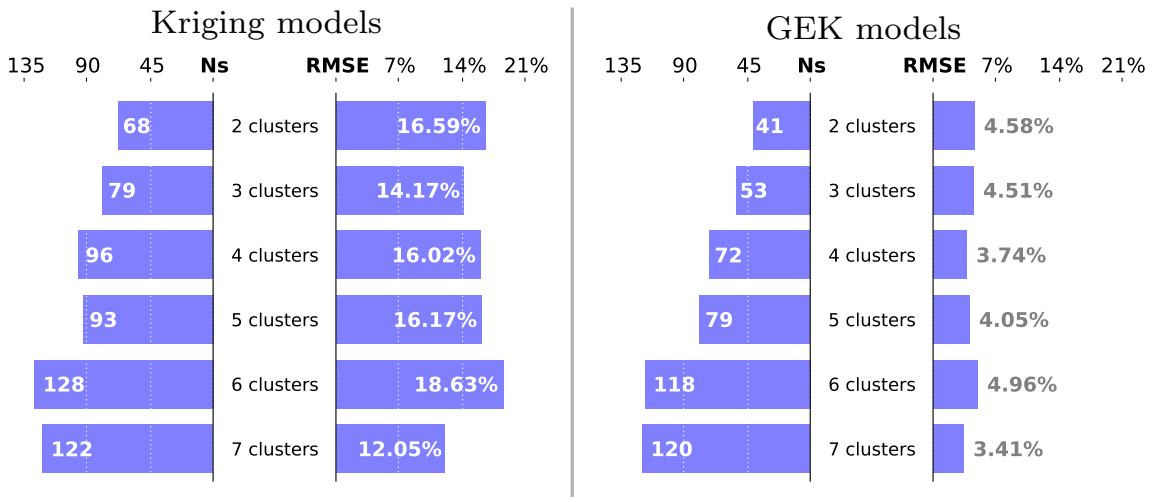
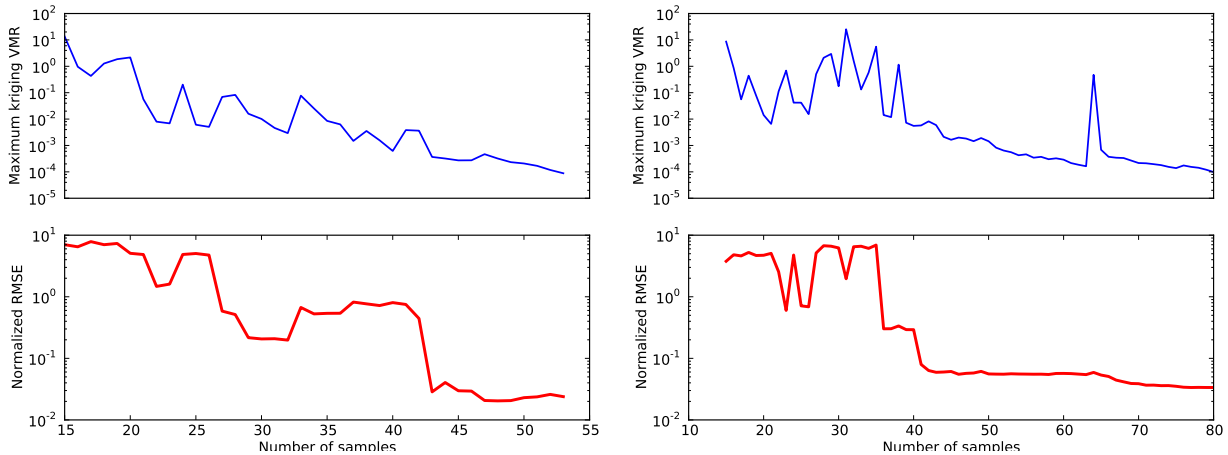


Figure 21: Mixture of experts result summary with $\partial C_D/\partial M$ clustering criterion to approximate the C_D profile of the CRM configuration in a four-dimensional space.



(a) $\hat{C}_L$ **(b)** $\hat{C}_M$

Figure 22: The convergence plots for the maximum VMR and normalized RMS error for kriging model with adaptive sampling to approximate C_L and C_M profiles of the CRM configuration in a four-dimensional space.

mixture of experts approach comes with added computational complexity and more parameters to tune, such as the number of clusters and the clustering criterion. Applying the principle of parsimony, it is wise to use simple global models whenever sufficient, to avoid the unnecessary complexity that is inherent in the mixture of experts approach.

We demonstrated using the derived surrogate models in the surrogate-based mission analysis for the CRM configuration, where we use a mixture of experts to approximate C_D, and kriging models to approximate C_L and C_M (with adaptive sampling). Using this detailed fuel burn computation procedure, solving one mission requires 6.6 millions function evaluations, which would be computationally prohibitive without using surrogate models. With our approach, the number of aerodynamic performance evaluations is dramatically reduced to the number of samples required to build the surrogate models (including the clustering training data should mixtures of experts are used). When we use this mission analysis procedure in an optimization problem, we can fix the sample locations. Therefore, the clustering algorithm and the adaptive sampling procedure only need to be performed once, prior to the optimization. At each optimization iteration, we draw samples only at the predetermined sample locations.

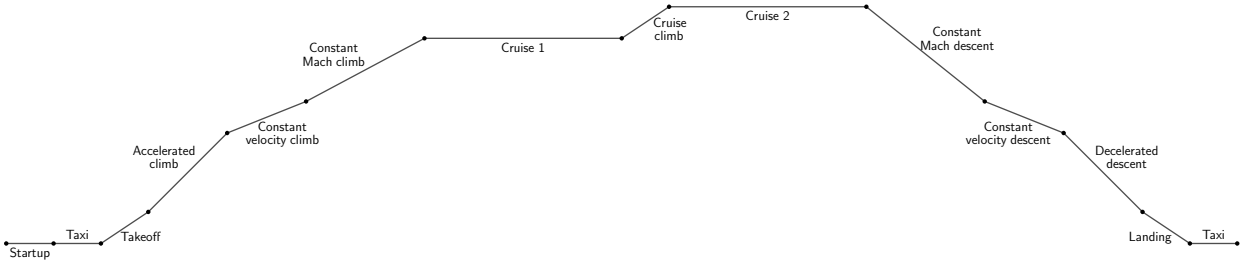


Figure 23: Typical mission profile for a long-range configuration.

As the next step, we need to assess how the surrogate models' levels of accuracy translate to the accuracy in the fuel burn computation with the mission analysis. This information could in turn guide us in the surrogate model selection and training, to achieve the desired accuracy in the optimization results. We will also look into improving the sample selection by using the information on the distribution of points in the input space that are evaluated in the mission analysis procedure. We believe that focusing the sample distribution around this important area could further improve the accuracy of surrogate models, while reducing the number of samples required to train them.

Acknowledgments

The authors are grateful for the funding provided by the Vanier Canada Graduate Scholarships. The computations were performed on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; the Ontario Research Fund—Research Excellence; and the University of Toronto. The authors would like to recognize the other members of our research group, especially Charles Mader, Gaetan Kenway, Graeme Kennedy, Edmund Lee, and Peter Lyu for their contributions to the solvers and framework we use in this work.

References

[1] Lee, J. J., "Can we accelerate the improvement of energy efficiency in aircraft systems?" *Energy Conversion and Management*, Vol. 51, 2010, pp. 189–196.

[2] Nidumolu, R., Prahalad, C. K., and Rangaswami, M. R., "Why Sustainability Is Now the Key Driver of Innovation," *Harvard Business Review*, Vol. 87, No. 9, 2009, pp. 56–64.

[3] Lee, J. J., *Historical and Future Trends in Aircraft Performance, Cost, and Emissions*, Master's thesis, Aeronautics & Astronautics Department and Technology & Policy Program, Massachusetts Institute of Technology, September 2000.

[4] Randle, W. E., Hall, C. A., and Vera-Morales, M., "Improved Range Equation Based on Aircraft Flight Data," *Journal of Aircraft*, Vol. 48, No. 4, July–August 2011, pp. 1291–1298. doi:10.2514/1.C031262.

[5] Roskam, J., *Airplane Design Part I: Preliminary Sizing of Airplanes*, Roskam Aviation and Engineering Corporations, Ottawa, KS, 1985.

[6] Yan, B., Jansen, P. W., and Perez, R. E., "Multidisciplinary Design Optimization of Airframe and Trajectory Considering Cost and Emissions," *14th AIAA/ISSMO Multidisciplinary Analysis and Optimization (MAO) Conference*, Indianapolis, IN, September 2012. doi:10.2514/6.2012-5494, AIAA 2012-5494.

[7] Simpson, T. W., Booker, A. J., Ghosh, D., Giunta, A. A., Koch, P. N., and Yang, R. J., "Approximation methods in multidisciplinary analysis and optimization: a panel discussion," *Struct Multidisc Optim*, Vol. 27, 2004, pp. 302–313.

[8] Simpson, T. W., Toropov, V., Balabanov, V., and Viana, F. A. C., "Design and Analysis of Computer Experiments in Multidisciplinary Design Optimization: A Review of How Far We Have Come—or Not," *12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Victoria, BC, Canada, September 2008. doi:10.2514/6.2008-5802, AIAA 2008-5802.

[9] Sobieszczanski-Sobieski, J. and Haftka, R. T., "Multidisciplinary aerospace design optimization: survey of recent developments," *Structural Optimization*, Vol. 14, 1997, pp. 1–23. doi:10.1007/BF01197554.

[10] Chung, H. S. and Alonso, J. J., "Design of a Low-Boom Supersonic Business Jet Using Cokriging Approximation Models," *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Atlanta, GA, September 2002, AIAA Paper 2002–5598.

[11] Chung, H. S. and Alonso, J. J., "Using Gradients to Construct Cokriging Approximation Models for High-Dimensional Design Optimization Problems," *9th AIAA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Reno, NV, January 2002, AIAA Paper 2002–0317.

[12] Toal, D. J. J. and Keane, A. J., "Efficient Multipoint Aerodynamic Design Optimization via Cokriging," *Journal of Aircraft*, Vol. 48, No. 5, September–October 2011, pp. 1685–1695. doi:10.2514/1.C031342.

[13] Zimmermann, R. and Görtz, S., "Non-linear reduced order models for steady aerodynamics," *Procedia Computer Science*, Vol. 1, 2010, pp. 165–174. doi:10.1016/j.procs.2010.04.019.

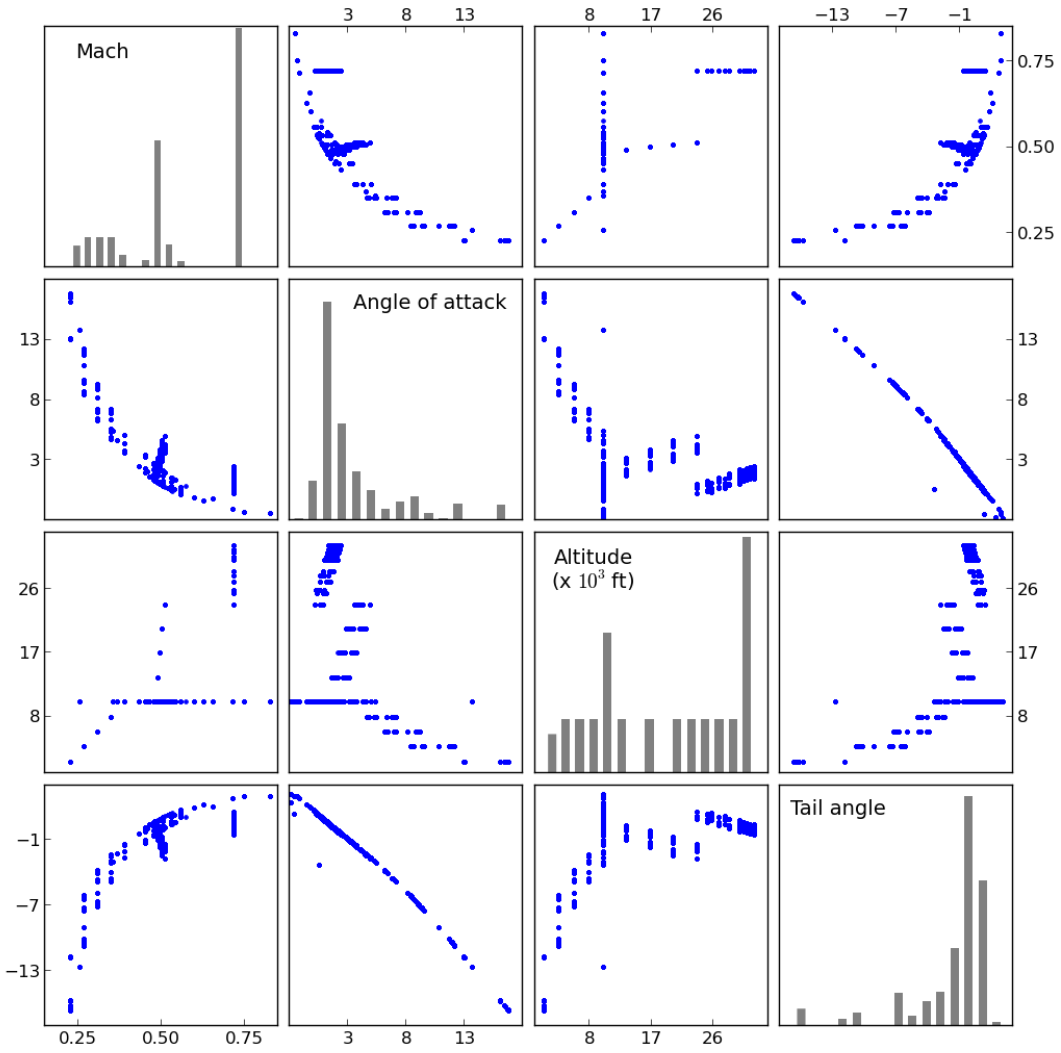


Figure 24: Scatterplot matrix showing the distribution of points in the four-dimensional surrogate model input space that are evaluated during the mission analysis.

[14] Amsallem, D., Cortial, J., and Farhat, C., "Toward Real-Time Computational-Fluid-Dynamics-Based Aeroelastic Computations Using a Database of Reduced-Order Information," *AIAA Journal*, Vol. 48, No. 9, September 2010, pp. 2029–2037.

[15] Koko, F., *Aerostructural and Trajectory Optimization of Morphing Wingtip Devices*, Master's thesis, Faculty of Aerospace Engineering, Delft University of Technology, October 2011.

[16] Liem, R. P., Kenway, G. K. W., and Martins, J. R. R. A., "Multi-point, multi-mission, high-fidelity aerostructural optimization of a long-range aircraft configuration," *14th AIAA/ISSMO Multidisciplinary Analysis and Optimization (MAO) Conference*, Indianapolis, IN, September 2012. doi:10.2514/6.2012-5706, AIAA 2012-5706.

[17] Liem, R. P., Mader, C. A., Lee, E., and Martins, J. R. R. A., "Aerostructural design optimization of a 100-passenger regional jet with surrogate-based mission analysis," *AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, Los Angeles, CA, August 2013. doi:10.2514/6.2013-4372.

[18] Lyu, Z. and Martins, J. R. R. A., "Aerodynamic Design Optimization Studies of a Blended-Wing-Body Aircraft," *Journal of Aircraft*, 2014. doi:10.2514/1.C032491, (In press).

[19] Bishop, C. M., *Pattern Recognition and Machine Learning*, Springer, 2006.

[20] Jin, R., Chen, W., and Simpson, T. W., "Comparative studies of metamodelling techniques under multiple modelling criteria,"

Structural and Multidisciplinary Optimization, Vol. 23, 2001, pp. 1–13. doi:10.1007/S00158-001-0160-4.
[21] Laurenceau, J. and Sagaut, P., "Building Efficient Response Surfaces of Aerodynamic Functions with Kriging and Cokriging," *AIAA Journal*, Vol. 46, No. 2, 2008, pp. 498–507.
[22] Laurenceau, J. and Meaux, M., "Comparison of gradient and response surface based optimization frameworks using adjoint method," *4th AIAA Multidisciplinary Design Optimization Specialist Conference*, Schaumburg, IL, 2008.
[23] Laurenceau, J., Meaux, M., Montagnac, M., and Sagaut, P., "Comparison of gradient-based and gradient-enhanced response-surface-based optimizers," *AIAA Journal*, Vol. 48, No. 5, 2010, pp. 981–994.
[24] Journel, A. G. and Rossi, M. E., "When Do We Need a Trend Model in Kriging?" *Mathematical Geology*, Vol. 21, No. 7, 1989, pp. 715–739. doi:10.1007/BF00893318.
[25] Coffin, J. G., "A Study of Airplane Range and Useful Loads," NACA-TR-69, NACA, 1920.
[26] Breguet, L., "Calcul du Poids de Combustible Consummé par un Avion en Vol Ascendant," *Comptes Rendus Hebdomodaires des Séances de l'Académie des Sciences*, Vol. 177, 1923, pp. 870–872.
[27] McCormick, B. W., *Aerodynamics, Aeronautics, and Flight Mechanics*, John Wiley & Sons, New York, US, 1979.
[28] Lee, H. and Chatterji, G. B., "Closed-Form Takeoff Weight Estimation Model for Air Transportation Simulation," *10th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference*, Fort Worth, TX, Sept 13–15 2010. doi:10.2514/6.2010-9156, AIAA 2010-9156.
[29] Kroo, I. M., *Aircraft Design: Synthesis and Analysis*, Desktop Aeronautics, Palo Alto, CA, 1st ed., Sept 2006.
[30] Henderson, R. P., Martins, J. R. R. A., and Perez, R. E., "Aircraft Conceptual Design for Optimal Environmental Performance," *The Aeronautical Journal*, Vol. 116, 2012, pp. 1—22.
[31] PASS, "Program for Aircraft Synthesis Studies Software Package," Desktop Aeronautics, Inc., Palo Alto, CA, 2005.
[32] Roskam, J. and Lan, C. T. E., *Airplane Aerodynamics and Performance*, DARcorporation, Lawrence, KS, 1997.
[33] Giannakoglou, K. C., Papadimitriou, D. I., and Kampolis, I. C., "Aerodynamic shape design using evolutionary algorithms and new gradient-assisted metamodels," *Computer Methods in Applied Mechanics and Engineering*, Vol. 195, 2006, pp. 6312–6329. doi:10.1016/j.cma.2005.12.008.
[34] Meckesheimer, M., Booker, A. J., Barton, R. R., and Simpson, T. W., "Computationally Inexpensive Metamodel Assessment Strategies," *AIAA Journal*, Vol. 40, No. 10, 2002, pp. 2053–2060. doi:10.2514/2.1538.
[35] Eldred, M. S., Giunta, A. A., Collis, S. S., Alexandrov, N. A., and Lewis, R. M., "Second-Order Corrections for Surrogate-Based Optimization with Model Hierarchies," *11th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Albany, NY, August 30–September 1 2004. doi:10.2514/6.2004-4457, AIAA 2004-4457.
[36] Venter, G., Haftka, R. T., and Starnes, J. H., J., "Construction of Response Surface Approximations for Design Optimization," *AIAA Journal*, Vol. 36, No. 12, December 1998, pp. 2242–2249.
[37] Cressie, N., *Statistics of Spatial Data*, John Wiley and Sons, New York, 1991.
[38] Koehler, J. R. and Owen, A. B., "Computer Experiments," *Handbook of Statistics*, edited by S. Ghosh and C. Rao, Vol. 13, Elsevier Science, New York, 1996.
[39] Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P., "Design and Analysis of Computer Experiments," *Statistical Science*, Vol. 4, 1989, pp. 409–423. doi:10.1214/ss/1177012413.
[40] Forrester, A. I. J., Bresslof, N. W., and Keane, A. J., "Optimization using surrogate models and partially converged computational fluid dynamics simulations," *Proceedings of the Royal Society A*, Vol. 462, March 2000, pp. 2177–2204.
[41] McDonald, D., Grantham, W., Tabor, W., and Murphy, M., "Response surface model development for global/local optimization using radial basis functions," *8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Long Beach, CA, September 2000, AIAA Paper 2000–4776.
[42] Antoulas, A. C., *Approximation of large-scale dynamical systems*, SIAM, Philadelphia, 2005.
[43] Grimme, E., *Krylov Projection Methods for Model Reduction*, Ph.D. thesis, Coordinated-Science Laboratory, University of Illinois at Urbana-Champaign, 1997.
[44] Gugercin, S. and Antoulas, A., "A survey of model reduction by balanced truncation and some new results," *International Journal of Control*, Vol. 77, 2004, pp. 748–766.
[45] Sorensen, D. C. and Antoulas, A. C., "The Sylvester equation and approximate balanced reduction," *Linear Algebra and its Applications*, 2002, pp. 351–352:671–700.
[46] Bui-Thanh, T., Damodaran, M., and Willcox, K. E., "Aerodynamic data reconstruction and inverse design using proper orthogonal decomposition," *Journal of Aircraft*, Vol. 42, No. 8, September–October 2004, pp. 1505–1516.
[47] Holmes, P., Lumley, J. L., and Berkooz, G., *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1996.
[48] Sirovich, L., "Turbulence and the dynamics of coherent structures. Part 1: Coherent structures," *Quarterly of Applied Mathematics*, Vol. 45, No. 3, 1987, pp. 561–571.
[49] Willcox, K. E. and Peraire, J., "Balanced Model Reduction via the proper orthogonal decomposition," *AIAA Journal*, Vol. 40, No. 11, 2002, pp. 2323–2330.
[50] Robinson, T. D., Eldred, M. S., Willcox, K. E., and Haimes, R., "Strategies for Multifidelity Optimization with Variable Dimensional Hierarchical Models," *47th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, Newport, RI, 1–4 May 2006, AIAA Paper 2006-1819.
[51] Robinson, T. D., *Surrogate-Based Optimization using Multifidelity Models with Variable Parameterization*, Ph.D. thesis, Massachusetts Institute of Technology, May 2007.
[52] Briggs, W. L., Henson, V. E., and McCormick, S. F., *A Multigrid Tutorial*, SIAM, Philadelphia, 2nd ed., 2000.
[53] Lewis, R. M. and Nash, S. G., "A Multigrid Approach to the Optimization of Systems Governed by Differential Equations," *8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Optimization*, Long Beach, CA, September 2000, AIAA

Paper 2000-4890.
[54] Alexandrov, N. M., Nielsen, E. J., Lewis, R. M., and Anderson, W. K., "First-Order Model Management with Variable Fidelity Physics Applied to Multi-Element Airfoil Optimization," *8th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, Long Beach, CA, September 2000, AIAA Paper 2000-4886.
[55] Thokala, P. and Martins, J., "Variable Complexity Methods Applied to Airfoil Design," *Engineering Optimization*, Vol. 39, No. 3, April 2006, pp. 271–286.
[56] Ahmed, M. Y. M. and Qin, N., "Surrogate-Based Aerodynamics Design Optimization: Use of Surrogates in Aerodynamics Design Optimization," *13th International Conference on Aerospace Science & Aviation Technology*, Cairo, Egypt, 26–28 May 2009.
[57] Keane, A. J. and Nair, P. B., *Computational Approaches for Aerospace Design, the Pursuit of Excellence*, John Wiley and Sons, Chichester, 2005.
[58] Forrester, A. I. J. and Keane, A. J., "Recent advances in surrogate-based optimization," *Progress in Aerospace Sciences*, Vol. 45, 2009, pp. 50–79.
[59] Lovison, A. and Rigoni, E., "Adaptive sampling with a Lipschitz criterion for accurate metamodeling," *Communications in Applied and Industrial Mathematics*, Vol. 1, No. 2, 2010, pp. 110–126. doi:10.1685/2010CAIM545.
[60] Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F., "Kriging Metamodels for Global Approximation in Simulation-Based Multidisciplinary Design Optimization," *AIAA Journal*, Vol. 39, No. 12, 2001, pp. 2233–2241.
[61] Wang, G. G. and Shan, S., "Review of Metamodeling Techniques in Support of Engineering Design Optimization," *Journal of Mechanical Design*, Vol. 129, No. 4, April 2007, pp. 370–380. doi:10.1115/1.2429697.
[62] McKay, M. D., Conover, W. J., and Beckman, R. J., "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," *Technometrics*, Vol. 21, No. 2, 1979, pp. 239–245.
[63] Sobol', I. M., "On the Systematic Search in a Hypercube," *SIAM Journal of Numerical Analysis*, Vol. 16, No. 5, October 1979, pp. 190–193.
[64] Halton, J. H., "On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals," *Numerische Mathematik*, Vol. 2, 1960, pp. 84–90. doi:10.1007/BF01386213.
[65] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D., "Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments," *Journal of the American Statistical Association*, Vol. 86, No. 416, December 1991, pp. 953–963.
[66] Mitchell, T. J. and Morris, M. D., "Bayesian design and analysis of computer experiments: Two examples," *Statistica Sinica*, Vol. 2, 1992, pp. 359–379.
[67] Krige, D. G., "A statistical approach to some basic mine valuation problems on the Witwatersrand," *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, Vol. 52, 1951, pp. 119–139.
[68] Matheron, G., "Principles of geostatistics," *Economic Geology*, Vol. 58, 1963, pp. 1246–1266.
[69] Olea, R. A., "Sampling Design Optimization for Spatial Functions," *Mathematical Geology*, Vol. 16, No. 4, 1984, pp. 369–392. doi:10.1007/BF01029887.
[70] Viana, F. A. C., Simpson, T. W., Balabanov, V., and Toropov, V., "Metamodeling in Multidisciplinary Design Optimization: How Far Have We Really Come?" *AIAA Journal*, Vol. 52, No. 4, 2014, pp. 670–690. doi:10.2514/1.J052375.
[71] Journel, A. G. and Huijbregts, C. J., *Mining Geostatistics*, Academic Press, London, 1978.
[72] Zimmermann, R., "Asymptotic Behavior of the Likelihood Function of Covariance Matrices of Spatial Gaussian Processes," *Journal of Applied Mathematics*, 2010. doi:10.115/2010/494070.
[73] Lophaven, S. N., Nielsen, H. B., and Søndergaard, J., "DACE – A Matlab Kriging Toolbox, Version 2.0." Tech. Rep. IMM-REP-2002-12, Informatics and Mathematical Modeling, Technical University of Denmark, 2002.
[74] Kitanidis, P. K., *Introduction to geostatistics: applications in hydrogeology*, Cambridge University Press, 1997.
[75] Cressie, N., "The Origins of Kriging," *Mathematical Geology*, Vol. 22, No. 3, 1990, pp. 239–253.
[76] Costa, J.-P., Rostaing, P., and Pitarque, T., "A comparison between kriging and radial basis function networks for nonlinear prediction." *International Workshop on Nonlinear Signal and Image Processing, NSIP'99.*, Antalya, Turkey., 1999.
[77] O'Hagan, A. and Kingman, J. F. C., "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 40, No. 1, 1978, pp. 1–42.
[78] Omre, H., "Bayesian kriging—Merging observations and qualified guesses in kriging," *Mathematical Geology*, Vol. 19, No. 1, 1987, pp. 25–39. doi:10.1007/BF01275432.
[79] Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D., "A Bayesian Approach to the Design and Analysis of Computer Experiments," Tech. Rep. ORNL-6498, National Technical Information Service, 5285 Port Royal Road, Springfield, VA 22161, 1988.
[80] Handcock, M. S. and Stein, M. L., "A Bayesian Analysis of Kriging," *Technometrics*, Vol. 35, No. 4, November 1993, pp. 403–410.
[81] Hooke, R. and Jeeves, T. A., ""Direct Search" Solution of Numerical and Statistical Problems," *Journal of the Association for Computing Machinery*, Vol. 8, 1961, pp. 212–229. doi:10.1145/321062.321069.
[82] Han, Z.-H., Görtz, S., and Zimmermann, R., "On Improving Efficiency and Accuracy of Variable-Fidelity Surrogate Modeling in Aero-data for Loads Context," *Proceedings of CEAS 2009 European Air and Space Conference*, Manchester, UK, October 26–29 2009.
[83] Rijpkema, J. J. M., Etman, L. F. P., and Schoofs, A. J. G., "Use of Design Sensitivity Information in Response Surface and Kriging Metamodels," *Optimization and Engineering*, Vol. 2, 2001, pp. 469–484.
[84] Shi, J. Q., Murray-Smith, R., and Titterington, D. M., "Hierarchical Gaussian process mixtures for regression," *Statistics and Computing*, Vol. 15, 2005, pp. 31–41.

[85] Masoudnia, S. and Ebrahimpour, R., "Mixture of experts: a literature survey," *Artificial Intelligence Review*, 2012, pp. 1–19. doi:10.1007/s10462-012-9338-y.
[86] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E., "Adaptive Mixtures of Local Experts," *Neural Computation*, Vol. 3, 1991, pp. 79–87.
[87] Tresp, V., "Mixtures of Gaussian Processes," *Advances in Neural Information Processing Systems*, Vol. 13, 2000, pp. 654–660.
[88] Rasmussen, C. E. and Ghahramani, Z., "Infinite Mixtures of Gaussian Process Experts," *Advances in Neural Information Processing Systems 14*, edited by T. Diettrich, S. Becker, and Z. Ghahramani, MIT Press, 2002.
[89] Quandt, R. E. and Ramsey, J. B., "Estimating Mixtures of Normal Distributions and Switching Regressions," *Journal of the American Statistical Association*, Vol. 73, No. 364, 1978, pp. 730–738. doi:10.1080/01621459.1978.10480085.
[90] Veaux, R. D. D., "Mixtures of linear regressions," *Computational Statistics and Data Analysis*, Vol. 8, 1989, pp. 227–245. doi:10.1016/0167-9473(89)90043-1.
[91] Faria, S. and Soromenho, G., "Fitting mixtures of linear regressions," *Journal of Statistical Computation and Simulation*, Vol. 80, No. 2, 2010, pp. 201–225. doi:10.1080/00949650802590261.
[92] Nguyen-Tuong, D., Peters, J., and Seeger, M., "Local Gaussian process regression for real time online model learning and control," *In Advances in Neural Information Processing Systems 22 (NIPS)*, 2008.
[93] Steinhaus, H., "Sur la division des corps materiels en parties," *Bull. Acad. Polon. Sci.*, Vol. 1, 1956, pp. 801–804.
[94] Lloyd, S. P., "Least Squares Quantization in PCM," *IEEE Transactions on Information Theory*, Vol. 28, No. 2, March 1982, pp. 129–137. doi:10.1109/TIT.1982.1056489.
[95] MacQueen, J., "Some methods for classification and analysis of multivariate observations," *Fifth Berkeley Symposium on Mathematics. Statistics and Probability.*, University of California Press., 1967, pp. 281–297.
[96] Reynolds, D., "Gaussian Mixture Models," *Encyclopedia of Biometric Recognition*, Springer, February 2008.
[97] McLachlan, G. J. and Basford, K. E., *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker Inc., 1988.
[98] McLachlan, G. J. and Peel, D., *Finite Mixture Models*, Wiley, 2000.
[99] Dempster, A. P., Laird, N. M., and Rubin, D. B., "Maximum Likelihood from Incomplete Data via the *EM* Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, 1977, pp. 1–38.
[100] McLachlan, G. J. and Krishnan, T., *The EM Algorithm and its Extensions*, Wiley, 1997.
[101] Vassberg, J. C., DeHaan, M. A., Rivers, S. M., and Wahls, R. A., "Development of a *Common Research Model* for Applied CFD Validation Studies," *26th AIAA Applied Aerodynamics Conference*, AIAA, Honolulu, HI, August 2008. doi:10.2514/6.2008-6919, AIAA 2008-6919.
[102] Kennedy, G. J. and Martins, J. R. R. A., "Parallel Solution Methods for Aerostructural Analysis and Design Optimization," *13th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*, Fort Worth, TX, September 2010. doi:10.2514/6.2010-9308, AIAA 2010-9308.
[103] Kennedy, G. J. and Martins, J. R. R. A., "A parallel aerostructural optimization framework for aircraft design studies," *Structural and Multidisciplinary Optimization*, 2014, (In press).
[104] Raymer, D. P., *Aircraft Design: A Conceptual Approach*, Education Series, AIAA, Washington, DC, 1992.
[105] Sadraey, M. H., *Aircraft Design: A Systems Engineering Approach*, John Wiley & Sons, Chichester, West Sussex, 2012.
[106] Kenway, G. K. W., Kennedy, G. J., and Martins, J. R. R. A., "A Scalable Parallel Approach for High-Fidelity Steady-State Aeroelastic Analysis and Derivative Computations," *AIAA Journal*, 2013, (In press).
[107] Martins, J. R. R. A. and Hwang, J. T., "Review and Unification of Methods for Computing Derivatives of Multidisciplinary Computational Models," *AIAA Journal*, Vol. 51, No. 11, 2013, pp. 2582–2599. doi:10.2514/1.J052184.