Projection-based multifidelity linear regression for data-poor applications

Vignesh Sella*, Julie Pham[†], Anirban Chaudhuri[‡], Karen Willcox [§]
University of Texas at Austin, TX 78712, USA

Surrogate modeling for systems with high-dimensional quantities of interest is important for many applications in science and engineering, but remains a challenge in situations where training data are expensive to acquire. This work develops multifidelity (MF) approaches for multivariate multi-output linear regression for data-poor applications with high-dimensional outputs. The MF approach combines information from many low-cost, low-fidelity (LF) model evaluations with limited expensive, high-fidelity (HF) model evaluations. We implement and contrast three MF linear regression methods with projections to a lower-dimensional space through proper orthogonal decomposition basis vectors. The three MF linear regression approaches developed in this work are: (i) an additive method based on the Kennedy-O'Hagan framework, (ii) a direct data augmentation using LF data, and (iii) a data augmentation method using explicit linear regression mapping between LF and HF data. The data augmentation technique combines the HF and the LF data into one training data set and the linear regression model is trained using weighted least squares with different weights for the different fidelity models. We apply the projection-enabled MF linear regression methods to approximate the surface pressure field on a hypersonic vehicle in flight. The MF linear regression outperforms the single-fidelity linear regression in the low data regime of $3-10$ HF samples with an improvement in the range of approximately $3-12\%$ in median accuracy for similar computational cost.

I. Introduction

An important challenge in scientific machine learning research is to develop methods that can exploit and maximize the amount of learning possible from scarce data [1]. The need for such research arises often in science and engineering, and especially in the case of computational fluid dynamics (CFD), since expensive-to-evaluate high-fidelity (HF) models make many-query problems such as uncertainty quantification, risk analysis, optimization, and optimization under uncertainty computationally prohibitive [2]. Thus, there is a need for surrogate models that can approximate the solutions to HF models to accelerate the design and analysis process. However, lack of sufficient HF data adversely affects the accuracy of surrogate models. We propose multifidelity linear regression methods that can use multiple cheaper lower-fidelity (LF) information sources along with the limited HF data for training linear regression models.

Linear regression [3–6], including polynomial regression or the response surface methodology (RSM) [7], is a regression analysis technique that has been extensively used for surrogate modelling and prediction in aerospace applications. Madsen et al. [8] utilized RSM using polynomials for diffuser shape optimization with up to five input design variables. Nakamura et al. [9] applied linear regression methods to predict fluid-flow over two-dimensional cylinders and the state-estimation from wall characteristics in turbulent flow using training samples on the order of $O(10^3)$. Another line of work seeks to use low-order polynomial RSM approximations as a tool within multidisciplinary design optimization frameworks in order to deal with data plagued by high computational cost and noise in aerospace design applications [10, 11]. Recently, there has also been a growing wealth of literature on using even more data to train and create surrogate models. Ladický et al. [12] investigated the feasibility of random forests to approximate fluid particle behaviour in time and compared it to the state-of-the-art position based fluid approach using $O(10^9)$ training samples. In many of these studies, the amount of training data procured to train the respective surrogate models was a computationally intensive effort. While these levels of data procuration are possible, they are not always plausible without access to significant computing resources. In this work, we propose multifidelity versions of linear regression that can alleviate the issue of a lack of data.

*Graduate Research Assistant, Oden Institute of Computational Engineering and Sciences, AIAA Student Member

[†]Graduate Research Assistant, Deptartment of Aerospace Engineering and Engineering Mechanics, AIAA Student Member

[‡]Research Associate, Oden Institute of Computational Engineering and Sciences, AIAA Member

[§]Director, Oden Institute of Computational Engineering and Sciences, AIAA Fellow

Multifidelity (MF) linear regression methods that can fuse information from cheaper LF data with the sparse HF data have been explored to address the prohibitive data requirement for regression in many applications. A MF linear regression technique was proposed by Balabanov et al. [13] that used a quadratic RSM created using many coarse finite element structural model simulations and a correction RSM using a small number of finer mesh finite element model simulations for civilian transport wing design. Along a similar vein, the work in [14–17] on surrogate modelling in computational fluid dynamics and structural mechanics for low-dimensional problems found MF linear regression to to be more efficient and have a higher accuracy than a single-fidelity approach given the same computational cost. There has also been development on MF surrogate modelling using artificial neural networks [18, 19] and using Gaussian process regression [20, 21]. The focus of this paper is on linear regression that can tackle problems in the ultra low-data regime and with high-dimensional outputs.

In this work, we develop and compare three MF linear regression methods: (i) additive structure using the Kennedy-O'Hagan [22] approach, (ii) data augmentation of LF data directly, and (iii) data augmentation via an explicit mapping between LF outputs and HF outputs. The MF regression algorithms can be generalized to any other choice of underlying regression technique. However, we implement the MF regression algorithms using multi-output linear (or polynomial) regression since we are dealing with a limited number of HF samples. For problems involving high-dimensional quantities of interest, we show how the MF linear regression methods combine with data-driven projection techniques to improve the performance and accuracy. We apply the projection-enabled MF linear regression techniques to predict the pressure loads over the surface of a hypersonic vehicle.

The remainder of this paper is organized as follows. Section II describes the MF regression problem setup. Section III describes the different MF linear regression methods developed in this work. Section IV describes the hypersonic vehicle application and provides a detailed analysis of the performance of the algorithms. Finally, Section V concludes the paper with remarks on the efficacy and performance of the proposed methods.

II. Problem Setup

This paper considers linear regression problems wherein the d inputs to a system are $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ defined on the input space $\mathcal{X}$, and the output quantity of interest is m-dimensional $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^m$ defined on the output space $\mathcal{Y}$. In our target applications, $\boldsymbol{y}$ is a high-dimensional quantity with m in thousands. For a single-fidelity case, a set of N input-output samples $(\boldsymbol{X}, \boldsymbol{Y})$, where $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N] \in \mathbb{R}^{d \times N}$ and $\boldsymbol{Y} = [\boldsymbol{y}_1, ..., \boldsymbol{y}_N] \in \mathbb{R}^{m \times N}$, are available for training the surrogate model through linear regression using ordinary least squares (OLS). In many aerospace applications, only sparse data is available for training these surrogates due to the high computational cost associated with the HF model evaluations. The goal of this work is to develop a multifidelity linear regression method that has the potential to build accurate surrogates in a data-poor regime using linear regression as the underlying regression method. Given input-output data of varying fidelity levels, multifidelity linear regression can efficiently approximate the HF quantity of interest by incorporating information from the cheaper LF models. The training data consists of N_{HF} samples from the HF model $(\boldsymbol{X}_{\mathrm{HF}}, \boldsymbol{Y}_{\mathrm{HF}})$ and N_{LF} samples from the LF model $(\boldsymbol{X}_{\mathrm{LF}}, \boldsymbol{Y}_{\mathrm{LF}})$. In this work, we will be considering bifidelity problems, wherein there are only two models with different levels of fidelity, but the general idea can be extended to more than two fidelity levels.

III. Multifidelity Linear Regression

In this section, we first discuss how we implement dimensionality reduction to efficiently approximate output quantities III.A. We then describe two overarching multifidelity linear regression methods, both of which incorporate dimensionality reduction for high-dimensional outputs: an additive method in Section III.B and data augmentation via synthetic data generation in Section III.C. Furthermore, all the methods presented here work for the general case of *non-collocated* samples, i.e. the $\boldsymbol{x}_{\mathrm{LF}}$ and $\boldsymbol{x}_{\mathrm{HF}}$ in each input training set do not need to be at the same locations.

A. Output Dimensionality Reduction Using Proper Orthogonal Decomposition

This work targets applications with high-dimensional output quantity of interest $\boldsymbol{y}$ and with limited training data for the surrogate. We use dimensionality reduction techniques such as proper orthogonal decomposition (POD) to reduce output dimensions and train a compact surrogate model with limited data. Similar projection-enabled techniques have been used for parametric reduced order models [23–25] and for neural networks [26]. In this work, the POD basis vectors are obtained by taking the singular value decomposition (SVD) of the training data matrix consisting of N samples $\boldsymbol{Y} \in \mathbb{R}^{m \times N}$ centered by the sample average mean of the training data $\overline{\boldsymbol{Y}}$. In our target applications, m is in thousands and the number of samples $N \ll m$. The thin SVD of the centered training data matrix is

$$\boldsymbol{Y} - \overline{\boldsymbol{Y}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top, \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{m\times N}$ and $\boldsymbol{V} \in \mathbb{R}^{N\times N}$ are orthogonal matrices, and $\boldsymbol{\Sigma} \in \mathbb{R}^{N\times N}$ is a square diagonal matrix consisting of the singular values, σ, of the centered training data matrix. The reduced basis for projection to a lower-dimensional subspace of size $k \ll m$ (and $k \leq N$) $\boldsymbol{U}_k \in \mathbb{R}^{m\times k}$ consists of the first k columns of the left singular vectors $\boldsymbol{U}$. The projection of an output sample $\boldsymbol{y}$ on the low-dimensional subspace is given by the reduced state $\boldsymbol{c} = \boldsymbol{U}_k^\top \boldsymbol{y}$. The dimension k is chosen such that the cumulative energy captured by the first k POD modes is larger than a specified tolerance of ϵ as given by

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{N} \sigma_i^2} > \epsilon, \tag{2}$$

where σ_i is the i-th singular value.

B. Multifidelity Linear Regression Using an Additive Structure

In this method, we use the Kennedy-O'Hagan approach [22] to build the additive multifidelity linear regression. This method assumes that the relationship between the LF and the HF data can be well modeled linearly. We begin by training the LF surrogate model, $\boldsymbol{f}_{\text{LF}}$, that fits the reduced state of the LF outputs $(\boldsymbol{U}_k^{\text{LF}})^\top(\boldsymbol{Y}_{\text{LF}} - \overline{\boldsymbol{Y}}_{\text{LF}})$ using OLS, where $\boldsymbol{U}_k^{\text{LF}}$ is the LF reduced basis and $\overline{\boldsymbol{Y}}_{\text{LF}}$ is the mean of the LF training data (see Eqn. (1)). Then, we use the LF surrogate to predict data at the same input locations as the HF data. We reconstruct the LF surrogate predictions in the full-dimensional space as $(\boldsymbol{U}_k^{\text{LF}}\boldsymbol{f}_{\text{LF}}(\boldsymbol{X}_{\text{HF}}) + \overline{\boldsymbol{Y}}_{\text{LF}})$. We use the predicted data at the same locations to estimate the discrepancy as

$$\boldsymbol{\delta}(\boldsymbol{X}_{\text{HF}}) = \boldsymbol{Y}_{\text{HF}} - (\boldsymbol{U}_k^{\text{LF}}\boldsymbol{f}_{\text{LF}}(\boldsymbol{X}_{\text{HF}}) + \overline{\boldsymbol{Y}}_{\text{LF}}). \tag{3}$$

We fit a linear regression for discrepancy, $\boldsymbol{f}_\delta$, to the projected states of the discrepancy data via similar steps as for $\boldsymbol{f}_{\text{LF}}$. We can then use the discrepancy surrogate along with the LF surrogate to construct our MF linear regression surrogate as

$$\boldsymbol{f}_{\text{MF}}(\boldsymbol{x}) = \boldsymbol{U}_k^{\delta}\boldsymbol{f}_\delta(\boldsymbol{x}) + \overline{\boldsymbol{\delta}} + \boldsymbol{U}_k^{\text{LF}}\boldsymbol{f}_{\text{LF}}(\boldsymbol{x}) + \overline{\boldsymbol{Y}}_{\text{LF}} \tag{4}$$

where $\boldsymbol{U}_k^{\delta}$ is reduced basis obtained from the SVD on the discrepancy data and $\overline{\boldsymbol{\delta}}$ is the mean of the discrepancy data. We summarize the additive multifidelity linear regression approach in Alg. 1.

Algorithm 1 Multifidelity linear regression via additive method (with output dimensionality reduction)

```
Input: HF and LF training data ($\boldsymbol{X}_{\text{LF}}$, $\boldsymbol{Y}_{\text{LF}}$) and ($\boldsymbol{X}_{\text{HF}}$, $\boldsymbol{Y}_{\text{HF}}$), new input locations for prediction $\boldsymbol{X}^*$
Output: Output predictions $\widehat{\boldsymbol{Y}}_{\text{MF}}$ at inputs $\boldsymbol{X}^*$ from MF surrogate
Use $\boldsymbol{U}_k^{\text{LF}}$ from the SVD of $\boldsymbol{Y}_{\text{LF}}$ (Eqn. 1) to obtain the reduced states $\boldsymbol{C}_{\text{LF}} \in \mathbb{R}^{k\times N_{\text{LF}}}$:
    $\boldsymbol{C}_{\text{LF}} = \left(\boldsymbol{U}_k^{\text{LF}}\right)^\top \left(\boldsymbol{Y}_{\text{LF}} - \overline{\boldsymbol{Y}}_{\text{LF}}\right)$
Train LF linear regression surrogate model $\boldsymbol{f}_{\text{LF}}$ on $(\boldsymbol{X}_{\text{LF}}, \boldsymbol{C}_{\text{LF}})$ using OLS linear regression
Predict and reconstruct LF outputs at the HF input locations $(\boldsymbol{U}_k^{\text{LF}}\boldsymbol{f}_{\text{LF}}(\boldsymbol{X}_{\text{HF}}) + \overline{\boldsymbol{Y}}_{\text{LF}})$
Estimate discrepancy data $\boldsymbol{\delta}$ using Eqn. (3)
Use $\boldsymbol{U}_k^{\delta}$ from the SVD of $\boldsymbol{\delta}$ to project the discrepancy to the reduced state: $\boldsymbol{C}_\delta = (\boldsymbol{U}_k^{\delta})^\top(\boldsymbol{\delta} - \overline{\boldsymbol{\delta}})$
Train discrepancy linear regression $\boldsymbol{f}_\delta$ on ($\boldsymbol{X}_{\text{HF}}$, $\boldsymbol{C}_\delta$) using OLS linear regression
Set $\boldsymbol{f}_{\text{MF}}(\boldsymbol{X}_{\text{HF}})$ as the linear combination of $\boldsymbol{f}_\delta(\boldsymbol{X}_{\text{HF}})$, $\boldsymbol{f}_{\text{LF}}(\boldsymbol{X}_{\text{HF}})$, and the known bias terms using Eqn. (4)
Predict outputs $\widehat{\boldsymbol{Y}}_{\text{MF}}$ at new input locations $\boldsymbol{X}^*$:
    $\widehat{\boldsymbol{Y}}_{\text{MF}} = \boldsymbol{f}_{\text{MF}}(\boldsymbol{X}^*)$
```

C. Multifidelity Linear Regression Through Data Augmentation

In the data augmentation method, we generate synthetic data from LF model evaluations and augment the HF training set with the synthetic data. Synthetic data in this context refers to data which comes from a different source or fidelity level than the HF model. We describe two approaches for obtaining synthetic data in Section III.C.1 and the general data augmentation method for multifidelity linear regression using the generated synthetic data in Section III.C.2.

1. Synthetic data for data augmentation

We explore two ways of using the LF training data to create the synthetic data: (i) augmenting the LF training data directly, and (ii) using an explicit linear regression mapping of LF data to HF data. In the first approach, the LF training data is directly fed in as synthetic data to be augmented to the HF training data as $(\boldsymbol{X}_{\mathrm{LF}}^{\mathrm{syn}}, \boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}}) = (\boldsymbol{X}_{\mathrm{LF}}, \boldsymbol{Y}_{\mathrm{LF}})$. There is no need to build a LF linear regression to implement the direct data augmentation. In the second approach, an explicit map between the LF and HF data is created. To create the explicit map, we need HF and LF data at the same input locations. However, in the general case, we do not assume that LF and HF samples are available at the same locations. Thus, we construct the LF surrogate model $\boldsymbol{f}_{\mathrm{LF}}$ and use this to predict LF surrogate outputs at the HF input locations similar to the process shown for the additive approach in Steps 1-3 in Alg. 1. Note that if LF and HF data were available at the same locations one can skip the initial step to create a LF surrogate model and directly train the explicit map on the collocated data. Then, we train the explicit linear regression map $\boldsymbol{f}_{\mathrm{LF}\mapsto\mathrm{HF}}$ from reduced LF outputs to reduced HF outputs using data at the same location $(\widehat{\boldsymbol{C}}_{\mathrm{LF}}, \boldsymbol{C}_{\mathrm{HF}})$ as detailed in Alg. 2. After constructing the explicit map, we can generate the synthetic data at all the LF input locations as $\boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}} = \boldsymbol{U}_k^{\mathrm{HF}} \boldsymbol{f}_{\mathrm{LF}\mapsto\mathrm{HF}}(\boldsymbol{C}_{\mathrm{LF}}) + \overline{\boldsymbol{Y}}_{\mathrm{HF}}$ to obtain the training set $(\boldsymbol{X}_{\mathrm{LF}}^{\mathrm{syn}} = \boldsymbol{X}_{\mathrm{LF}}, \boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}})$. We summarize this process in Alg. 2. Once the HF training data is augmented with the synthetic data, the multifidelity linear regression is trained through the process described in the next section.

Algorithm 2 Synthetic data generation via an explicit mapping using linear regression

```
Input: HF and LF training data $(\boldsymbol{X}_{\mathrm{LF}}, \boldsymbol{Y}_{\mathrm{LF}})$ and $(\boldsymbol{X}_{\mathrm{HF}}, \boldsymbol{Y}_{\mathrm{HF}})$
Output: Synthetic data $\boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}}$ at inputs $\boldsymbol{X}_{\mathrm{LF}}$ from the LF to HF surrogate map
Reduce $\boldsymbol{Y}_{\mathrm{HF}}$ using $\boldsymbol{U}_k^{\mathrm{HF}}$ from the SVD of $\boldsymbol{Y}_{\mathrm{HF}}$ to obtain the reduced states $\boldsymbol{C}_{\mathrm{HF}} \in \mathbb{R}^{k \times N_{\mathrm{HF}}}$:
    $\boldsymbol{C}_{\mathrm{HF}} = \left(\boldsymbol{U}_k^{\mathrm{HF}}\right)^{\top} \left(\boldsymbol{Y}_{\mathrm{HF}} - \overline{\boldsymbol{Y}}_{\mathrm{HF}}\right)$
Obtain $(\boldsymbol{U}_k^{\mathrm{LF}} \boldsymbol{f}_{\mathrm{LF}}(\boldsymbol{X}_{\mathrm{HF}}) + \overline{\boldsymbol{Y}}_{\mathrm{LF}})$ from Alg. 1 (steps 1-3) and project it to obtain the reduced states for explicit map training:
    $\widehat{\boldsymbol{C}}_{\mathrm{LF}} = \left(\boldsymbol{U}_k^{\mathrm{HF}}\right)^{\top} \left(\left(\boldsymbol{U}_k^{\mathrm{LF}} \boldsymbol{f}_{\mathrm{LF}}(\boldsymbol{X}_{\mathrm{HF}}) + \overline{\boldsymbol{Y}}_{\mathrm{LF}}\right) - \overline{\boldsymbol{Y}}_{\mathrm{HF}}\right)$
Train LF $\mapsto$ HF linear regression model $\boldsymbol{f}_{\mathrm{LF}\mapsto\mathrm{HF}}$ on $(\widehat{\boldsymbol{C}}_{\mathrm{LF}}, \boldsymbol{C}_{\mathrm{HF}})$ using OLS linear regression
Use $\boldsymbol{U}_k^{\mathrm{LF}}$ from the SVD of $\boldsymbol{Y}_{\mathrm{LF}}$ to project all the LF data to the reduced state $\boldsymbol{C}_{\mathrm{LF}}$
Generate synthetic data $\boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}}$ at $\boldsymbol{X}_{\mathrm{LF}}$ locations and reconstruct the output in the full-dimensional space:
    $\boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}} = \boldsymbol{U}_k^{\mathrm{HF}} \boldsymbol{f}_{\mathrm{LF}\mapsto\mathrm{HF}}\left(\boldsymbol{C}_{\mathrm{LF}}\right) + \overline{\boldsymbol{Y}}_{\mathrm{HF}}$
```

2. Data augmentation with weighted least squares

The data augmentation method augments the HF training dataset with the synthetic data generated from the LF training data and performs weighted least squares (WLS) linear regression to train the MF surrogate model. In this work, we use the two approaches described in Section III.C.1 for generating the synthetic data, $(\boldsymbol{X}_{\mathrm{LF}}^{\mathrm{syn}}, \boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}})$. We then augment our HF training dataset with the transformed LF synthetic data to get the training data of size $N_{\mathrm{HF}} + N_{\mathrm{LF}}$ for the MF linear regression as $([\boldsymbol{X}_{\mathrm{HF}}, \boldsymbol{X}_{\mathrm{LF}}^{\mathrm{syn}}], [\boldsymbol{Y}_{\mathrm{HF}}, \boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}}])$ where typically, $N_{\mathrm{LF}} > N_{\mathrm{HF}}$. While we can train our MF surrogate model directly with this augmented dataset, we know that the synthetic data is lower-fidelity. OLS has an underlying assumption of homoscedasticity, or constant variance in the errors. Since we know that $\boldsymbol{Y}_{\mathrm{LF}}^{\mathrm{syn}}$ is inherently a LF approximation to the true HF quantity, we use different sample weights to account for the expected heteroscedasticity when training on data from different sources. The sample weight matrix $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_{N_{\mathrm{HF}}+N_{\mathrm{LF}}})$ is used in the MF linear regression training through WLS [5]. For all HF samples we use the sample weight $w_i = 1, i = 1, \ldots, N_{\mathrm{HF}}$ and for all synthetic data samples generated from LF data we use equal sample weights of $w_i = w_{\text{syn}} < 1, i = N_{\text{HF}} + 1, \ldots, N_{\text{HF}} + N_{\text{LF}}$. Furthermore, in the results section we note the impact of various sample weighting schemes on the numerical examples we have applied MF regression to. We summarize the data augmentation method for MF linear regression in Alg. 3. As we show in the results, using the direct LF data approach is more sensitive to the choice of weight parameter w_{syn} in WLS as compared to the explicit linear regression map approach. One way to further improve the performance for both the approaches would be select an optimal hyperparameter w_{syn} by minimizing the leave-one-out cross-validation error.

Algorithm 3 Multifidelity linear regression via data augmentation

```
Input: HF and LF training data $(\boldsymbol{X}_{\text{LF}}, \boldsymbol{Y}_{\text{LF}})$ and $(\boldsymbol{X}_{\text{HF}}, \boldsymbol{Y}_{\text{HF}})$, synthetic sample weight $w_{\text{syn}}$, new input locations
    for prediction $\boldsymbol{X}^*$
Output: Output predictions $\widehat{\boldsymbol{Y}}_{\text{MF}}$ at inputs $\boldsymbol{X}^*$ from MF surrogate
Generate synthetic data by transforming the LF data: $(\boldsymbol{X}_{\text{LF}}, \boldsymbol{Y}_{\text{LF}}) \mapsto (\boldsymbol{X}_{\text{LF}}^{\text{syn}}, \boldsymbol{Y}_{\text{LF}}^{\text{syn}})$ using Alg. 2
Augment the training dataset to contain $N_{\text{HF}} + N_{\text{LF}}$ samples: $([\boldsymbol{X}_{\text{HF}}, \boldsymbol{X}_{\text{LF}}^{\text{syn}}], [\boldsymbol{Y}_{\text{HF}}, \boldsymbol{Y}_{\text{LF}}^{\text{syn}}])$
Use $\boldsymbol{U}_k^{\text{HF}}$ from the SVD of $\boldsymbol{Y}_{\text{HF}}$ (Eqn. 1) to obtain the reduced states of MF training data outputs $\boldsymbol{C}_{\text{MF}} \in \mathbb{R}^{k\times(N_{\text{HF}}+N_{\text{LF}})}$:
    $$\boldsymbol{C}_{\text{MF}} = \left(\boldsymbol{U}_k^{\text{HF}}\right)^\top \left([\boldsymbol{Y}_{\text{HF}}, \boldsymbol{Y}_{\text{LF}}^{\text{syn}}] - \overline{\boldsymbol{Y}}_{\text{HF}}\right)$$
Set up sample weight matrix $\boldsymbol{W}$ using $w_i = 1, i = 1, \ldots, N_{\text{HF}}$ and $w_j = w_{\text{syn}}, j = N_{\text{HF}} + 1, \ldots, N_{\text{HF}} + N_{\text{LF}}$
Train MF linear regression surrogate model $\boldsymbol{f}_{\text{MF}}$ on $([\boldsymbol{X}_{\text{HF}}, \boldsymbol{X}_{\text{LF}}^{\text{syn}}], \boldsymbol{C}_{\text{MF}})$ with sample weights $\boldsymbol{W}$ using WLS linear
    regression
Predict $\widehat{\boldsymbol{Y}}_{\text{MF}}$ at new input locations $\boldsymbol{X}^*$ by reconstructing the output of $\boldsymbol{f}_{\text{MF}}(\boldsymbol{X}^*)$ in the full-dimensional space:
    $$\widehat{\boldsymbol{Y}}_{\text{MF}} = \boldsymbol{U}_k^{\text{HF}} \boldsymbol{f}_{\text{MF}}(\boldsymbol{X}^*) + \overline{\boldsymbol{Y}}_{\text{HF}}$$
```

IV. Multifidelity Linear Regression for Adapted IC3X Testbed Problem

In the following sections, we present the results for a testbed problem in the computational fluid dynamics domain described in Section IV.A. The HF and the LF model used for the MF linear regression are described in Section IV.B. Then, we present results for the projection-enabled MF linear regression techniques proposed in this work in Section IV.C.

A. IC3X Problem Description

The hypersonic vehicle considered in this work is the Initial Concept 3.X (IC3X). The IC3X was initially proposed by Pasiliao et al. [27], and a finite element model for the vehicle was developed by Witeof et al. [28]. The distributed pressure load over the surface of a hypersonic vehicle, which is a primary quantity of interest, varies as a function of Mach number, angle of attack, and sideslip angle. Based on a nominal mission trajectory for this geometry, we consider the interval of Mach numbers $M \in [5, 7]$, angles of attack $\alpha \in [0, 8]$, and sideslip angles $\beta \in [0, 8]$. The pressure field is simulated by solving the inviscid Euler equations via a Cartesian volume mesh using the flow solver package Cart3D [29–31]. The pressure field solution computed by Cart3D is a vector of dimension $m = 55966$. A pressure field solution at a given operating condition of $M = 7$, $\alpha = 4$, and $\beta = 0$ is shown in Figure 1.

In order to gain design insights for performance, stability, and reliability of a hypersonic vehicle, CFD simulations are required over a range of flight conditions. For example, stability analyses for a hypersonic vehicle requires an understanding of the pressure field surrounding the vehicle over the operating range of Mach number, angle of attack, and sideslip angle of the vehicle. However, high-fidelity CFD solutions are computationally intensive due to the fine mesh size required to adequately capture the physics of hypersonic flight. In this work, we address the prohibitive computational cost through constructing cheaper approximations using MF linear regression techniques that reduce the number of HF model evaluations required to make accurate predictions of the pressure fields by combining with data from cheaper LF models.

B. Model Specifications and Data Generation

We can construct different levels of fidelity for the pressure field solution by leveraging Cart3D's inherent mesh adaptation over multiple adaptation cycles. We control the number of adaptations and the error tolerance. In this work,

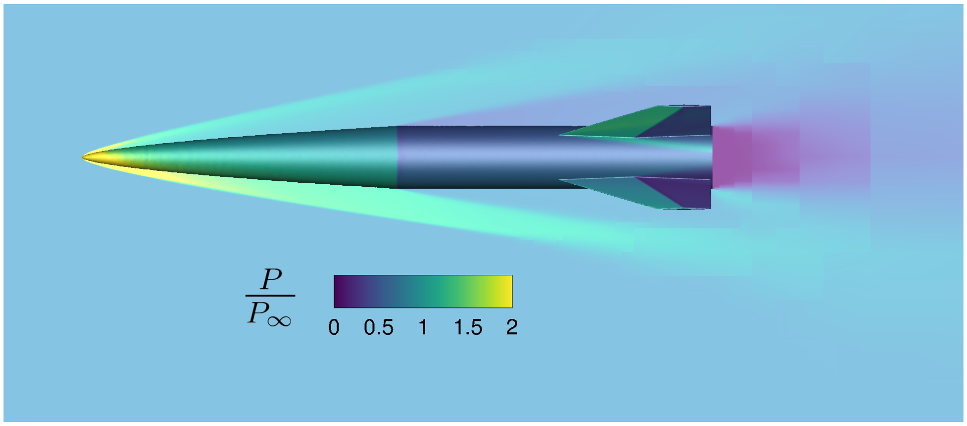


Fig. 1 Cart3D pressure solution at $M = 7$, $\alpha = 4$, and $\beta = 0$

we define two levels of fidelity for simulating the pressure field: (i) the HF model with a finer volume mesh after more mesh adaptations and (ii) the LF model with a coarser volume mesh after fewer mesh adaptations and with a lower error tolerance. We control the maximum number of mesh refinements ("Max Refinement"), the maximum number of adaptation processes ("Max Adaptations"), error tolerance, and the number of iterations per adaptation process ("Cycles/Adaptation") to generate the different fidelity levels. The specifications for the HF model and the LF model used in this work are described in Table 1. We also provide the relative computational cost in terms of one HF model evaluation.

While the choice of HF and LF sample sizes is problem- and resource-dependent, in this case we use a very limited number of HF samples, $N_{\mathrm{HF}} \in \{3, 5, 10\}$, a LF training sample size of $N_{\mathrm{LF}} = 80$, and a HF testing sample size of $N_{\mathrm{HF}}^{\mathrm{test}} = 50$ to analyze the algorithms' effectiveness in the ultra low-data regime. We present the results while accounting for the computational cost of using the additional 80 LF samples given by $80/127 = 0.63$ equivalent HF samples. In all of our results, we bootstrap the available dataset to provide a measure of robustness of each method over 50 repetitions of the training and the testing samples.

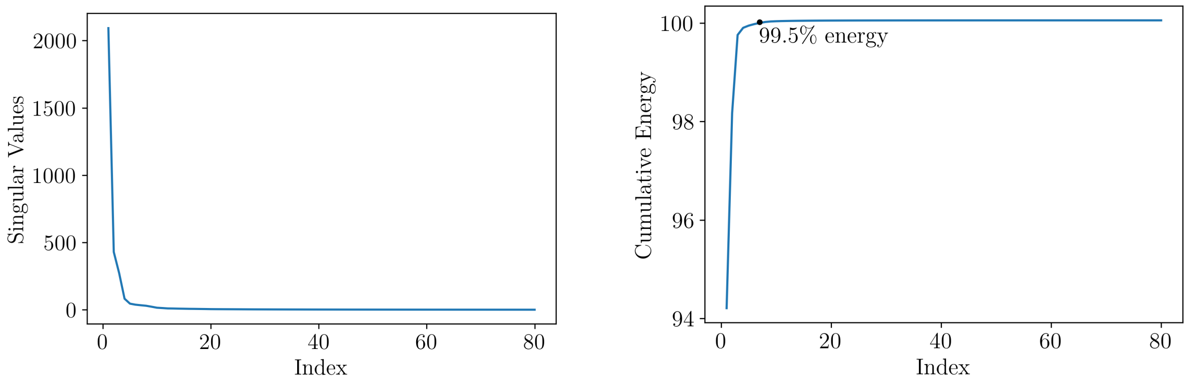
Table 1 Model specifications

Model Type	Max Refinement	Max Adaptations	Error Tolerance	Cycles/Adaptation	Cost (# HF)
HF	7	12	1e-3	175	1
LF	5	2	5e-3	50	1/127

C. Results for Projection-Enabled Multifidelity Linear Regression

We first analyze the dimensionality reduction on our training datasets of $N_{\mathrm{HF}} = 10$ and $N_{\mathrm{LF}} = 80$ to select an appropriate lower-dimensional subspace size. Figures 2 and 3 show the singular value decay and the cumulative energy plots for the LF and HF data, respectively. We show the median of 50 repetitions of SVD computations and the 25th and 75th percentile shaded around the median curve. There is not much variability in the singular values as seen from the plots. We use a tolerance of $\epsilon = 0.995$ for the cumulative energy to decide the size of the low-dimensional subspace using Eqn. (2), which leads to $k = 7$ for most LF training datasets and $k = 4$ for most HF training data sets. This facilitates the use of lower dimensional representations of the data for the surrogate models to be trained on, without significant loss of information.

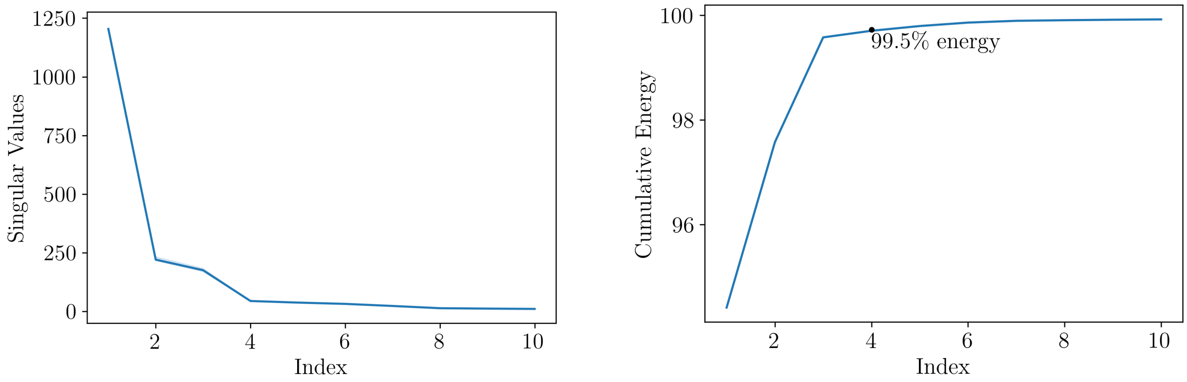
We apply the three MF linear regression methods described in Section III to the prediction of the pressure field on the IC3X testbed hypersonic vehicle. We evaluate the performance of a surrogate model through the normalized L2



(a) Singular Values

(b) Cumulative energy

Fig. 2 SVD on 80 LF training data



(a) Singular Values

(b) Cumulative energy

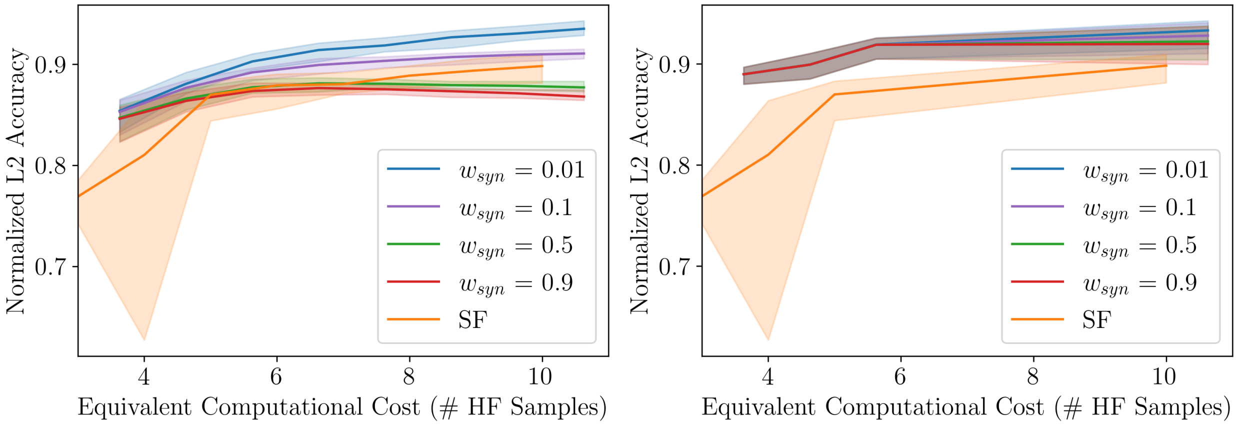
Fig. 3 SVD on 10 HF training data

accuracy metric given by $(1 - \epsilon_{\text{L2}})$, where the normalized L2 error ϵ_{L2} is defined by

$$\epsilon_{\text{L2}} = \frac{1}{N_{\text{HF}}^{\text{test}}} \sum_{i=1}^{N_{\text{HF}}^{\text{test}}} \frac{||\boldsymbol{y}_i - \hat{\boldsymbol{y}}_i||_2}{||\boldsymbol{y}_i||_2}, \tag{5}$$

where $||.||_2$ is the L2 vector norm, $\boldsymbol{y}_i$ is the HF model solution at i^{th} test sample, and $\hat{\boldsymbol{y}}_i$ is the surrogate prediction at i^{th} test sample. Note that the results for the single-fidelity (SF) surrogate model refer to the linear regression which was trained on the HF pressure field data only. Since the surrogate models were trained on 50 repetitions of the training and test dataset, we present the median, 25th, and 75th percentiles of the test accuracies. For the SF model, the order of the polynomial was limited by the number of samples available – limiting the choice to a linear equation in all cases. The MF linear regression with the additive structure also used a linear polynomial since it is trained on the same amount of HF data albeit with the discrepancy added. Lastly, both the MF surrogate models using the data augmentation methods were able to be trained using a polynomial of order two since the number of samples available to train was larger by the nature of the algorithms.

We first study the impact of different sample weighting schemes on the results of the two data augmentation methods as shown in Figure 4. To illustrate the effect of sample weighting schemes, a weighting scheme of $w_{\text{HF}} = 1, w_{\text{syn}} = \{0.9, 0.5, 0.1, 0.01\}$ is tested. We find that the direct data augmentation method is sensitive to the choice of w_{syn}, with a variation of up to 10% in median accuracy. On the contrary, the explicit map data augmentation method is less sensitive to changes in the sample weight, with a variation of up to 2% in median accuracy. In both cases, optimal selection of the synthetic sample weight could improve the accuracy of the MF linear regression using data augmentation.



(a) Direct Augmentation

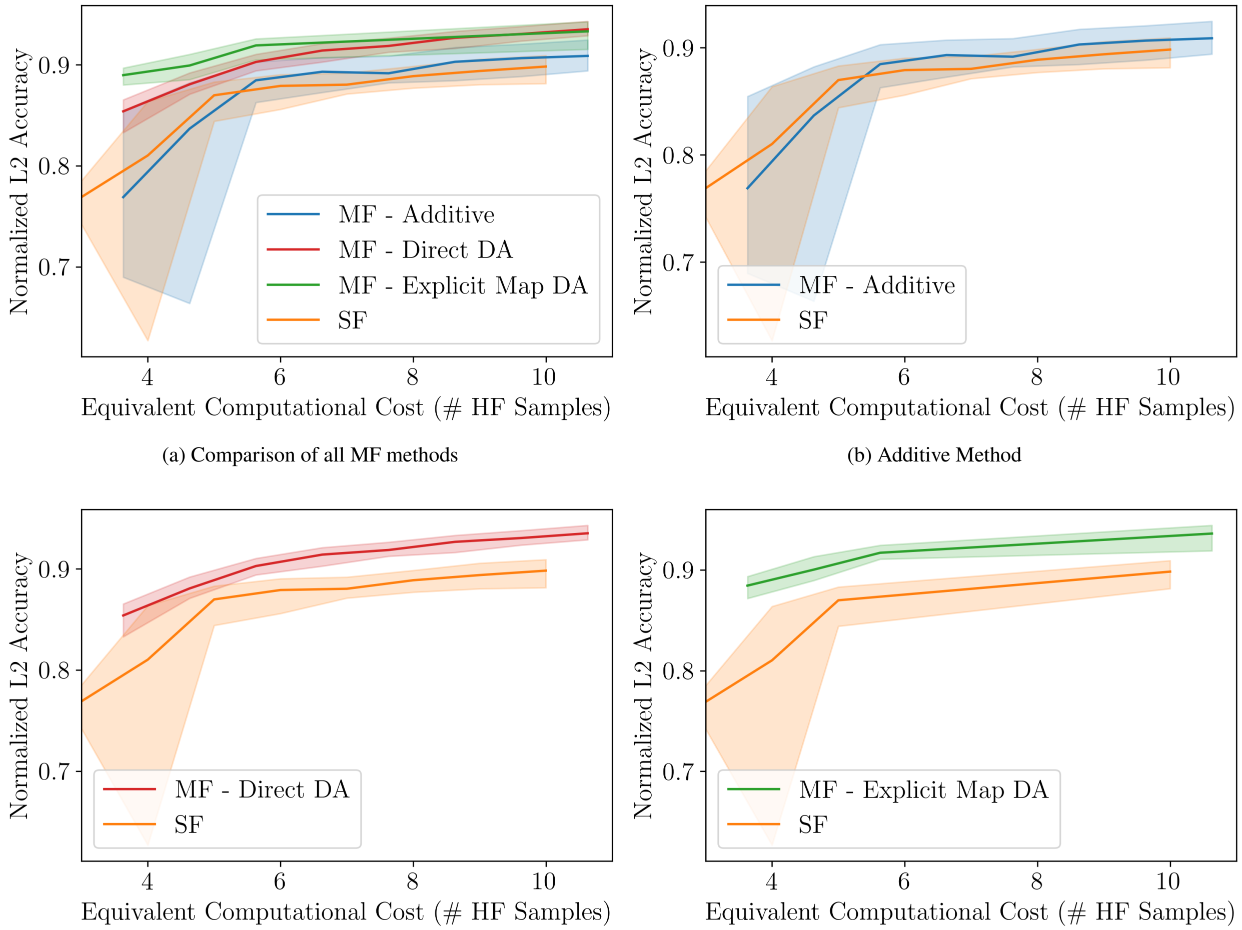
(b) Explicit Map

Fig. 4 Comparison of sample weighting schemes for MF linear regression using data augmentation

Figure 5 shows the comparison of the three different MF linear regression methods proposed in this work with the SF surrogate model. The additive MF method performs similar to the SF linear regression and does not offer significant increase in accuracy for this application. The additive approach shows modest improvements for $N_{\mathrm{HF}} \geq 5$ samples. In contrast, both the data augmentation techniques (using $w_{\mathrm{syn}} = 0.01$) perform better than the additive approach and show significant improvement in accuracy over the SF linear regression for equivalent computational cost. Furthermore, the robustness of both the MF linear regression models with data augmentation is markedly better than the SF surrogate model. This is to be expected as the MF surrogate model sees more data when training the linear regression model. The extra LF samples in data augmentation methods are of course not fully representative of the HF model, so we use sample weights of 1 for the HF samples and $w_{\mathrm{syn}} = 0.01$ for the synthetic data generated from the LF samples. The MF method with explicit map for data augmentation performs the best of the three presented here, specifically in the fewer data regime of $N_{\mathrm{HF}} \in [3, 5]$ HF samples. Table 2 provides the median accuracies of each regression method for $N_{\mathrm{HF}} =$ 3, 5, and 10 HF samples. We can see that the data augmentation technique using explicit map leads to an improvement of approximately 12% compared to the SF model for $N_{\mathrm{HF}} = 3$ HF samples and 3.6% compared to the SF model for $N_{\mathrm{HF}} = 10$ HF samples.

Table 2 Multifidelity linear regression results

Model Type	# LF Samples	# HF Samples	Median Normalized L2 Test Accuracy
SF	-	3	0.769
	-	5	0.870
	-	10	0.898
MF - Additive	80	3	0.765
		5	0.892
		10	0.919
MF - Direct data augmentation	80	3	0.854
		5	0.903
		10	0.935
MF - Explicit map data augmentation	80	3	0.890
		5	0.919
		10	0.934



(a) Comparison of all MF methods

(b) Additive Method

(c) Data augmentation: direct augmentation

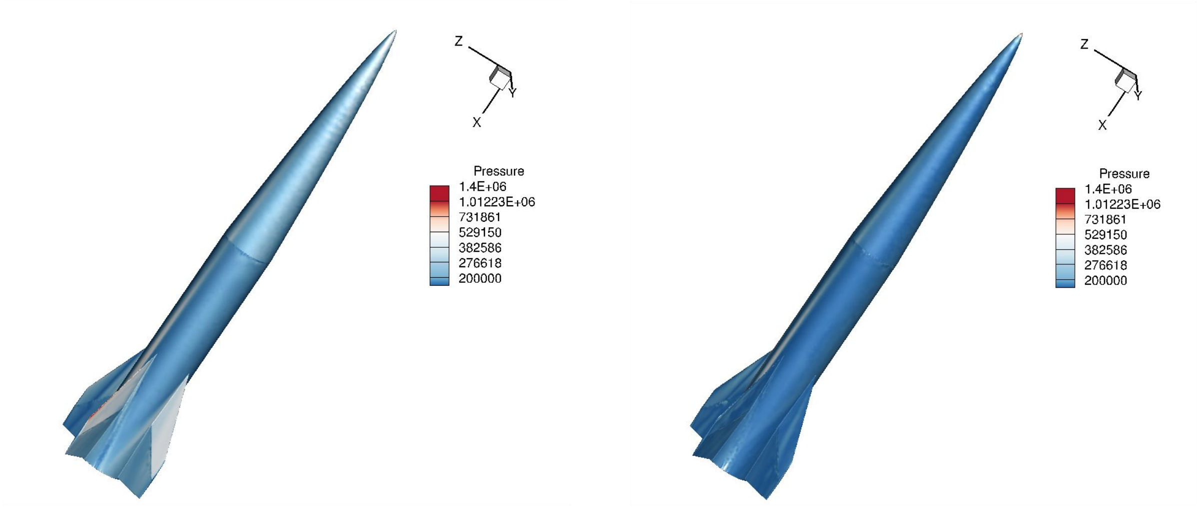
(d) Data augmentation: explicit mapping

Fig. 5 Comparison of MF linear regression methods to baseline SF linear regression (DA denotes data augmentation in the legends)

Finally, we look at a comparison of the absolute errors in pressure prediction between the SF surrogate and the MF surrogate using the direct data augmentation method with explicit map. For a random test sample, we predict the pressure field using the surrogates and show the absolute error compared to the HF model simulation. We show a contour plot of the errors on the vehicle body in Figure 6, providing some context for the gains the MF surrogate model nets.

V. Conclusion

This work presents and contrasts multifidelity linear regression methods for problems in the ultra low-data regime with two overarching ideas: (i) using discrepancy/additive structure and (ii) using data augmentation. In the additive structure for MF linear regression, we use a linear regression model to calibrate the LF data and better align it to the HF data based on the Kennedy O'Hagan approach. In the MF linear regression using data augmentation, we transform the LF data in two different ways and augment the transformed data to the HF dataset to perform a weighted least squares linear regression. In all these methods we embed dimensionality reduction through the proper orthogonal decomposition to tackle high-dimensional outputs. A numerical example on the prediction of the pressure load upon a hypersonic vehicle in-flight is used to compare and contrast the various MF approaches. For this application and HF samples in the range of 3 to 10, we find that the additive approach does not substantially improve the accuracy compared to the baseline SF surrogate model. The data augmentation techniques produce robust and accurate surrogate models,



(a) SF Surrogate Model Errors

(b) MF Surrogate Model Errors

Fig. 6 Comparison of errors in pressure field prediction

with up to approximately 3 – 12% in median accuracy gain in the low-data regime as compared to the SF surrogate. The direct data augmentation method had comparable accuracy to the explicit mapping method, but showed more sensitivity to the selection of the synthetic data weight in the weighted least squares regression. One possible future direction is to explore improvements in the performance of the data augmentation methods through an optimal hyperparameter selection strategy for the weight associated with the synthetic data in the weighted least squares regression.

Acknowledgments

This work has been supported in part by the ARPA-E Differentiate award number DE-AR0001208 and AFOSR grant FA9550-21-1-0089 under the NASA University Leadership Initiative (ULI).

References

[1] Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Willcox, K., and Lee, S., “Workshop Report on Basic Research Needs for Scientific Machine Learning: Core Technologies for Artificial Intelligence,” 2019. https://doi.org/10.2172/1478744.

[2] Peherstorfer, B., Willcox, K., and Gunzburger, M., “Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization,” *SIAM Review*, Vol. 60, No. 3, 2018, pp. 550–591. https://doi.org/10.1137/16M1082469.

[3] Myers, R., and Montgomery, D., *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, Wiley Series in Probability and Statistics, Wiley, 1995. URL https://books.google.com/books?id=7xvvAAAAMAAJ.

[4] Seber, G. A., and Lee, A. J., *Linear regression analysis*, John Wiley & Sons, 2012.

[5] Montgomery, D., Peck, E., and Vining, G., *Introduction to Linear Regression Analysis*, Wiley Series in Probability and Statistics, Wiley, 2021. URL https://books.google.com/books?id=tCIgEAAAQBAJ.

[6] Simpson, T., Poplinski, J., Koch, P. N., and Allen, J., “Metamodels for Computer-based Engineering Design: Survey and recommendations,” *Engineering with Computers*, Vol. 17, No. 2, 2001, pp. 129–150. https://doi.org/10.1007/pl00007198.

[7] Box, G. E. P., and Wilson, K. B., “On the Experimental Attainment of Optimum Conditions,” *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 13, No. 1, 1951, pp. 1–38. https://doi.org/https://doi.org/10.1111/j.2517-6161.1951.tb00067.x.

[8] Madsen, J. I., Shyy, W., and Haftka, R. T., “Response Surface Techniques for Diffuser Shape Optimization,” *AIAA Journal*, Vol. 38, No. 9, 2000, pp. 1512–1518. https://doi.org/10.2514/2.1160.

[9] Nakamura, T., Fukami, K., and Fukagata, K., "Identifying key differences between linear stochastic estimation and neural networks for fluid flow regressions," *Scientific Reports*, Vol. 12, No. 1, 2022. https://doi.org/10.1038/s41598-022-07515-7.

[10] Hosder, S., Watson, L. T., Grossman, B., Mason, W. H., Kim, H., Haftka, R. T., and Cox, S. E., "Polynomial Response Surface Approximations for the Multidisciplinary Design Optimization of a High Speed Civil Transport," *Optimization and Engineering*, Vol. 2, No. 4, 2001, pp. 431–452. https://doi.org/10.1023/a:1016094522761.

[11] Mack, Y., Goel, T., Shyy, W., and Haftka, R., "Surrogate Model-Based Optimization Framework: A Case Study in Aerospace Design," *Studies in Computational Intelligence*, Springer Berlin Heidelberg, 2007, pp. 323–342. https://doi.org/10.1007/978-3-540-49774-5_14.

[12] Ladický, L., Jeong, S., Solenthaler, B., Pollefeys, M., and Gross, M., "Data-Driven Fluid Simulations Using Regression Forests," *ACM Trans. Graph.*, Vol. 34, No. 6, 2015. https://doi.org/10.1145/2816795.2818129.

[13] Balabanov, V., Grossman, B., Watson, L., Mason, W., and Haftka, R., "Multifidelity response surface model for HSCT wing bending material weight," *7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization*, 1998, p. 4804.

[14] Madsen, J. I., and Langthjem, M., "Multifidelity Response Surface Approximations for the Optimum Design of Diffuser Flows," *Optimization and Engineering*, Vol. 2, No. 4, 2001, pp. 453–468. https://doi.org/10.1023/a:1016046606831.

[15] Vitali, R., Haftka, R. T., and Sankar, B. V., "Multi-fidelity design of stiffened composite panel with a crack," *Structural and Multidisciplinary Optimization*, Vol. 23, No. 5, 2002, pp. 347–356.

[16] Zhang, Y., Kim, N. H., Park, C., and Haftka, R. T., "Multifidelity Surrogate Based on Single Linear Regression," *AIAA Journal*, Vol. 56, No. 12, 2018, pp. 4944–4952. https://doi.org/10.2514/1.j057299.

[17] Park, C., Haftka, R. T., and Kim, N. H., "Remarks on multi-fidelity surrogates," *Structural and Multidisciplinary Optimization*, Vol. 55, No. 3, 2016, pp. 1029–1050. https://doi.org/10.1007/s00158-016-1550-y.

[18] Meng, X., and Karniadakis, G. E., "A composite neural network that learns from multi-fidelity data: Application to function approximation and inverse PDE problems," *Journal of Computational Physics*, Vol. 401, 2020, p. 109020. https://doi.org/https://doi.org/10.1016/j.jcp.2019.109020.

[19] Guo, M., Manzoni, A., Amendt, M., Conti, P., and Hesthaven, J. S., "Multi-fidelity regression using artificial neural networks: Efficient approximation of parameter-dependent output quantities," *Computer Methods in Applied Mechanics and Engineering*, Vol. 389, 2022, p. 114378. https://doi.org/10.1016/j.cma.2021.114378.

[20] Forrester, A. I., Sóbester, A., and Keane, A. J., "Multi-fidelity optimization via surrogate modelling," *Proceedings of the royal society a: mathematical, physical and engineering sciences*, Vol. 463, No. 2088, 2007, pp. 3251–3269.

[21] Le Gratiet, L., and Garnier, J., "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity," *International Journal for Uncertainty Quantification*, Vol. 4, No. 5, 2014.

[22] Kennedy, M. C., and O'Hagan, A., "Predicting the Output from a Complex Computer Code When Fast Approximations Are Available," *Biometrika*, Vol. 87, No. 1, 2000, pp. 1–13. URL http://www.jstor.org/stable/2673557.

[23] Benner, P., Gugercin, S., and Willcox, K., "A survey of projection-based model reduction methods for parametric dynamical systems," *SIAM Review*, Vol. 57, No. 4, 2015, pp. 483–531.

[24] Swischuk, R., Mainini, L., Peherstorfer, B., and Willcox, K., "Projection-based model reduction: Formulations for physics-based machine learning," *Computers & Fluids*, Vol. 179, 2019, pp. 704–717. https://doi.org/https://doi.org/10.1016/j.compfluid.2018.07.021.

[25] Guo, M., and Hesthaven, J. S., "Data-driven reduced order modeling for time-dependent problems," *Computer Methods in Applied Mechanics and Engineering*, Vol. 345, 2019, pp. 75–99.

[26] O'Leary-Roseberry, T., Villa, U., Chen, P., and Ghattas, O., "Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs," *Computer Methods in Applied Mechanics and Engineering*, Vol. 388, 2022, p. 114199. https://doi.org/10.1016/j.cma.2021.114199.

[27] Pasiliao, C. L., Sytsma, M. J., Neergaard, L., Witeof, Z., and Trolier, J. W., "Preliminary Aero-thermal Structural Simulation," *14th AIAA Aviation Technology, Integration, and Operations Conference*, American Institute of Aeronautics and Astronautics, 2014. https://doi.org/10.2514/6.2014-2292.

[28] Witeof, Z., and Neergaard, L., “Initial concept 3.0 finite element model definition,” *Eglin Air Force Base, Air Force Research Laboratory*, AFRL-RWWV-TN-2014-0013, FL, 2014.

[29] NASA Ames Research Center, “Automated Triangle Geometry Processing for Surface Modeling and Cartesian Grid Generation (Cart3D),” , Accessed 2022. URL https://software.nasa.gov/software/ARC-14275-1.

[30] Aftosmis, M., Berger, M., and Adomavicius, G., “A parallel multilevel method for adaptively refined Cartesian grids with embedded boundaries,” *38th AIAA Aerospace Sciences Meeting and Exhibit*, 2000, pp. 2000–0808. https://doi.org/10.2514/6.2000-808.

[31] Aftosmis, M. J., Berger, M. J., and Melton, J. E., “Robust and Efficient Cartesian Mesh Generation for Component-Based Geometry,” *AIAA Journal*, Vol. 36, No. 6, 1998, pp. 952–960. https://doi.org/10.2514/2.464.