

Projection-based multifidelity linear regression for data-scarce applications

Vignesh Sella^{1*}, Julie Pham², Karen Willcox^{1,2,3},
Anirban Chaudhuri¹

^{1*}Oden Institute for Computational Engineering and Sciences,
University of Texas at Austin, 78712, TX, USA.

²Aerospace Engineering and Engineering Mechanics, University of Texas
at Austin, 78712, TX, USA.

³Santa Fe Institute, Santa Fe, 87501, NM, USA.

*Corresponding author(s). E-mail(s): vignesh.sella@austin.utexas.edu;
Contributing authors: julie.pham@austin.utexas.edu;
kwillcox@oden.utexas.edu; anirbanc@oden.utexas.edu;

Abstract

Surrogate modeling for systems with high-dimensional quantities of interest remains challenging, particularly when training data are costly to acquire. This work develops multifidelity methods for multiple-input multiple-output linear regression targeting data-limited applications with high-dimensional outputs. Multifidelity methods integrate many inexpensive low-fidelity model evaluations with limited, costly high-fidelity evaluations. We introduce two projection-based multifidelity linear regression approaches with linear and nonlinear features that leverage principal component basis vectors for dimensionality reduction and combine multifidelity data through: (i) a direct data augmentation using low-fidelity data, and (ii) a data augmentation incorporating explicit linear corrections between low-fidelity and high-fidelity data. The data augmentation approaches combine high-fidelity and low-fidelity data into a unified training set and train the linear regression model through weighted least squares with fidelity-specific weights. We introduce a proximity-based weighting scheme with automatic weight selection strategy through cross-validation. The proposed multifidelity linear regression methods are demonstrated on approximating the surface pressure field of a hypersonic vehicle in flight and the temperature field on an aircraft disc braking system. In an ultra low-data regime of no more than twelve high-fidelity samples, multifidelity linear regression achieves approximately 2% – 12%

improvement in median accuracy and a higher R^2 score relative to single-fidelity methods at comparable computational cost.

Keywords: multifidelity, linear regression, scientific machine learning, surrogate modeling, principal component analysis, data augmentation

1 Introduction

An important challenge in scientific machine learning is to develop methods that can exploit and maximize the amount of learning possible from scarce data [1–4]. The need for such methods arises often in science and engineering, 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 [5]. Surrogate models that approximate the solutions to HF models can facilitate the design and analysis process; however, lack of sufficient HF data in tandem with high-dimensional quantities of interest adversely affect surrogate model accuracy. We propose multifidelity (MF) linear regression methods that leverage abundant low-cost, lower-fidelity (LF) data alongside severely limited HF data to construct linear regression models. These models operate within a reduced-dimensional subspace, obtained through the principal component analysis (PCA), to effectively handle both training data scarcity and the high dimensionality (on the order of tens of thousands of quantities of interest) inherent in our problem setting.

Linear regression has been widely utilized as a surrogate modeling approach in aerospace applications due to its simplicity and interpretability. We note that linear regression is a parametric surrogate modeling technique that encompasses a broad class of models that are *linear in their parameters but can include features that are arbitrarily nonlinear functions of the input variables* [6]. Traditionally, linear regression methods such as the response surface methodology (RSM) employ low-order polynomial approximations for problems characterized by a modest number of input variables (typically fewer than ten) and limited datasets ($\Theta(10^2)$ to $\Theta(10^3)$) due to computational costs [7–10]. In addition, many works have explored more data-intensive approaches, such as random forests or neural networks, that leverage significantly larger datasets, demonstrating strong predictive capabilities but requiring substantial computational resources [11, 12]. However, acquiring extensive HF training data often remains impractical for typical aerospace design applications without considerable computational investment. This motivates alternative approaches capable of working effectively in ultra low-data regime.

MF linear regression techniques that leverage data of varying fidelity levels provide an effective means to reduce reliance on costly HF training data. Early MF approaches modeled the relationship between low- and high-fidelity outputs through additive or multiplicative correction factors applied to the LF model [13–16]. These correction, or discrepancy, models were later extended using polynomial RSMs and other surrogate forms to capture more complex relationships [17–21]. Broadly, such methods

treat the HF response as a modified version of the LF prediction combined with an additional term that models the residual error [22–24]. Recent work has also explored linear (and polynomial) MF regression strategies that utilize ordinary least squares (OLS)- [25] and moving least squares (MLS)-based methods [26]. Both approaches incorporate scalar LF model outputs as either a basis function or a feature with additive or multiplicative correction terms. These correction-based methods are effective for scalar outputs, but their achievable model complexity is restricted by the number of available HF samples and dimensionality of the problem at hand. While in many engineering analyses only integrated or aggregate quantities (e.g., lift, drag, or maximum stress) are ultimately reported, access to the full-field response remains valuable. Full-field predictions provide flexibility for downstream tasks that can require different derived quantities or spatially localized information without the need to retrain or rerun simulations (or surrogates). They can enable identification of localized critical regions (e.g., areas of high pressure, stress, or temperature) that are important for tasks related to reliability, safety, and design optimization. In addition, full-field predictions can serve as inputs to coupled multi-physics analyses, where spatial fields of quantities such as pressure, velocity, or temperature are used as inputs. Our work introduces a unified framework using weighted least squares (WLS) that combines LF and HF data in a single global regression with fidelity-informed weighting, enabling higher-order polynomial fits without the co-location or correction constraints of prior methods and can predict high-dimensional outputs of interest.

In the parametric multifidelity surrogate modeling context, recent studies on multifidelity neural networks have primarily focused on using larger quantities of data, often via discrepancy modeling, to improve estimation through a variety of machine learning techniques [27–31]. These methods typically require datasets on the order of $\Theta(10^2)$ – $\Theta(10^3)$ samples and are thus not applicable within the targeted sparse data regime. For the limited data setting, Bayesian nonparametric surrogate models, such as multifidelity Gaussian process regression or co-kriging, can be another surrogate modeling approach and have been investigated for low-dimensional output setting [32, 33]. However, given the cubic complexity with respect to the number of samples, Gaussian process regression can be computationally expensive as compared to linear regression when the training data contains numerous LF samples [34].

In this work, we develop a MF linear regression method that can address regression problems involving high-dimensional outputs with sparse training data. The proposed method projects the outputs onto a lower-dimensional subspace obtained via PCA, facilitating effective modeling despite severely limited HF data. Note that the dimensionality of the input space in the applications considered in this work is significantly smaller than the output space. Our key methodological contribution is a data augmentation method that integrates weighted least squares (WLS) with the dimensionality reduction and an automated weighting mechanism to explicitly incorporate LF data within a unified training process. We provide a closed-form expression for the projection-based MF linear regression coefficients following the derivation of the WLS solution. Specifically, we introduce and analyze two methods for MF data augmentation: (i) direct data augmentation by combining fidelity sources, and

(ii) data augmentation employing an explicit mapping between low- and HF outputs. We present a proximity-based weighting scheme for WLS and an automatic weight selection strategy using cross-validation. The WLS approach and adaptive data-driven weighting schemes enable appropriate utilization of LF data to improve predictive accuracy in high-dimensional, data scarce regimes. We benchmark our proposed methods against state-of-the-art MF method with additive structure [25, 35]. We incorporate output dimensionality reduction into the additive MF method to provide a fair basis for comparison. We demonstrate the effectiveness of the proposed MF methods through their application on two engineering applications: predicting pressure load distributions on a hypersonic testbed vehicle and predicting the temperature field of aircraft disc brakes after a braking period.

The remainder of this paper is structured as follows. Section 2 introduces the MF regression problem setup and the dimensionality reduction. Section 3 details the proposed MF linear regression method using data augmentation and an extension of the state-of-the-art additive MF regression strategy with projections is presented in Appendix A for comparison. We present the numerical experiments on a hypersonic testbed vehicle application in Section 4 and on an aircraft disc brake application in Section 5. Finally, Section 6 provides concluding remarks.

2 Multifidelity regression: Background and problem formulation

We consider an MF regression setting involving training datasets obtained from models with different fidelity levels: an HF model that provides accurate predictions but is computationally expensive, and LF models that are computationally less costly but yield less accurate predictions.

2.1 Background: Linear regression with dimensionality reduction

Let the d -dimensional inputs to a system be denoted by $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$, where \mathcal{X} is the input space, and the output quantity of interest be $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^m$, defined on the output space \mathcal{Y} . In our target applications, \mathbf{y} is a high-dimensional quantity, with m typically on the order of tens of thousands. Due to the high-dimensionality of the outputs and limited HF data, we employ PCA to reduce the output dimension prior to regression. Similar projection-based approaches have been applied in the context of parametric reduced-order models [36–38], as well as in neural network-based models [29, 39].

Dimensionality reduction via PCA. For the training data matrix $\mathbf{Y} \in \mathbb{R}^{m \times N}$ with N samples, the PCA basis vectors are obtained by standard PCA projection [6]. We compute the principal components through the singular value decomposition (SVD). Given the row-wise sample mean of the training data $\bar{\mathbf{y}} \in \mathbb{R}^{m \times 1}$, we define $\bar{\mathbf{Y}} = \bar{\mathbf{y}} \mathbf{1}_N^\top$ where $\mathbf{1}_N \in \mathbb{R}^N$ is the length- N column vector with all entries set to unity. For $N < m$, the thin SVD of the centered training data matrix is written as

$$\mathbf{Y} - \bar{\mathbf{Y}} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \quad (1)$$

where $\mathbf{U} \in \mathbb{R}^{m \times N}$ and $\mathbf{V} \in \mathbb{R}^{N \times N}$ are orthogonal matrices, and $\mathbf{\Sigma} \in \mathbb{R}^{N \times N}$ is a diagonal matrix with non-decreasing entries of the singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_N \geq 0$. Given the left singular vectors \mathbf{U} , the reduced basis for projection to a lower-dimensional subspace of size $k \leq N$ is the first k columns $\mathbf{U}_k \in \mathbb{R}^{m \times k}$. The projection of the set of output samples \mathbf{Y} on the low-dimensional subspace is given by the reduced states $\mathbf{C} \in \mathbb{R}^{k \times N}$, defined as

$$\mathbf{C} = \mathbf{U}_k^\top (\mathbf{Y} - \bar{\mathbf{Y}}). \quad (2)$$

The dimension k is chosen such that the cumulative variance captured by the first k principal components is larger than a specified tolerance of ϵ as given by

$$\frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^N \sigma_i^2} > \epsilon, \quad (3)$$

where σ_i is the i -th singular value.

Projection-based linear regression. The regression problem considered in this work is a linear-regression-based surrogate model in the reduced-dimensional space $f: \mathbb{R}^d \rightarrow \mathbb{R}^k$, parameterized by regression coefficients β . To address the high dimensionality of the output space, we perform regression in the reduced space defined by the first k principal components obtained from PCA. Note that one could also apply dimensionality reduction to the inputs in addition to the outputs as shown by Sun [40]. Alternate methods for dimensionality reduction in multivariate linear regression [41, 42] are also feasible and composable with the MF methods presented in the following section. We employ PCA because it is interpretable, widely used, and has been shown to be effective across a broad range of applications [43].

We use training data projected to the reduced space using Eq. (2) to obtain the projection-based linear regression model $f(\mathbf{x}; \beta)$ for k -dimensional outputs as

$$f(\mathbf{x}; \beta) = \Phi(\mathbf{x})^\top \beta,$$

where $\Phi(\mathbf{x})^\top \in \mathbb{R}^{p+1}$ is a $p+1$ -dimensional feature vector that can include nonlinear transformations of the input (e.g., polynomial basis terms) and $\beta \in \mathbb{R}^{p+1 \times k}$ is the matrix of regression coefficients to be estimated. The surrogate model is therefore linear in the regression coefficients and can be trained using either ordinary or weighted least squares, depending on the MF regression methodology presented in the next section. We reconstruct the full-dimensional output from the regression predictions as

$$\hat{\mathbf{y}}(\mathbf{x}^*) = \mathbf{U}_k f(\mathbf{x}^*; \beta) + \bar{\mathbf{y}}, \quad (4)$$

where $\hat{\mathbf{y}}(\mathbf{x}^*)$ is the approximation of the true HF output for any new input $\mathbf{x}^* \in \mathcal{X}$.

2.2 Multifidelity regression problem formulation

For ease of exposition, we consider a bifidelity setup, but the general idea can be extended to more than two fidelity levels. To distinguish between data originating

from the HF and LF models, we define $\mathbf{X}_{\text{HF}} := [\mathbf{x}_1^{(\text{HF})}, \dots, \mathbf{x}_{N_{\text{HF}}}^{(\text{HF})}] \in \mathbb{R}^{d \times N_{\text{HF}}}$ and $\mathbf{Y}_{\text{HF}} := [\mathbf{y}_1^{(\text{HF})}, \dots, \mathbf{y}_{N_{\text{HF}}}^{(\text{HF})}] \in \mathbb{R}^{m \times N_{\text{HF}}}$ with analogous definitions for the LF data $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$. In the applications of interest, we have $N_{\text{HF}} \ll N_{\text{LF}}$ and $N_{\text{HF}} \ll m$, reflecting the high computational cost of HF evaluations and the high-dimensionality of the output space.

The core supervised learning problem in this work is to construct a linear-regression-based surrogate model that accurately predicts the HF output $\mathbf{y}_{\text{HF}}(\mathbf{x}^*)$ for new inputs \mathbf{x}^* , by leveraging the bifidelity training dataset $(\mathbf{X}_{\text{HF}}, \mathbf{Y}_{\text{HF}})$ and $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$. The challenge of high output dimensionality ($m \gg N_{\text{HF}}$), and the limited number of HF samples makes direct regression in \mathbb{R}^m ill-posed. Our approach mitigates these challenges by first projecting the HF and LF outputs to a lower-dimensional subspace using PCA, and then developing multifidelity regression methods in this reduced space as described in the following section.

3 Projection-based multifidelity linear regression via data augmentation

In this section, we develop a MF linear regression approach via data augmentation using the projected data. We first present two ways of synthetic data generation for data augmentation in Section 3.1 followed by the proximity-based weighting technique and the WLS approach developed for the MF linear regression in Section 3.2. We then present an automated weight selection strategy through cross-validation in Section 3.3.

3.1 Synthetic data for data augmentation

The sparsity of HF data poses a significant challenge when attempting to fit more expressive surrogate models, such as polynomial regression with higher-order basis functions, because the HF dataset alone may not sufficiently constrain the model or allow for meaningful generalization. In contrast, training a surrogate model only on LF data is more feasible and less expensive, albeit at the cost of reduced accuracy. This work addresses these limitations by developing an MF linear regression method that utilizes an augmented training dataset, consisting of HF data and synthetic data derived from LF evaluations, to better constrain the regression model. The MF regression via data augmentation provides additional information about the underlying system response in regions of the input space insufficiently covered by HF samples. Furthermore, the MF approach facilitates the training of regression models with larger number of regression coefficients, for example, enabling the use of higher-degree polynomial bases beyond what the HF data alone would support.

Let $(\mathbf{X}_{\text{LF}}^{\text{syn}}, \mathbf{Y}_{\text{LF}}^{\text{syn}})$ denote the synthetic data used in data augmentation. We construct the synthetic data by one of two approaches:

1. *Direct augmentation*: The LF data are used directly as synthetic training data, i.e., $(\mathbf{X}_{\text{LF}}^{\text{syn}}, \mathbf{Y}_{\text{LF}}^{\text{syn}}) = (\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$.
2. *Explicit mapping*: A learned linear correction map is applied to the LF data to approximate the HF behavior at \mathbf{X}_{LF} . This mapping is constructed by training a

linear model g between reduced-order representations of the LF and HF outputs in a shared low-dimensional space.

The remainder of this section defines the explicit mapping approach, where we choose to model the relationship between LF and HF outputs in the subspace spanned by the reduced basis derived from the HF data as a linear transformation. We define a linear model $g : \mathbb{R}^k \rightarrow \mathbb{R}^k$ that maps reduced LF states to reduced HF states using the HF reduced basis \mathbf{U}_k^{HF} to perform the dimensionality reduction.

To train the model g , we need co-located HF and LF samples. When the LF samples are not co-located with the HF samples, we use a LF surrogate model to obtain the LF predictions at \mathbf{X}_{HF} . Let $f_{\text{LF}} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ denote the linear regression surrogate model trained on the projected LF dataset $(\mathbf{X}_{\text{LF}}, \mathbf{C}_{\text{LF}})$, where $\mathbf{C}_{\text{LF}} \in \mathbb{R}^{k \times N_{\text{LF}}}$ are the reduced LF states obtained via PCA as given by Eq. (2). The predictions of reduced LF states at the HF input locations \mathbf{X}_{HF} are obtained by $f_{\text{LF}}(\mathbf{X}_{\text{HF}})$ and reconstructed to the full-dimensional output space as $\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}}$ to obtain co-located LF output predictions. Since the linear mapping g operates within the subspace spanned by the HF reduced basis, we project the co-located LF predictions using the HF reduced basis to obtain the coordinate-transformed reduced LF states, $\hat{\mathbf{C}}_{\text{LF}}$, as

$$\hat{\mathbf{C}}_{\text{LF}} = (\mathbf{U}_k^{\text{HF}})^\top ((\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}}) - \bar{\mathbf{Y}}_{\text{HF}}). \quad (5)$$

The projection step in Eq. (5) serves to express the LF predictions in the HF reduced basis. Alternatively, one could explore methods such as manifold alignment to align the two subspaces and potentially provide better mappings between the two reduced states [44, 45]. Simultaneously, we compute the HF reduced states as

$$\mathbf{C}_{\text{HF}} = (\mathbf{U}_k^{\text{HF}})^\top (\mathbf{Y}_{\text{HF}} - \bar{\mathbf{Y}}_{\text{HF}}). \quad (6)$$

The linear mapping model g is then trained via OLS on the co-located dataset $(\hat{\mathbf{C}}_{\text{LF}}, \mathbf{C}_{\text{HF}})$, where we are choosing to model this as a low-rank linear relationship between the LF and HF reduced outputs. Note that if co-located data is already available, then one does not need to fit the LF surrogate model f_{LF} and can directly obtain $\hat{\mathbf{C}}_{\text{LF}} = (\mathbf{U}_k^{\text{HF}})^\top (\mathbf{Y}_{\text{LF}}(\mathbf{X}_{\text{HF}}) - \bar{\mathbf{Y}}_{\text{HF}})$ for training g .

Once trained, g is used to generate synthetic data, $\mathbf{Y}_{\text{LF}}^{\text{syn}}$, at all the LF input locations \mathbf{X}_{LF} by mapping the LF outputs as

$$\mathbf{Y}_{\text{LF}}^{\text{syn}} = \mathbf{U}_k^{\text{HF}} g(\mathbf{C}_{\text{LF}}) + \bar{\mathbf{Y}}_{\text{HF}} = \mathbf{U}_k^{\text{HF}} g\left((\mathbf{U}_k^{\text{LF}})^\top (\mathbf{Y}_{\text{LF}} - \bar{\mathbf{Y}}_{\text{LF}})\right) + \bar{\mathbf{Y}}_{\text{HF}}. \quad (7)$$

This produces the synthetic dataset $(\mathbf{X}_{\text{LF}}^{\text{syn}} = \mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}}^{\text{syn}})$ by explicit mapping, which is used for data augmentation in the MF regression method. We summarize this process in Alg. 1.

3.2 Weighted least squares using proximity-based weights

Given an augmented training dataset incorporating synthetic LF-derived samples, we train the MF surrogate model using weighted least squares regression to account for

Algorithm 1 Synthetic data generation via explicit linear mapping model

Input: HF and LF training data $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$ and $(\mathbf{X}_{\text{HF}}, \mathbf{Y}_{\text{HF}})$

Output: Synthetic data $\mathbf{Y}_{\text{LF}}^{\text{syn}}$ at inputs \mathbf{X}_{LF} from the LF to HF surrogate map

- 1: Project \mathbf{Y}_{LF} to obtain the reduced states $\mathbf{C}_{\text{LF}} = (\mathbf{U}_k^{\text{LF}})^\top (\mathbf{Y}_{\text{LF}} - \bar{\mathbf{Y}}_{\text{LF}})$ \triangleright see Eq. (2)
 - 2: Train the LF regression model f_{LF} on $(\mathbf{X}_{\text{LF}}, \mathbf{C}_{\text{LF}})$ using OLS
 - 3: Generate co-located LF predictions as $\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}}$ at HF sample location \mathbf{X}_{HF}
 - 4: Project co-located LF predictions to obtain the coordinate-transformed LF reduced states $\hat{\mathbf{C}}_{\text{LF}}$ via Eq. (5)
 - 5: Project \mathbf{Y}_{HF} to obtain the reduced states \mathbf{C}_{HF} using Eq. (6)
 - 6: Train LF \mapsto HF linear regression model g on $(\hat{\mathbf{C}}_{\text{LF}}, \mathbf{C}_{\text{HF}})$ using OLS
 - 7: Generate synthetic data $\mathbf{Y}_{\text{LF}}^{\text{syn}}$ at \mathbf{X}_{LF} locations using Eq. (7)
-

fidelity-dependent variance. Ordinary least squares (OLS) assumes homoscedasticity, or constant variance in the residuals, which does not hold in this setting, as synthetic samples derived from LF data are known *a priori* to be a less accurate approximation. To account for this expected heteroscedasticity, we instead apply WLS [46] with distinct weights assigned to HF and synthetic training samples. Specifically, we define a diagonal weight matrix $\mathbf{W} = \text{diag}(w_1, \dots, w_{N_{\text{HF}} + N_{\text{LF}}})$, where weights are assigned as

$$w_i = \begin{cases} 1, & i = 1, \dots, N_{\text{HF}} \\ h(w_{\text{syn}}) < 1, & i = N_{\text{HF}} + 1, \dots, N_{\text{HF}} + N_{\text{LF}}, \end{cases} \quad (8)$$

where $h(w_{\text{syn}})$ is a weighting function for LF training samples defined using the hyperparameters w_{syn} .

In the context of MF surrogate modeling, the HF model is regarded as the best available source of ground truth for a given quantity of interest. Thus, when applying least-squares linear regression, LF samples located near HF samples in the input space can be considered redundant or uninformative, since continuity ensures that proximity in the input space yields proximity of the outputs. Consequently, the LF data may introduce noise rather than useful information due to their inherent lower fidelity. This issue is particularly relevant when LF and HF datasets are fixed, which is the setting considered in this paper. In this context, the LF data introduces position-dependent variance, an instance of heteroscedasticity, which is precisely what WLS was designed to handle [46]. To mitigate this effect, we introduce a *proximity-based weighting scheme* that down-weights LF samples located near HF samples. The sample weight assigned to a given LF point depends on (1) its distance to the nearest HF point and (2) whether it originates from the LF or HF source. This approach allows the model to emphasize LF samples that fill gaps (alleviate epistemic uncertainty) in the HF dataset while discounting those that are likely redundant. We compare the proximity-based weighting scheme with a fixed weighting scheme. The weighting

function is then defined as

$$h(w_{\text{syn}}) = \begin{cases} w_{\text{syn}} & \text{fixed weights} \\ \sigma(\rho(\mathbf{x}^{\text{LF}}, \mathbf{x}^{\text{HF}}); w_{\text{syn}}) & \text{proximity-based weights,} \end{cases} \quad (9)$$

where $\rho : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a distance function (e.g., Euclidean distance) and $\sigma : \mathbb{R} \rightarrow [0, 1]$ is a monotonic transformation that maps distances to normalized weights. Suitable choices include Heaviside step functions, sigmoids, or any other similar function. In this work, we use a Heaviside step function to define $\sigma(\rho(\mathbf{x}^{\text{LF}}, \mathbf{x}^{\text{HF}}); w_{\text{syn}}) = w_{\text{syn}} \mathbf{1}_{\rho(\mathbf{x}^{\text{LF}}, \mathbf{x}^{\text{HF}}) \geq \tau}$, where $\mathbf{1}$ is an indicator function that sets the maximum value to w_{syn} and the minimum value to 0 depending on whether the distance from HF samples exceeds a threshold value of τ . We use Euclidean distance as the distance function $\rho(\cdot)$. The value of w_{syn} significantly impacts model performance and is selected via cross-validation, as described in Section 3.3.

The surrogate is trained in the projected output space defined by the HF reduced basis (see Section 2). The MF linear regression model denoted by $f_{\text{MF}} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ is given by

$$f_{\text{MF}}(\mathbf{x}; w_{\text{syn}}) = \Phi(\mathbf{x})^\top \hat{\beta}_{\text{MF}}(w_{\text{syn}}), \quad (10)$$

where $\hat{\beta}_{\text{MF}}(w_{\text{syn}})$ are the regression coefficients and has the explicit dependence on the synthetic sample weight since they are estimated using WLS. The regression model f_{MF} is trained on the augmented training dataset containing $N_{\text{HF}} + N_{\text{LF}}$ samples given by $([\mathbf{X}_{\text{HF}}, \mathbf{X}_{\text{LF}}^{\text{syn}}], [\mathbf{Y}_{\text{HF}}, \mathbf{Y}_{\text{LF}}^{\text{syn}}])$ as defined in Section 3.1. For brevity, when utilizing the data augmentation method, we define $\mathbf{X}_{\text{MF}} := [\mathbf{X}_{\text{HF}}, \mathbf{X}_{\text{LF}}^{\text{syn}}]$ as the independent variables. Similarly, we define $\mathbf{Y}_{\text{MF}} := [\mathbf{Y}_{\text{HF}}, \mathbf{Y}_{\text{LF}}^{\text{syn}}]$. Projecting these outputs yields reduced states,

$$\mathbf{C}_{\text{MF}} = (\mathbf{U}_k^{\text{HF}})^\top (\mathbf{Y}_{\text{MF}} - \bar{\mathbf{y}}_{\text{HF}} \mathbf{1}_{N_{\text{HF}} + N_{\text{LF}}}^\top). \quad (11)$$

The optimal regression coefficients when the MF linear regression model is trained on $(\mathbf{X}_{\text{MF}}, \mathbf{C}_{\text{MF}})$ using WLS with weights \mathbf{W} can be obtained in closed-form as

$$\hat{\beta}_{\text{MF}}^*(w_{\text{syn}}) = \left(\Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} \Phi(\mathbf{X}_{\text{MF}}) \right)^{-1} \Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} \mathbf{C}_{\text{MF}} \quad (12)$$

$$= \left(\Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} \Phi(\mathbf{X}_{\text{MF}}) \right)^{-1} \Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} (\mathbf{U}_k^{\text{HF}})^\top (\mathbf{Y}_{\text{MF}} - \bar{\mathbf{y}}_{\text{HF}} \mathbf{1}_{N_{\text{HF}} + N_{\text{LF}}}^\top), \quad (13)$$

where the derivation for the closed-form expression in Eq.(12) follows from the known WLS solution [6] and Eq.(13) substitutes the reduced states. The prediction at any new input location \mathbf{x}^* is made in the reduced space and then lifted to the full-dimensional output space as

$$\begin{aligned} \hat{\mathbf{y}}_{\text{MF}}(\mathbf{x}^*; w_{\text{syn}}) &= \mathbf{U}_k^{\text{HF}} f_{\text{MF}}(\mathbf{x}^*; w_{\text{syn}}) + \bar{\mathbf{Y}}_{\text{HF}} \\ &= \mathbf{U}_k^{\text{HF}} \Phi(\mathbf{x}^*)^\top \hat{\beta}_{\text{MF}}^*(w_{\text{syn}}) + \bar{\mathbf{y}}_{\text{HF}} \mathbf{1}_{N_{\text{HF}} + N_{\text{LF}}}^\top \end{aligned} \quad (14)$$

$$= \mathbf{U}_k^{\text{HF}} \Phi(\mathbf{x}^*)^\top \underbrace{(\Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} \Phi(\mathbf{X}_{\text{MF}}))^{-1} \Phi(\mathbf{X}_{\text{MF}})^\top \mathbf{W} \mathbf{C}_{\text{MF}}}_{\hat{\beta}_{\text{MF}}^*(w_{\text{syn}})} + \bar{\mathbf{y}}_{\text{HF}} \mathbf{1}_{N_{\text{HF}} + N_{\text{LF}}}^\top. \quad (15)$$

We summarize the data augmentation method for MF linear regression in Alg. 2.

Algorithm 2 Multifidelity linear regression via data augmentation

Input: HF and LF training data $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$ and $(\mathbf{X}_{\text{HF}}, \mathbf{Y}_{\text{HF}})$, synthetic sample weighting parameter w_{syn} , new input location for prediction \mathbf{x}^*

Output: Output predictions $\hat{\mathbf{y}}_{\text{MF}}$ at inputs \mathbf{x}^* from MF surrogate

- 1: Generate synthetic data by transforming the LF data: $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}}) \mapsto (\mathbf{X}_{\text{LF}}^{\text{syn}}, \mathbf{Y}_{\text{LF}}^{\text{syn}})$ \triangleright use Alg. 1 for the explicit mapping method
 - 2: Augment the training dataset to contain $N_{\text{HF}} + N_{\text{LF}}$ samples: $([\mathbf{X}_{\text{HF}}, \mathbf{X}_{\text{LF}}^{\text{syn}}], [\mathbf{Y}_{\text{HF}}, \mathbf{Y}_{\text{LF}}^{\text{syn}}])$
 - 3: Project $[\mathbf{Y}_{\text{HF}}, \mathbf{Y}_{\text{LF}}^{\text{syn}}]$ to obtain the reduced states of MF training data outputs \mathbf{C}_{MF} using Eq. (11)
 - 4: Set up sample weight matrix \mathbf{W} based on choice of sample weighting scheme using Eqs. (8) and (9)
 - 5: Train MF linear regression surrogate model f_{MF} on $([\mathbf{X}_{\text{HF}}, \mathbf{X}_{\text{LF}}^{\text{syn}}], \mathbf{C}_{\text{MF}})$ with weights \mathbf{W} using WLS \triangleright closed-form expression in Eq. (13)
 - 6: Predict $\hat{\mathbf{y}}_{\text{MF}}(\mathbf{x}^*)$ by reconstructing the output of $f_{\text{MF}}(\mathbf{x}^*; w_{\text{syn}})$ in the full-dimensional space defined in Eq. (14)
-

3.3 Cross-validation for optimal sample weight selection

The synthetic sample weighting function in Eq. (9) has a tunable hyperparameter $w_{\text{syn}} \in (0, 1)$. Rather than arbitrarily selecting w_{syn} or utilizing a heuristic-based approach, we select w_{syn} for the proximity-based weighting scheme by minimizing the prediction error using leave-one-out cross-validation (LOOCV). LOOCV was chosen rather than general k -fold cross-validation since we are working with a limited number of HF samples. For each HF training sample $i \in \{1, \dots, N_{\text{HF}}\}$, a model $f_{\text{MF}}(\cdot; w_{\text{syn}})$ is trained on the remaining data and the validation error for the held-out sample is defined as

$$\epsilon_{\text{LOOCV}}(\mathbf{y}_i^{\text{HF}}; w_{\text{syn}}) := \frac{\|\mathbf{y}_i^{\text{HF}} - \hat{\mathbf{y}}_i^{\text{MF}}(w_{\text{syn}})\|_2}{\|\mathbf{y}_i^{\text{HF}}\|_2}, \quad (16)$$

where $\hat{\mathbf{y}}_i^{\text{MF}}(w_{\text{syn}})$ denotes the prediction at \mathbf{x}_i^{HF} made by the model trained without sample i . The optimal weight hyperparameter w_{syn}^* minimizes the mean LOOCV error over the HF training set as given by

$$w_{\text{syn}}^* = \arg \min_{w_{\text{syn}} \in (0, 1)} \frac{1}{N_{\text{HF}}} \sum_{i=1}^{N_{\text{HF}}} \epsilon_{\text{LOOCV}}(\mathbf{y}_i^{\text{HF}}; w_{\text{syn}}), \quad (17)$$

where $\epsilon_{\text{LOOCV}}(\cdot; w_{\text{syn}})$ is the error function defined in (16). The optimization in Eq. (17) is performed using the BFGS algorithm [47]. As shown in Section 4.3, this procedure is critical for the robust performance of the data augmentation methods with proximity-based weighting, which are sensitive to the choice of w_{syn} .

4 Numerical demonstration: hypersonic aerodynamics application

In this section, we present the results for a hypersonic testbed vehicle problem in the CFD domain described in Section 4.1. The HF and the LF models used for the MF linear regression are described in Section 4.2. Then, we present results for the projection-based MF linear regression methods proposed in this work in Section 4.3. We compare against the state-of-the-art MF regression method using the additive approach presented in [25, 35]. We provide a fair comparison by combining the additive MF method with dimensionality reduction of the outputs (see Appendix A), and compare it against the MF methods presented in Section 3.

4.1 IC3X hypersonic vehicle problem description

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 require an understanding of the surface pressure field as a function of the operating flight conditions, namely, the Mach number, angle of attack, and sideslip angle of the vehicle. However, HF 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 over a range of operating conditions by introducing data from cheaper LF models.

To demonstrate the MF linear regression methods, we consider the Initial Concept 3.X (IC3X) hypersonic vehicle. The IC3X was initially proposed by Pasiliao et al. [48], and a finite element model for the vehicle was developed by Witeof et al. [49]. A primary quantity of interest is the distributed aerodynamic pressure load over the surface of the vehicle at various flight condition parameters. Based on a nominal mission trajectory for this geometry, we consider the range of Mach numbers $M \in [5, 7]$, angles of attack $\alpha \in [0, 8]$, and sideslip angles $\beta \in [0, 8]$. The surface pressure field is computed at a particular flight condition by solving the inviscid Euler equations using the flow solver package Cart3D [50–52] over an adaptive multilevel Cartesian mesh. The mesh adaptation scheme provides a natural hierarchy of model fidelity through various levels of mesh refinement. The discretized surface mesh remains constant, and contains $m = 55966$ nodes, which is the dimension of the output surface pressure vector. An example non-dimensional surface pressure field solution computed by Cart3D at flight conditions of $M = 6$, $\alpha = 4$, and $\beta = 0$ is visualized in Figure 1.

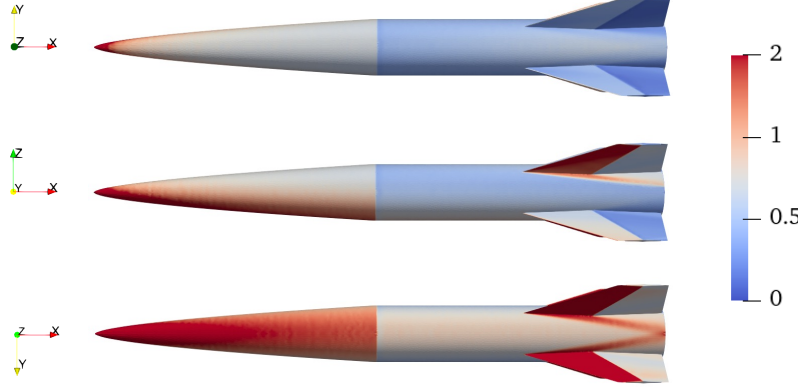


Fig. 1: Top, side, and bottom view of surface pressure (non-dimensional) at flight conditions $M = 6$, $\alpha = 4$, and $\beta = 0$.

4.2 Model specifications and data generation

We can construct different levels of fidelity for the pressure field solution by leveraging Cart3D’s mesh adaptation, which refines the Cartesian volume mesh over multiple adaptation steps. We define two levels of fidelity for simulating the surface 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. Specifically, we control the maximum number of initial mesh refinements (“Max Refinement”), the maximum number of adaptation processes (“Max Adaptations”), error tolerance, and the number of cycles 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. Here, cost refers to the wall-clock time of running the HF and LF simulation on the same hardware. Note that we do not consider the LF simulations to be negligible cost and instead account for the cost of evaluating the LF samples when reporting the computational costs.

Table 1: Model specifications of hypersonic testbed problem

Model Type	Max Initial Refinement	Max Adaptations	Error Tolerance	Cycles/Adaptation	Cost Ratio
HF	7	12	1×10^{-3}	175	1
LF	5	2	5×10^{-3}	50	1/127

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 $N_{\text{HF}} \in [3, 10]$, a LF training sample size of $N_{\text{LF}} = 80$, and a HF testing sample size of $N_{\text{HF}}^{\text{test}} = 50$ to analyze the effectiveness of the proposed methods in the ultra low-data regime. A large sample pool of 100 HF samples are drawn by Latin hypercube sampling (LHS). The testing set is then sampled via conditioned LHS [53] from these points, and was fixed across all repetitions of the dataset. We bootstrap the remaining dataset by using conditioned LHS with different random seeds to create varying combinations of the training dataset and provide a measure of robustness of each method over 50 repetitions of the training samples (which entails the points for training are randomly distributed across the domain). 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.

4.3 Results and discussion

We first analyze the dimensionality reduction on our training datasets of $N_{\text{HF}} = 10$ and $N_{\text{LF}} = 80$ 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. As is evident from the plots, there is not much variability in the singular values across the 50 different dataset draws. We use a tolerance of $\epsilon = 0.995$ for the cumulative energy to decide the size of the low-dimensional subspace using Eq. (3). This leads to $k = 7$ for most LF training datasets. For most HF training datasets with $N_{\text{HF}} > 4$, we get $k = 4$; otherwise, k is bounded by number of HF training samples when $N_{\text{HF}} \leq 4$. This facilitates the use of lower dimensional representations of the data for the surrogate models to be trained on, without significant loss of information.

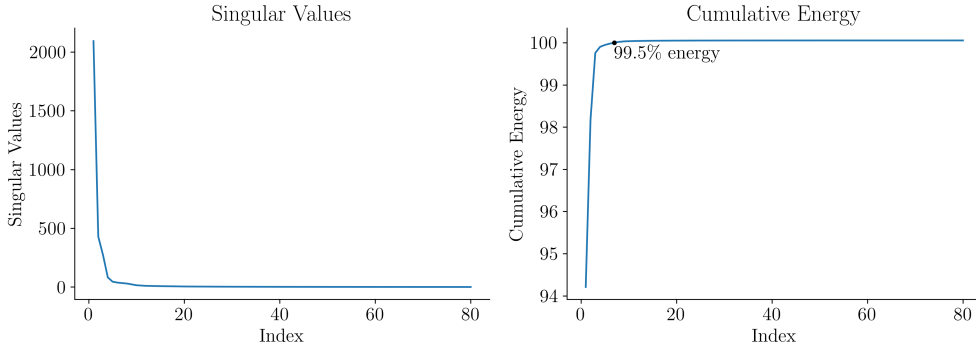


Fig. 2: SVD on 50 repetitions of $N_{\text{LF}} = 80$ LF training data

We apply the three MF linear regression methods described in Section 3 to the prediction of the surface pressure field upon the IC3X testbed hypersonic vehicle. We evaluate the performance of a surrogate model through the normalized L2 accuracy

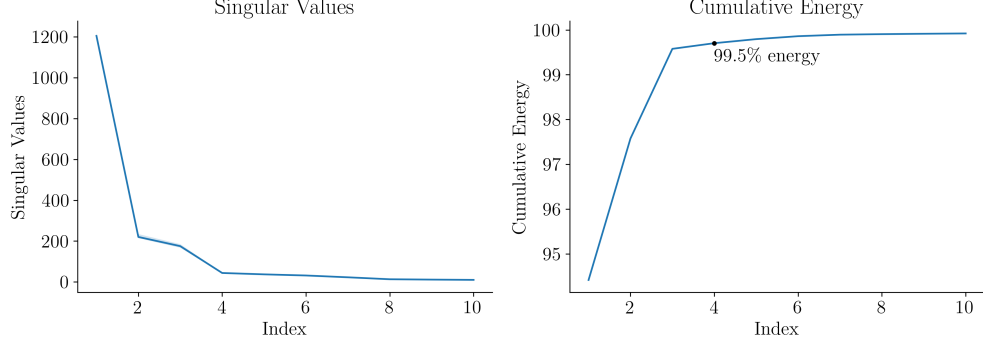


Fig. 3: SVD on 50 repetitions of $N_{\text{HF}} = 10$ HF training data

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

$$\epsilon_{\text{L2}} := \frac{1}{N_{\text{HF}}^{\text{test}}} \sum_{i=1}^{N_{\text{HF}}^{\text{test}}} \frac{\|\mathbf{y}_i^{\text{HF}} - \hat{\mathbf{y}}_i\|_2}{\|\mathbf{y}_i\|_2}, \quad (18)$$

where $\|\cdot\|_2$ is the L2 vector norm, \mathbf{y}_i^{HF} is the HF model solution at i^{th} test sample, and $\hat{\mathbf{y}}_i$ is the surrogate prediction at i^{th} test sample. The R^2 score is another generalizable and useful metric when comparing errors between high-dimensional quantities for regression methods [6]. We present the results using the R^2 metric in Appendix B. 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 varying repetitions of the training 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.

Next, we analyze the impact of different sample weighting schemes on the results of the two data augmentation methods in Figure 4. Setting the weights associated with the HF training samples to 1, we compare the fixed weighting scheme, where $w_{\text{syn}} \in \{0.01, 0.1, 0.9\}$ and $h(w_{\text{syn}}) = w_{\text{syn}}$, against the LOOCV with proximity-based weighting method described in Eq (9), where $h(w_{\text{syn}}) = \sigma(\cdot; w_{\text{syn}}^*)$. Recall that w_{syn}^* is the optimal weighting function hyperparameter value for proximity-based weighting obtained through the LOOCV procedure described in Section 3.3. For this application we use the Heaviside step function to implement $\sigma(\cdot; w_{\text{syn}}^*)$, with a threshold τ set to eliminate the furthest 10th percentile of LF samples (by minimum Euclidean distance to HF samples). We observe that the direct data augmentation method is sensitive to the choice of w_{syn} for the fixed weighting scheme, with a variation of up to $\sim 10\%$ 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. We find that the LOOCV method for determining w_{syn}^* for each repetition of the training dataset performs close to the best fixed weighting scheme option for both data augmentation methods. This highlights the effectiveness of automatic weight selection based on the underlying training dataset. Figure 5 shows the distribution of optimized LF sample weights across 50 repetitions of HF and LF training datasets, following the LOOCV-based optimization procedure. The plotted quantity corresponds to the value of w_{syn}^* obtained through LOOCV for proximity-based weighting function described in Eq. (9) (here, w_{syn}^* is the maximum weight possible when using the Heaviside function). The distribution of w_{syn}^* across the 50 training repetitions is generally bimodal in our setting with $N_{\text{HF}} \leq 10$. This bimodality arises in part because the LOOCV optimization is initialized at 10^{-1} , which explains the presence of a higher mode near this magnitude. As the HF sample size increases, the resulting weight distribution shifts toward smaller magnitudes, suggesting that the added HF data reduces the reliance on LF information for accurate prediction for this application.

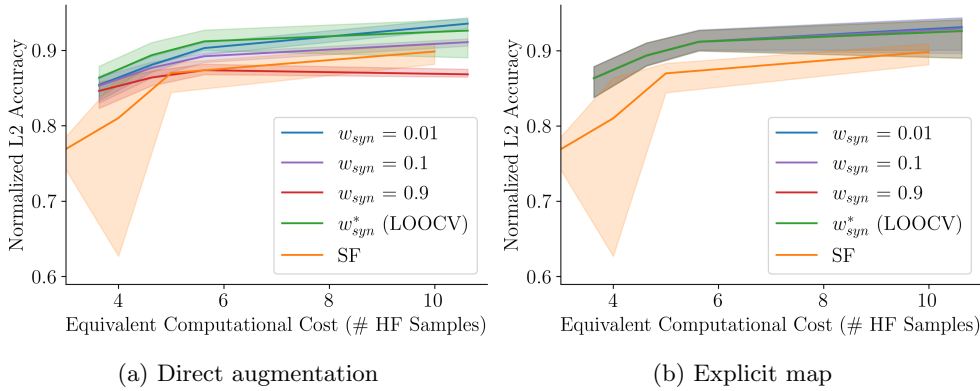


Fig. 4: Comparison of weighting schemes for MF linear regression using data augmentation on 50 repetitions of the training dataset. Here, w_{syn}^* refers to the optimal hyperparameter obtained after LOOCV for proximity-based weighting as described in Eq. (9), and the other weights follow the fixed scheme.

Figure 6 shows the comparison of the two MF linear regression methods proposed in this work and the additive MF method presented in Appendix A, in comparison 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. In contrast, both the data augmentation techniques (using the optimal w_{syn} after LOOCV and proximity-based weights) perform better than the additive approach and show a substantial 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 likely due to the fact that the MF linear regression model sees a larger variety

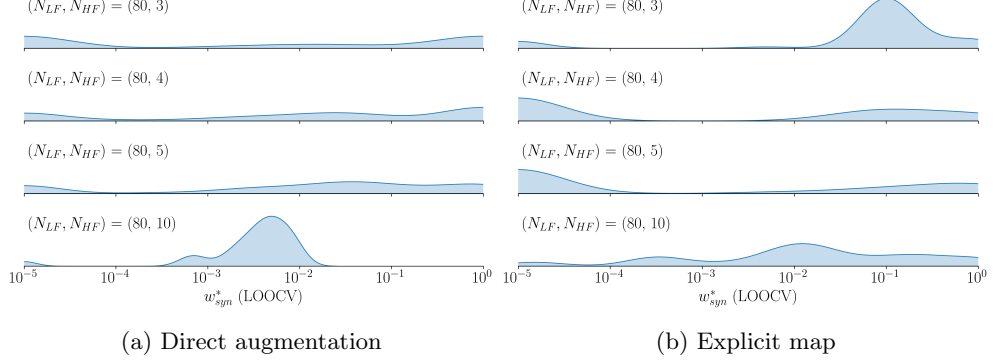


Fig. 5: Comparison of w_{syn}^* distributions obtained from LOOCV for MF linear regression using data augmentation with proximity-based weighting on 50 repetitions of the training dataset. Plotted are the kernel density estimates for each combination of LF and HF data.

of data during the training phase. The extra LF samples in the data augmentation methods are of course not fully representative of the HF model, as indicated by sample weights of 1 for the HF samples and $w_{\text{syn}}^* < 1$ for the synthetic data generated from the LF samples as seen in Figure 5. The MF method with explicit map for data augmentation performs the best with few samples, while the direct augmentation had the highest accuracy with the largest amount of training data. Table 2 provides the median test accuracies and R^2 score of each regression method for $N_{\text{HF}} = 3, 5$, and 10 HF samples. We find that the data augmentation technique using explicit map leads to an improvement of approximately 9.5% compared to the SF model for $N_{\text{HF}} = 3$ HF samples and 3.2% compared to the SF model for $N_{\text{HF}} = 10$ HF samples. Interpolating at the first sample size tested for the MF methods, $N_{\text{HF}} = 3.63$ (3 HF, 80 LF samples), yields a 12.4% improvement in accuracy compared to the SF method. With respect to the R^2 values, the results of the data augmentation methods were shown to have similar improvements upon the SF and additive-based approaches.

Finally, we look at a comparison of the absolute errors in pressure prediction between the SF surrogate and the MF surrogate methods. For an arbitrary 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 7, providing some visual context for the gains the MF surrogate model nets.

5 Numerical demonstration: aircraft disc brakes

In this section, we present the results for an aircraft disc brake thermal modeling problem described in Section 5.1. The HF and the LF models used for the MF linear regression are described in Section 5.2. Then, we present results for the projection-based MF linear regression methods proposed in this work in Section 5.3 and show the comparison to the state-of-the-art additive approach for MF regression.

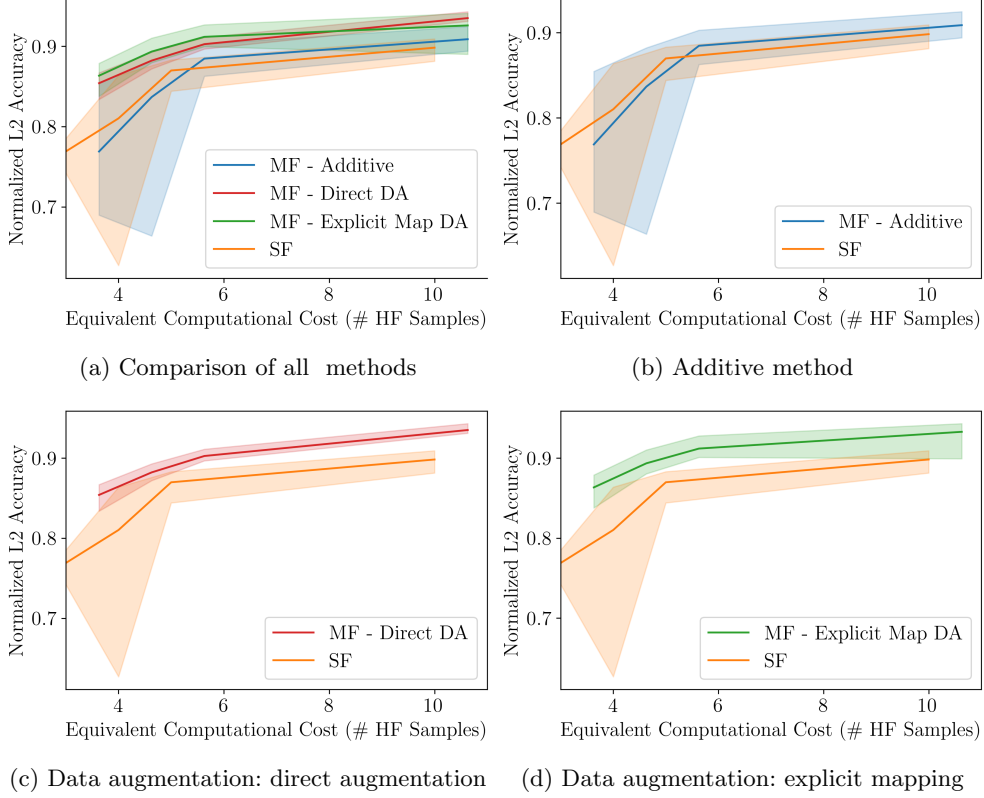


Fig. 6: Comparison of proposed MF linear regression methods to baseline SF linear regression and the additive MF method on 50 repetitions of the training dataset (DA denotes a data augmentation method implemented with LOOCV and the proximity-based weighting)

5.1 Aircraft disc brake thermal problem description

Aircraft wheel braking systems are designed to be capable of ensuring safe deceleration and stopping performance following landing or during rejected takeoff (RTO) scenarios, where air brakes are less impactful. During braking, the kinetic energy of the aircraft is rapidly converted into thermal energy through friction at the brake rotors and stators in a multi-disc brake setup, leading to substantial temperature rises within the brake stack. For aircraft such as the Boeing 737 (B737), each of the four main landing gear wheels is equipped with four rotor-stator pairs, which collectively absorb tens of megajoules of kinetic energy during a typical stop [54]. When braking pressure is applied, multi-disc aircraft brakes dissipate kinetic energy through thermal energy by clamping each of the rotor-stator pairs together through a hydraulic piston. A notional drawing of aircraft brakes in a typical transport-aircraft is depicted in Figure 8. Understanding the temperature evolution and spatial distribution within

Table 2: Selected multifidelity linear regression results (DA denotes a data augmentation method implemented with LOOCV and the proximity-based weighting)

Model Type	# LF Samples	# HF Samples	Median Normalized L2 Test Accuracy	Median Test Data R^2 Score
SF	-	3	0.768	0.38
	-	5	0.870	0.85
	-	10	0.898	0.90
MF - Additive	80	3	0.763	0.52
		5	0.875	0.88
		10	0.909	0.91
MF - Direct DA (LOOCV w_{syn}^*)	80	3	0.854	0.70
		5	0.903	0.86
		10	0.935	0.94
MF - Explicit map DA (LOOCV w_{syn}^*)	80	3	0.863	0.80
		5	0.912	0.93
		10	0.930	0.95

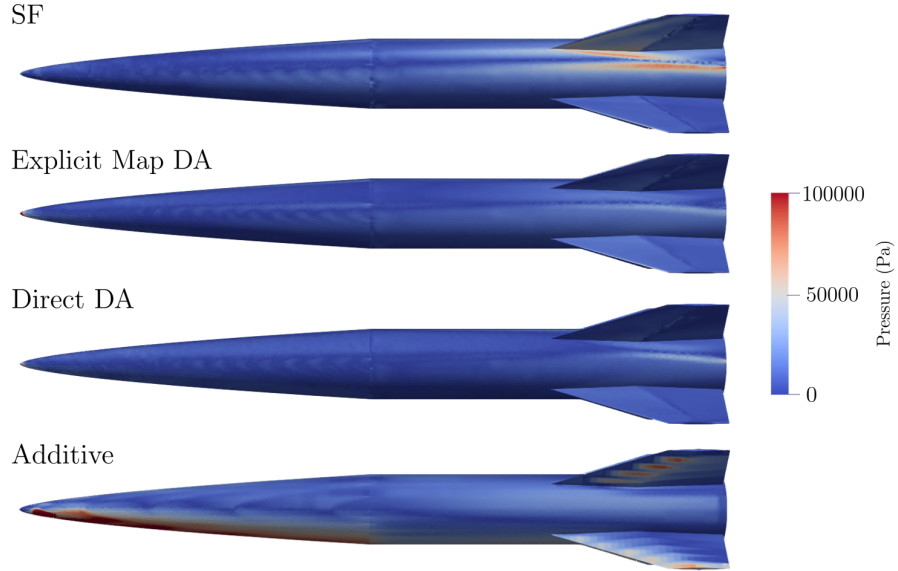


Fig. 7: Comparison of errors in pressure field prediction at **Mach 6.79**, $\alpha = 4.97^\circ$, $\beta = 4.74^\circ$

the brake discs is essential for assessing brake deterioration and the turnaround time required before the next takeoff. Additionally, efficient modeling of the temperature field given braking duration, aircraft mass, approach velocity, and other conditions can be utilized to expedite optimization and design cycles.

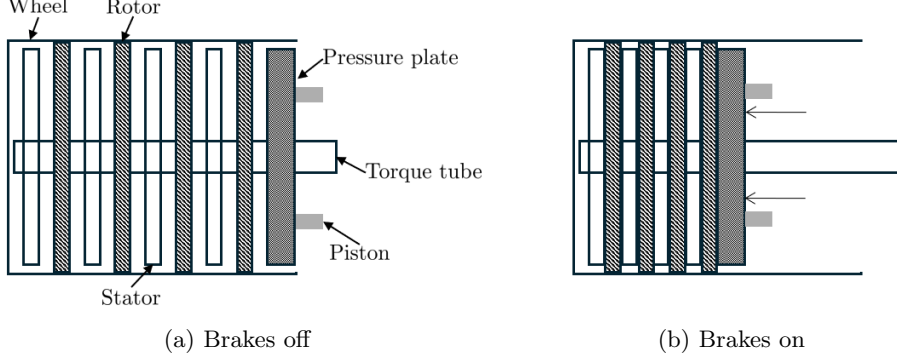


Fig. 8: Notional illustration of multi-disc transport-aircraft brakes, adapted from [55]

We develop an axisymmetric finite element model to simulate the transient temperature field within the brake stack of a representative narrow-body transport aircraft, based on the B737 geometry and braking configuration. The model resolves the temperature over an annular rotor–stator stack during a braking event of duration τ_b , by representing the frictional heating as a heat source term. Furthermore, convective cooling along all the exposed surfaces is modeled. For the MF linear regression, the primary quantity of interest is the temperature field $T(r, z)$ at the end of the braking period where r and z are the radial and axial coordinates in meters. One potential application of such a surrogate model is to rapidly evaluate designs by checking where maximum temperature occurs and if maximum allowable temperature limits are met in the brake assembly given design changes. We consider six model inputs for the surrogate model that include physical and environmental parameters in the form of aircraft mass, approach velocity (or ground velocity in the case of an RTO), braking duration, material thermal diffusivity, ambient temperature, and initial temperature. Table 3 summarizes the parameter definitions and ranges used in this example which were chosen to reflect typical flight conditions of the B737 aircraft.

Table 3: Input parameters and their ranges for the aircraft brake temperature field problem

Input Parameter (Units)	Range
Aircraft mass, m (kg)	[67720, 85140]
Landing velocity, V (m/s)	[30.0, 75.0]
Braking duration, τ_b (s)	[10, 30]
Thermal diffusivity, α (m ² /s)	$[6 \times 10^{-6}, 2 \times 10^{-5}]$
Ambient temperature, T_∞ (K)	[253.15, 319.15]
Initial disc temperature, T_{initial} (K)	[373.15, 473.15]

5.2 Model specifications and data generation

The thermal analysis of the aircraft disc brakes involves solving the 2D transient axisymmetric heat equation represented by the governing equation

$$\begin{aligned} \rho c_p \frac{\partial T}{\partial t} - \nabla \cdot (k \nabla T) &= f(r, z, t), \\ T(r, z, 0) &= T_{\text{initial}}, \end{aligned} \quad (19)$$

where $T(r, z, t)$ K is the temperature field at time t , ρ kg/m³ is the density, c_p J/kg·K is the isobaric specific heat capacity, k W/m·K is the thermal conductivity, and $f(\cdot)$ W/m³ is the volumetric heat source term which represents the frictional heating at the rotor-stator interfaces. The initial condition represents the disc brakes having an initial uniform temperature T_{initial} . The boundary conditions representing the convective cooling through airflow around the exterior surfaces and an adiabatic boundary at the axle or torque tube connection is given by

$$\begin{aligned} -k \nabla T \cdot n &= h(T - T_\infty) \quad \text{on outer radius and end-plates} \\ -k \nabla T \cdot n &= 0 \quad \text{on inner radius,} \end{aligned} \quad (20)$$

where h is the convective heat transfer coefficient of air.

The frictional heating source term is given by the following expression,

$$f(r, z, t) = \begin{cases} \sum_{k=1}^{N-1} \frac{q_k(t)}{\varepsilon} \mathbf{1}_k(r, z), & 0 \leq t \leq \tau_b \\ 0, & t > \tau_b \end{cases} \quad (21)$$

$$\mathbf{1}_k(r, z) = \begin{cases} 1, & |z - z_k| \leq \frac{\varepsilon}{2} \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

where $q_k(t)$ (W/m²) is the heat flux as a function of time at the k -th interface between a rotor-stator pair and ε (m) is chosen to be sufficiently large enough to encapsulate enough elements at the interface locations. Given N rotor-stators, frictional heating will occur at the $N - 1$ interfaces between them. We are assuming that perfect contact occurs between all rotor-stator pairs without resistance and the rotor-stators are of equal size as in [56, 57]; hence, we sum the individual heating contributions equally. In Eq. (22), an indicator function $\mathbf{1}_k(r, z)$ is used to ensure heating only occurs at the small epsilon-sized band centered around each rotor-stator contact interface. We show in Appendix C that this formulation correctly conserves the braking energy the brakes must dissipate.

The governing equation in Eq. (19) and the boundary conditions are used to construct the variational form of the problem, which is passed to the FEniCSx platform [58–61] to solve numerically. We use quadrilateral mesh elements and the backward Euler method for discretization in time. To define two levels of fidelity for this problem, we set up the HF model with the finest discretization in time and space

and the LF model with a far coarser discretization. The exact specifications for the HF and LF model are detailed in Table 4. As a result of the spatial discretization there are $m = 51681$ nodes in the HF simulation. We again consider a data-sparse environment with limited HF samples, specifically $N_{\text{HF}} \in [4, 12]$ and a LF training size of $N_{\text{LF}} = 25$. The same procedure using bootstrapping as described in Section 4.2 is used to generate the results and provide a measure of robustness, wherein a pool of 150 HF samples are drawn and then sampled from. The HF testing set is fixed across all training dataset repetitions and is of size $N_{\text{HF}}^{\text{test}} = 100$.

Table 4: Model specifications of aircraft brake problem

Model Type	n_r (radial elements)	n_z (axial elements)	Δt (s)	Cost Ratio
HF	320	160	2.5×10^{-2}	1
LF	40	20	5×10^{-1}	1/344

5.3 Results and discussion

We begin the analysis by confirming that our tolerance $\epsilon = 0.9999$ yields a feasible number of dimensions in the reduced-space by observing the singular value decay and cumulative energy for both the LF and HF data. We can see from Figures 9–10 that a reduced-dimension of size $k = 4$ would be selected. In the following results and figures, the data augmentation based methods were fitted via a quadratic polynomial while the SF and additive based method had a linear fit. The best performing comparator was plotted for the fairest comparison, because utilizing a higher-order fit for the SF and additive based method resulted in a worse overall accuracy. This highlights the advantage of the WLS approach which allows for higher-order fits before being under-determined. The data in this problem led to both methods being largely invariant with respect to the choice of weighting as seen by Figure 11. Importantly, the LOOCV procedure, which relies entirely on training data, still resulted in close to the best performing surrogate when robustness is taken into account. The distribution of the weighting hyperparameter w_{syn} that were selected using LOOCV is shown in Figure 12. For each repetition of the training dataset, after the cross-validation procedure, weights close to $w_{\text{syn}}^* = 1$ were selected most often. We note that the six-dimensional input space in this problem is greater than the three-dimensional input space in the hypersonics example, which drives the weights for the LF sample towards one for most case due to sparsity of available data. Thus, when fitting a quadratic (degree $p = 2$) polynomial, any additional data would be prioritized by the LOOCV procedure during the weight selection. In this application, applying the explicit mapping initially led to a worse fit compared to direct DA with low HF samples, which again highlights the issues with learning an explicit map for a higher-dimensional input space with limited data. As more HF data was added, the learned map improves and the mapped data becomes more informative as reflected by better accuracy of the MF

method. This can also be seen in the weight distribution plot in Figure 12b as smaller weights were picked more often than in the DA case.

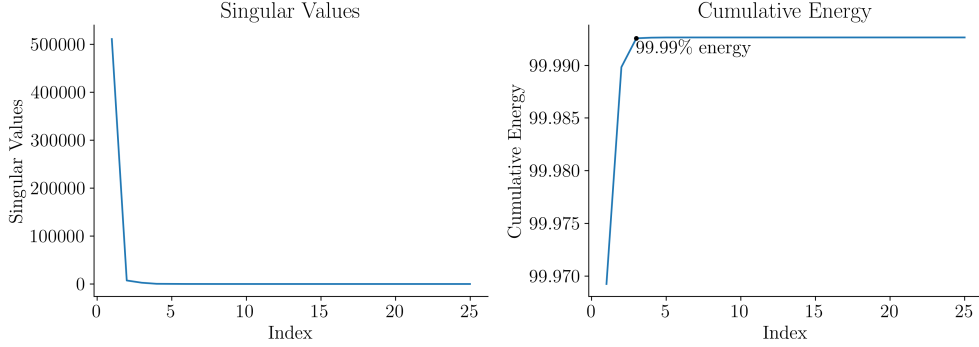


Fig. 9: SVD on 50 repetitions of $N_{LF} = 25$ LF training data

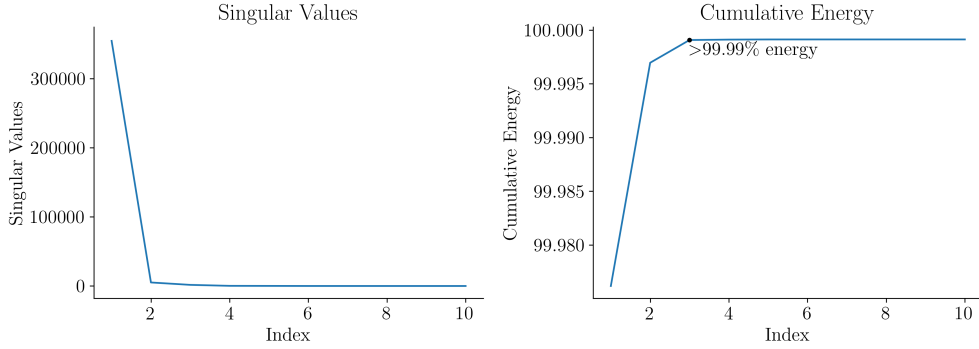


Fig. 10: SVD on 50 repetitions of $N_{HF} = 12$ HF training data

Figure 13 shows the comparison of the two variants of the MF linear regression method proposed in this work with the additive MF method and the standard SF regression. The additive MF method performs similar to the SF linear regression and does not offer significant increase in accuracy for this application. In contrast, both the data augmentation techniques (using the optimal w_{syn} after LOOCV and proximity-based weights) perform better than the additive approach and the SF linear regression for equivalent computational cost. Both the proposed MF techniques demonstrate more robustness to variations in training data with $\sim 2\%$ improvement in the normalized L2 test accuracy. However, the MF linear regression methods show a significant improvement with respect to the R^2 score as seen from Figure B2. The remainder of the results utilizing R^2 values are again found in Appendix B. A selected summary of results for this application can be found in Table 5.

Lastly, we plot the absolute errors in temperature prediction between the SF surrogate and the MF surrogate methods. For an arbitrary test sample, we predict the

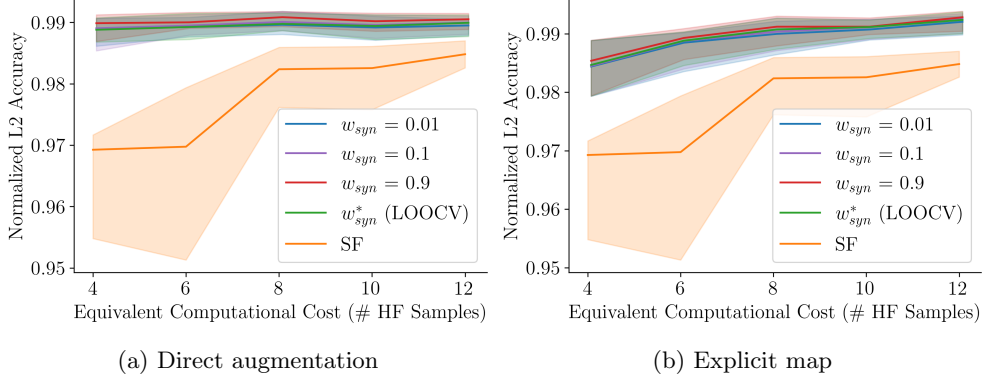


Fig. 11: Comparison of weighting schemes for MF linear regression using data augmentation on 50 repetitions of the training dataset for the aircraft brakes thermal modeling problem. Here, w_{syn}^* refers to the optimal hyperparameter obtained after LOOCV with proximity-based weighting as described in Eq. (9), and the other weights follow the fixed scheme.

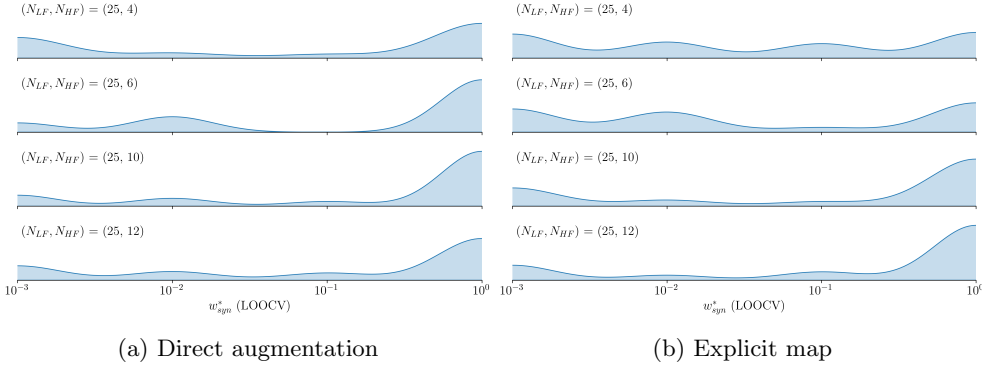


Fig. 12: Comparison of w_{syn}^* distributions obtained from LOOCV for MF linear regression using data augmentation with proximity-based weighting on 50 repetitions of the training dataset for the aircraft brakes thermal modeling problem. Plotted are the kernel density estimates for each combination of LF and HF data.

temperature 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 simplified cylindrical brake assembly in Figure 14, providing some visual context for the gains the MF surrogate model nets.

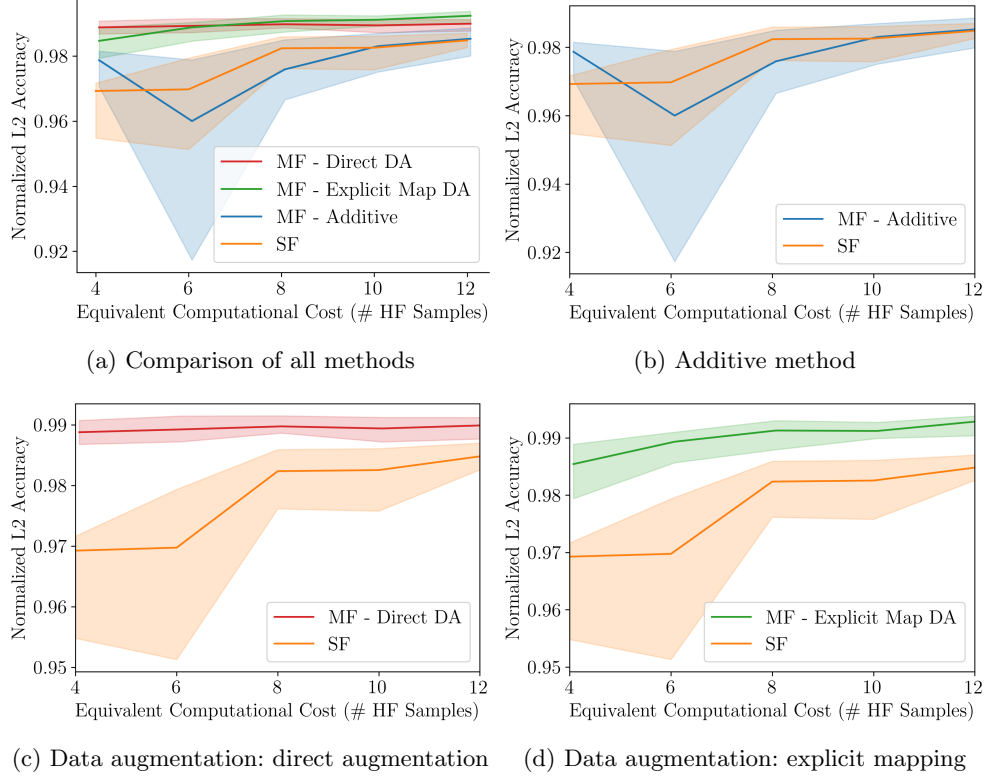


Fig. 13: Comparison of proposed MF linear regression methods to baseline SF linear regression and the additive MF method on 50 repetitions of the training dataset for the aircraft brake thermal modeling problem

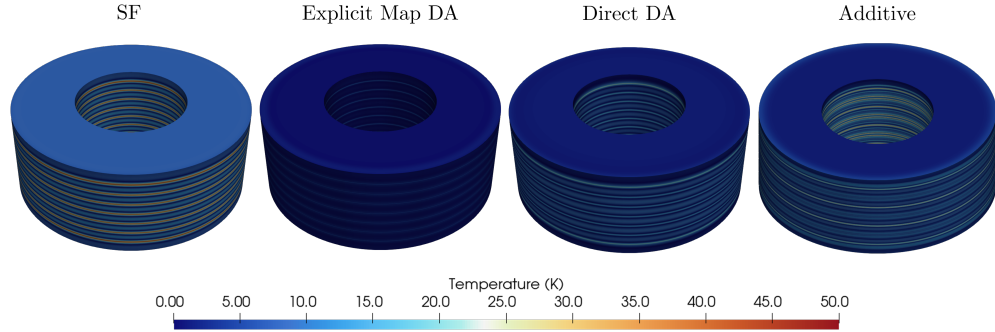


Fig. 14: Comparison of absolute errors in temperature field prediction at **Mass** = 84566 kg, **Velocity** = 70.6 m/s, $\tau_b = 27.1$ s, $\alpha = 7.27 \times 10^{-6}$ m²/s, $T_\infty = 285$ K, $T_{\text{initial}} = 441.6$ K

Table 5: Selected multifidelity linear regression results for the aircraft brake thermal modeling problem

Model Type	# LF Samples	# HF Samples	Median Normalized L2 Test Accuracy	Median Test Data R^2 Score
SF	-	4	0.968	0.75
	-	8	0.982	0.91
	-	12	0.984	0.93
MF - Additive	25	4	0.974	0.88
		8	0.971	0.83
		12	0.978	0.94
MF - Direct DA (LOOCV w_{syn}^*)	25	4	0.992	0.96
		8	0.994	0.98
		12	0.994	0.98
MF - Explicit map DA (LOOCV w_{syn}^*)	25	4	0.983	0.93
		8	0.993	0.97
		12	0.994	0.98

6 Conclusions

This work presents MF linear regression methods with linear and nonlinear features for problems in the ultra low-data regime with two approaches using data augmentation. We embed dimensionality reduction through the principal component analysis with the MF regression methods to tackle high-dimensional outputs. As a point of comparison, we compare with the additive method for MF linear regression that uses the Kennedy O’Hagan framework with discrepancy function to correct the LF regression model. 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. The MF method uses proximity-based weighting strategy with cross-validation to automatically select the optimal weighting parameters. To compare and contrast the MF approaches, two numerical examples are used: (1) the prediction of the pressure load on a hypersonic vehicle in-flight with a three-dimensional input space and (2) the prediction of the temperature field in aircraft disc-brakes after braking with a six-dimensional input space. For these applications with HF training samples in the range of three to twelve, we find that the data augmentation techniques with proximity-based weighting produce robust and accurate surrogate models leading to approximately 2 – 12% in median accuracy gain in the low-data regime as compared to the SF surrogate. The additive approach does not substantially improve the accuracy compared to the baseline SF surrogate model. The direct data augmentation method had comparable accuracy to the explicit mapping method for both examples, but showed more sensitivity to the selection of the synthetic data weight when applied to the hypersonics testbed problem. Both direct data augmentation and explicit mapping methods work robustly and accurately across variations in training data when used with proximity-based weighting and automatic weight selection through cross-validation. The proposed MF linear regression methods demonstrated the capability to robustly predict the high-dimensional quantities of interest in both numerical examples.

Future work can expand these MF regression methods to different underlying regression techniques, such as neural networks and regression trees. Another research direction would be to explore different coordinate transformation techniques for the explicit mapping method.

Declarations

Funding

This work has been supported in part by ARPA-E Differentiate award number DE-AR0001208, AFOSR grant FA9550-21-1-0089 under the NASA University Leadership Initiative (ULI), AFOSR grant FA9550-24-1-0327 under the Multidisciplinary University Research Initiatives (MURI), DARPA Automating Scientific Knowledge Extraction and Modeling (ASKEM) program award number DE-AC05-76RL01830, DARPA The Right Space (TRS) award number HR00112590114, and DOE ASCR grant DE-SC002317.

Appendix A Multifidelity linear regression with an additive structure

We use the state-of-the-art additive MF linear regression method based on the Kennedy–O’Hagan framework [35], adapted to operate on reduced-order representations of the outputs to provide a fair comparison with our proposed approach. The method in Ref. [25] is equipped with dimensionality reduction of the outputs and is used as a point of comparison for the data-augmentation-based MF methods proposed in this work. This method chooses to model the relationship between the LF and the HF data linearly and needs co-located data to estimate discrepancy by HF and LF models. The first component is a LF surrogate model f_{LF} , trained on the reduced LF outputs given by Eq. (2), using OLS on the dataset $(\mathbf{X}_{\text{LF}}, \mathbf{C}_{\text{LF}})$. Similar to the explicit mapping method (see Section 3.1), to obtain co-located LF output predictions, the predictions of reduced LF states at the HF input locations \mathbf{X}_{HF} are obtained by $f_{\text{LF}}(\mathbf{X}_{\text{HF}})$ and reconstructed to the full-dimensional output space as $\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}}$. The discrepancy data between the HF outputs and the co-located LF predictions is then computed as

$$\boldsymbol{\delta}(\mathbf{X}_{\text{HF}}) = \mathbf{Y}_{\text{HF}} - (\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}}). \quad (\text{A1})$$

A second surrogate model f_{δ} is trained via OLS on the reduced discrepancy data $(\mathbf{X}_{\text{HF}}, \mathbf{C}_{\delta})$ obtained through Eq. (A1) and Eq. (2). Then, the predictions from the additive MF regression model in the full-dimensional space at any new input location \mathbf{x}^* is given by

$$\begin{aligned} \hat{y}_{\text{MF}}(\mathbf{x}^*) &= \mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{x}^*) + \bar{\mathbf{Y}}_{\text{LF}} + \mathbf{U}_k^{\delta} f_{\delta}(\mathbf{x}^*) + \bar{\boldsymbol{\delta}} \\ &= \underbrace{\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{x}^*)}_{\text{LF model}} + \underbrace{\mathbf{U}_k^{\delta} f_{\delta}(\mathbf{x}^*)}_{\text{discrepancy model}} + \underbrace{(\bar{\mathbf{Y}}_{\text{LF}} + \bar{\boldsymbol{\delta}})}_{\text{bias}}, \end{aligned} \quad (\text{A2})$$

where \mathbf{U}_k^{δ} is the reduced basis obtained via PCA on the discrepancy data and $\bar{\boldsymbol{\delta}}$ is the sample mean. The procedure for the projection-based additive MF regression model is summarized in Alg. 3.

Appendix B Multifidelity results using the R^2 score

We also present the results showing the coefficient of determination (R^2 score) in Figure B1 for the hypersonics problem and Figure B2 for the aircraft disc brake problem. We see from Figure B1d that the explicit mapping method had the best R^2 for the hypersonics example and was more robust to changes in the training data according to this particular criterion. The explicit mapping and direct augmentation methods performed similarly for the aircraft disc brake application.

Algorithm 3 Multifidelity linear regression via an additive method

Input: HF and LF training data $(\mathbf{X}_{\text{LF}}, \mathbf{Y}_{\text{LF}})$ and $(\mathbf{X}_{\text{HF}}, \mathbf{Y}_{\text{HF}})$, new input locations for prediction \mathbf{x}^*

Output: Output predictions $\hat{\mathbf{y}}_{\text{MF}}(\mathbf{x}^*)$ at input location \mathbf{x}^* from MF surrogate

- 1: Project \mathbf{Y}_{LF} to obtain the reduced states $\mathbf{C}_{\text{LF}} = (\mathbf{U}_k^{\text{LF}})^\top (\mathbf{Y}_{\text{LF}} - \bar{\mathbf{Y}}_{\text{LF}})$ \triangleright see Eq. (2)
 - 2: Train LF linear regression model f_{LF} on $(\mathbf{X}_{\text{LF}}, \mathbf{C}_{\text{LF}})$ using OLS
 - 3: Predict and reconstruct LF outputs at the HF input locations $(\mathbf{U}_k^{\text{LF}} f_{\text{LF}}(\mathbf{X}_{\text{HF}}) + \bar{\mathbf{Y}}_{\text{LF}})$
 - 4: Estimate discrepancy data $\boldsymbol{\delta}(\mathbf{X}_{\text{HF}})$ using Eq. (A1)
 - 5: Use \mathbf{U}_k^δ from the SVD of $\boldsymbol{\delta}$ to project the discrepancy to the reduced state $\mathbf{C}_\delta = (\mathbf{U}_k^\delta)^\top (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}})$
 - 6: Train discrepancy linear regression model f_δ on $(\mathbf{X}_{\text{HF}}, \mathbf{C}_\delta)$ using OLS
 - 7: Predict outputs $\hat{\mathbf{y}}_{\text{MF}}(\mathbf{x}^*)$ at new input location \mathbf{x}^* as the linear combination of f_δ , f_{LF} , and the known bias terms using Eq. (A2)
-

Appendix C Conservation of energy for the heat source term

We show that the source term used to implement the frictional heating at the interfaces of the brake discs conserves the correct amount of energy. When an aircraft is braking, we assume the brakes will only have to dissipate the initial kinetic energy of the aircraft and no other sources. Thus, the total energy of the system at time $t = 0$ is $E_{\text{total}} = \frac{1}{2}mV^2$ where m is the mass of the aircraft (kg) and V is the velocity (m/s). Given that there are four braked wheels on the aircraft considered in this work, we define $W = 4$ to represent the number of brake stacks. Each brake stack contains N discs resulting in the energy needed to be dissipated by one stack as

$$E_{\text{stack}} = (N - 1)\tau_b q_k A_{\text{face}}, \quad (\text{C3})$$

where $N - 1$ is the number of interfaces between the N discs, τ_b (s) is the braking period, A_{face} (m^2) is the surface area of a rotor-stator interface, and q_k (W/m^2) is the heat flux as defined later in Eq. (C4). The total number of interfaces between brake discs undergoing heating is given by $W(N - 1)$. In our frictional heating source term in Eq. (21), we define each q_k such that the total energy of the aircraft is correctly applied within each stack with no variation over time as

$$q_k = \frac{E_{\text{total}}}{W(N - 1)A_{\text{face}}\tau_b}. \quad (\text{C4})$$

This formulation will result in the correct amount of power, and therefore energy, being applied at each interface. The power required to arrest the brakes is given by

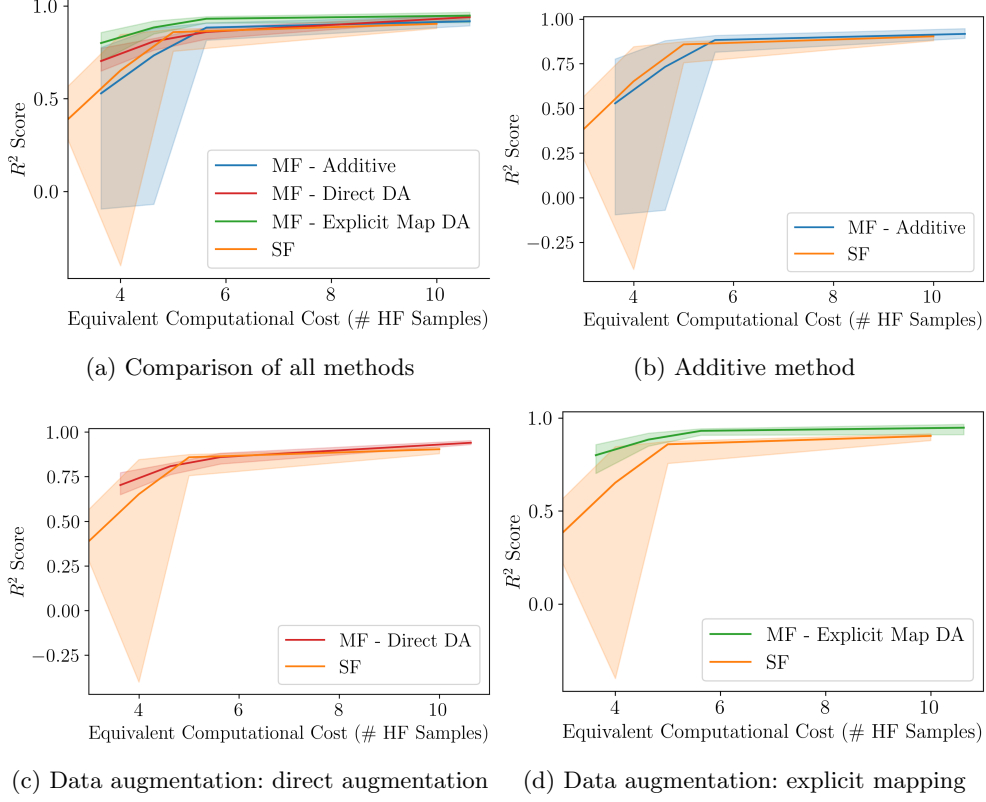


Fig. B1: Comparison of proposed MF linear regression methods to baseline SF linear regression and the additive MF method on 50 repetitions of the training dataset via the R^2 score for the hypersonic testbed problem

the kinetic energy being dissipated as heat through the heat source $f(r, z, t)$ as

$$\begin{aligned}
 P_{\text{interface}} &= \int_{r_i}^{r_o} \int_{z_k - \varepsilon/2}^{z_k + \varepsilon/2} f(r, z, t) 2\pi r \, dz \, dr \\
 &= \int_{r_i}^{r_o} \int_{z_k - \varepsilon/2}^{z_k + \varepsilon/2} \frac{q_k}{\varepsilon} 2\pi r \, dz \, dr \\
 &= \int_{r_i}^{r_o} q_k 2\pi r \, dr \\
 &= q_k A_{\text{face}}.
 \end{aligned}$$

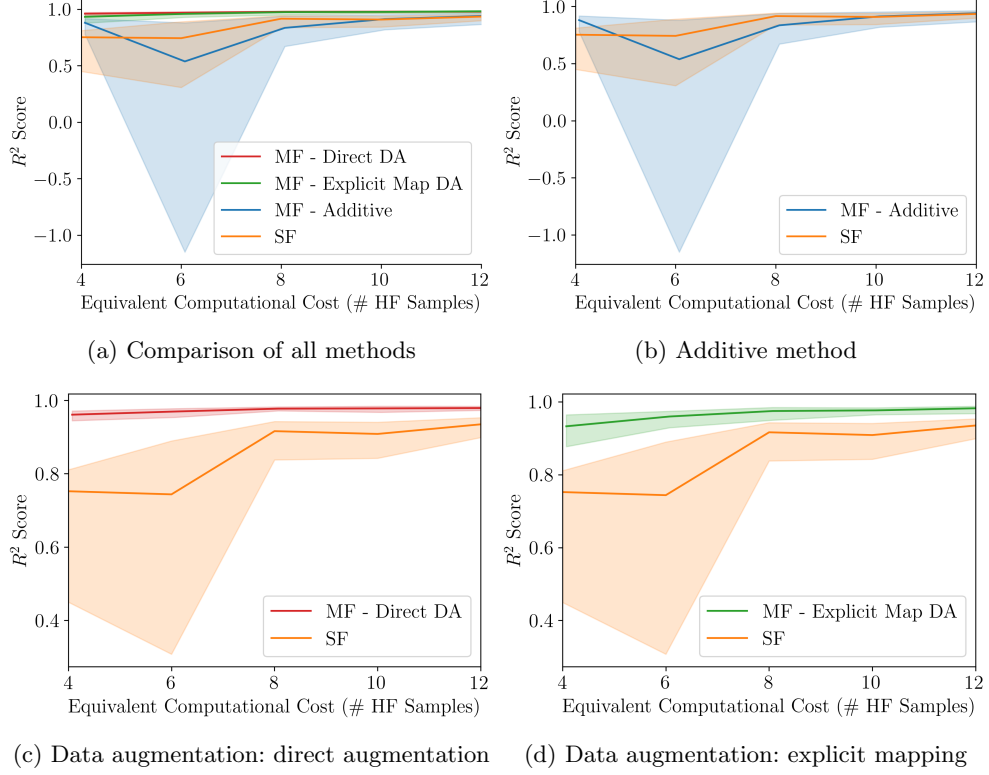


Fig. B2: Comparison of proposed MF linear regression methods to baseline SF linear regression and the additive MF method on 50 repetitions of the training dataset via the R^2 score for the aircraft brake thermal modeling problem

Thus, the energy dissipated over $N - 1$ interfaces in a stack over time τ_b is

$$(N - 1) \int_0^{\tau_b} P_{\text{interface}} dt = (N - 1) \tau_b q_k A_{\text{face}}.$$

which matches the expression for E_{stack} in Eq. (C3).

References

- [1] Baker, N., Alexander, F., Bremer, T., Hagberg, A., Kevrekidis, Y., Najm, H., Parashar, M., Patra, A., Sethian, J., Wild, S., Willcox, K., Lee, S.: Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC), Washington, D.C. (United States) (February 2019). <https://doi.org/10.2172/1478744>

- [2] Zhong, S., Zhang, K., Bagheri, M., Burken, J.G., Gu, A., Li, B., Ma, X., Marrone, B.L., Ren, Z.J., Schrier, J., Shi, W., Tan, H., Wang, T., Wang, X., Wong, B.M., Xiao, X., Yu, X., Zhu, J.-J., Zhang, H.: Machine learning: New ideas and tools in environmental science and engineering. *Environmental Science & Technology* **55**(19), 12741–12754 (2021) <https://doi.org/10.1021/acs.est.1c01339>
- [3] Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A.S., Al-dabbagh, B.S.N., Fadhel, M.A., Manoufali, M., Zhang, J., Al-Timemy, A.H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A., Gu, Y.: A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data* **10**(1), 46 (2023) <https://doi.org/10.1186/s40537-023-00727-2>
- [4] Carleo, G., Cirac, I., Cranmer, K., Daudet, L., Schuld, M., Tishby, N., Vogt-Maranto, L., Zdeborová, L.: Machine learning and the physical sciences. *Reviews of Modern Physics* **91**(4), 045002 (2019)
- [5] Peherstorfer, B., Willcox, K., Gunzburger, M.: Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review* **60**(3), 550–591 (2018) <https://doi.org/10.1137/16M1082469>
- [6] Hastie, T., Tibshirani, R., Friedman, J., Franklin, J.: The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer* **27**(2), 83–85 (2005)
- [7] Madsen, J.I., Shyy, W., Haftka, R.T.: Response surface techniques for diffuser shape optimization. *AIAA Journal* **38**(9), 1512–1518 (2000) <https://doi.org/10.2514/2.1160>
- [8] Nakamura, T., Fukami, K., Fukagata, K.: Identifying key differences between linear stochastic estimation and neural networks for fluid flow regressions. *Scientific Reports* **12**(1) (2022) <https://doi.org/10.1038/s41598-022-07515-7>
- [9] Hosder, S., Watson, L.T., Grossman, B., Mason, W.H., Kim, H., Haftka, R.T., Cox, S.E.: Polynomial response surface approximations for the multidisciplinary design optimization of a high speed civil transport. *Optimization and Engineering* **2**(4), 431–452 (2001) <https://doi.org/10.1023/a:1016094522761>
- [10] Mack, Y., Goel, T., Shyy, W., Haftka, R.: Surrogate model-based optimization framework: A case study in aerospace design, 323–342 (2007) https://doi.org/10.1007/978-3-540-49774-5_14
- [11] Ladický, L., Jeong, S., Solenthaler, B., Pollefeys, M., Gross, M.: Data-driven fluid simulations using regression forests. *ACM Trans. Graph.* **34**(6) (2015) <https://doi.org/10.1145/2816795.2818129>

- [12] Ibarra-Berastegi, G., Elias, A., Arias, R., Barona, A.: Artificial neural networks vs linear regression in a fluid mechanics and chemical modelling problem: Elimination of hydrogen sulphide in a lab-scale biofilter. In: 2007 IEEE/ACS International Conference on Computer Systems and Applications, pp. 584–587 (2007). <https://doi.org/10.1109/AICCSA.2007.370941>
- [13] Balabanov, V., Grossman, B., Watson, L., Mason, W., Haftka, R.: Multi-fidelity response surface model for hsct wing bending material weight. In: 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, p. 4804 (1998)
- [14] Haftka, R.T.: Combining global and local approximations. *AIAA Journal* **29**(9), 1523–1525 (1991) <https://doi.org/10.2514/3.10768>
- [15] Hutchison, M.G., Unger, E.R., Mason, W.H., Grossman, B., Haftka, R.T.: Variable-complexity aerodynamic optimization of a high-speed civil transport wing. *Journal of Aircraft* **31**(1), 110–116 (1994) <https://doi.org/10.2514/3.46462>
- [16] Alexandrov, N.M., Lewis, R.M., Gumbert, C.R., Green, L.L., Newman, P.A.: Approximation and model management in aerodynamic optimization with variable-fidelity models. *Journal of Aircraft* **38**(6), 1093–1101 (2001) <https://doi.org/10.2514/2.2877>
- [17] Madsen, J.I., Langthjem, M.: Multifidelity response surface approximations for the optimum design of diffuser flows. *Optimization and Engineering* **2**(4), 453–468 (2001) <https://doi.org/10.1023/a:1016046606831>
- [18] Vitali, R., Haftka, R.T., Sankar, B.V.: Multi-fidelity design of stiffened composite panel with a crack. *Structural and Multidisciplinary Optimization* **23**(5), 347–356 (2002)
- [19] Eldred, M., Giunta, A., Collis, S.: Second-order corrections for surrogate-based optimization with model hierarchies. In: 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, p. 4457 (2004)
- [20] Sun, G., Li, G., Zhou, S., Xu, W., Yang, X., Li, Q.: Multi-fidelity optimization for sheet metal forming process. *Structural and Multidisciplinary Optimization* **44**(1), 111–124 (2011)
- [21] Zhou, Q., Shao, X., Jiang, P., Zhou, H., Shu, L.: An adaptive global variable fidelity metamodeling strategy using a support vector regression based scaling function. *Simulation Modelling Practice and Theory* **59**, 18–35 (2015)
- [22] Berci, M., Toropov, V., Hewson, R., Gaskell, P.: Multidisciplinary multifidelity optimisation of a flexible wing aerofoil with reference to a small uav. *Structural and Multidisciplinary Optimization* **50**(4), 683–699 (2014)

- [23] Absi, G.N., Mahadevan, S.: Simulation and sensor optimization for multifidelity dynamics model calibration. *AIAA Journal* **58**(2), 879–888 (2020)
- [24] Batra, R., Pilania, G., Uberuaga, B.P., Ramprasad, R.: Multifidelity information fusion with machine learning: A case study of dopant formation energies in hafnia. *ACS Applied Materials & Interfaces* **11**(28), 24906–24918 (2019)
- [25] Zhang, Y., Kim, N.H., Park, C., Haftka, R.T.: Multifidelity surrogate based on single linear regression. *AIAA Journal* **56**(12), 4944–4952 (2018)
- [26] Wang, S., Liu, Y., Zhou, Q., Yuan, Y., Lv, L., Song, X.: A multi-fidelity surrogate model based on moving least squares: fusing different fidelity data for engineering design. *Structural and Multidisciplinary Optimization* **64**(6), 3637–3652 (2021)
- [27] Meng, X., Karniadakis, G.E.: A composite neural network that learns from multifidelity data: Application to function approximation and inverse PDE problems. *Journal of Computational Physics* **401**, 109020 (2020) <https://doi.org/10.1016/j.jcp.2019.109020>
- [28] Zhang, X., Xie, F., Ji, T., Zhu, Z., Zheng, Y.: Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Computer Methods in Applied Mechanics and Engineering* **373**, 113485 (2021) <https://doi.org/10.1016/j.cma.2020.113485>
- [29] Sella, V., O’Leary-Roseberry, T., Du, X., Guo, M., Martins, J.R., Ghattas, O., Willcox, K.E., Chaudhuri, A.: Improving neural network efficiency with multifidelity and dimensionality reduction techniques. In: *AIAA SciTech 2025 Forum*, p. 2807 (2025)
- [30] Guo, M., Manzoni, A., Amendt, M., Conti, P., Hesthaven, J.S.: Multi-fidelity regression using artificial neural networks: Efficient approximation of parameter-dependent output quantities. *Computer Methods in Applied Mechanics and Engineering* **389**, 114378 (2022) <https://doi.org/10.1016/j.cma.2021.114378>
- [31] Conti, P., Guo, M., Manzoni, A., Hesthaven, J.S.: Multi-fidelity surrogate modeling using long short-term memory networks. *Computer Methods in Applied Mechanics and Engineering* **404**, 115811 (2023)
- [32] Forrester, A.I., Sóbester, A., Keane, A.J.: Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **463**(2088), 3251–3269 (2007)
- [33] Le Gratiet, L., Garnier, J.: Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification* **4**(5) (2014)
- [34] Williams, C., Rasmussen, C.: Gaussian processes for regression. *Advances in*

- [35] Kennedy, M.C., O’Hagan, A.: Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**(1), 1–13 (2000)
- [36] Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Review* **57**(4), 483–531 (2015)
- [37] Swischuk, R., Mainini, L., Peherstorfer, B., Willcox, K.: Projection-based model reduction: Formulations for physics-based machine learning. *Computers & Fluids* **179**, 704–717 (2019) <https://doi.org/10.1016/j.compfluid.2018.07.021>
- [38] Guo, M., Hesthaven, J.S.: Data-driven reduced order modeling for time-dependent problems. *Computer Methods in Applied Mechanics and Engineering* **345**, 75–99 (2019)
- [39] O’Leary-Roseberry, T., Villa, U., Chen, P., Ghattas, O.: Derivative-informed projected neural networks for high-dimensional parametric maps governed by PDEs. *Computer Methods in Applied Mechanics and Engineering* **388**, 114199 (2022) <https://doi.org/10.1016/j.cma.2021.114199>
- [40] Sun, J.: A multivariate principal component regression analysis of nir data. *Journal of Chemometrics* **10**(1), 1–9 (1996)
- [41] Izenman, A.J.: Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* **5**(2), 248–264 (1975)
- [42] Li, K.-C., Aragon, Y., Shedden, K., Thomas Agnan, C.: Dimension reduction for multivariate response data. *Journal of the American Statistical Association* **98**(461), 99–109 (2003)
- [43] Abdi, H., Williams, L.J.: Principal component analysis. *Wiley interdisciplinary reviews: computational statistics* **2**(4), 433–459 (2010)
- [44] Ham, J., Lee, D., Saul, L.: Semisupervised alignment of manifolds. In: *International Workshop on Artificial Intelligence and Statistics*, pp. 120–127 (2005). PMLR
- [45] Wang, C., Mahadevan, S.: A general framework for manifold alignment. In: *AAAI Fall Symposium: Manifold Learning and Its Applications*, pp. 79–86 (2009)
- [46] Montgomery, D.C., Peck, E.A., Vining, G.G.: *Introduction to Linear Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ (2021). <https://books.google.com/books?id=tCIgEAAAQBAJ>
- [47] Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound

- constrained optimization. *SIAM Journal on Scientific Computing* **16**(5), 1190–1208 (1995) <https://doi.org/10.1137/0916069>
- [48] Pasiliao, C.L., Sytsma, M.J., Neergaard, L., Witeof, Z., Trolier, J.W.: Preliminary aero-thermal structural simulation. In: 14th AIAA Aviation Technology, Integration, and Operations Conference, p. 2292. American Institute of Aeronautics and Astronautics, Atlanta, GA (2014). <https://doi.org/10.2514/6.2014-2292>
 - [49] Witeof, Z., Neergaard, L.: Initial concept 3.0 finite element model definition. In: Eglin Air Force Base, Air Force Research Laboratory. AFRL-RWWV-TN-2014-0013, FL (2014)
 - [50] NASA Ames Research Center: Automated Triangle Geometry Processing for Surface Modeling and Cartesian Grid Generation (Cart3D) (Accessed 2022). <https://software.nasa.gov/software/ARC-14275-1>
 - [51] Aftosmis, M., Berger, M., Adomavicius, G.: A parallel multilevel method for adaptively refined cartesian grids with embedded boundaries. 38th AIAA Aerospace Sciences Meeting and Exhibit, 2000–0808 (2000) <https://doi.org/10.2514/6.2000-808>
 - [52] Aftosmis, M.J., Berger, M.J., Melton, J.E.: Robust and efficient cartesian mesh generation for component-based geometry. *AIAA Journal* **36**(6), 952–960 (1998) <https://doi.org/10.2514/2.464>
 - [53] Minasny, B., McBratney, A.B.: A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & Geosciences* **32**(9), 1378–1388 (2006)
 - [54] Brady, C.: The Boeing 737 Technical Guide. Tech Pilot Services, Ltd, U.K. (2015)
 - [55] Daidzic, N.E.: Modeling and computation of the maximum braking energy speed for transport category airplanes. *Journal of Aviation Technology and Engineering* **6**(2), 1 (2017)
 - [56] Ramadan, E., Hussien, A., Youssef, A., Abd El-Badia, T.: Numerical simulation of temperature field in a c/c composite multidisk brake during aircraft braking. In: International Conference on Aerospace Sciences and Aviation Technology, vol. 19, pp. 1–14 (2021). The Military Technical College
 - [57] Meunier, C., Bauzin, J.-G., Laraqi, N., Gapin, A., Diebold, J.-F.: Thermal characterization of the braking and cooling stages of an aircraft brake using identification techniques and a life-size experimental test bench. *International Journal of Heat and Mass Transfer* **196**, 123277 (2022)
 - [58] Baratta, I.A., Dean, J.P., Dokken, J.S., Habera, M., Hale, J.S., Richardson, C.N.,

- Rognes, M.E., Scroggs, M.W., Sime, N., Wells, G.N.: DOLFINx: the next generation FEniCS problem solving environment. preprint (2023). <https://doi.org/10.5281/zenodo.10447666>
- [59] Scroggs, M.W., Dokken, J.S., Richardson, C.N., Wells, G.N.: Construction of arbitrary order finite element degree-of-freedom maps on polygonal and polyhedral cell meshes. *ACM Transactions on Mathematical Software* **48**(2), 18–11823 (2022) <https://doi.org/10.1145/3524456>
- [60] Scroggs, M.W., Baratta, I.A., Richardson, C.N., Wells, G.N.: Basix: a runtime finite element basis evaluation library. *Journal of Open Source Software* **7**(73), 3982 (2022) <https://doi.org/10.21105/joss.03982>
- [61] Alnaes, M.S., Logg, A., Ølgaard, K.B., Rognes, M.E., Wells, G.N.: Unified form language: A domain-specific language for weak formulations of partial differential equations. *ACM Transactions on Mathematical Software* **40** (2014) <https://doi.org/10.1145/2566630>