ETH zürich

A COMPREHENSIVE FRAMEWORK FOR MULTI-FIDELITY SURROGATE MODELING WITH NOISY DATA: A GRAY-BOX PERSPECTIVE

K. Giannoukou, S. Marelli and B. Sudret

CHAIR OF RISK, SAFETY AND UNCERTAINTY QUANTIFICATION
STEFANO-FRANSCINI-PLATZ 5
CH-8093 ZÜRICH

Risk, Safety & Uncertainty Quantification

Data Sheet

Journal:	-
Report Ref.:	RSUQ-2024-001A
Arxiv Ref.:	https://arxiv.org/abs/2401.06447 [stat.ME] [stat.ML] [stat.CO]
DOI:	-
Date submitted:	January 12, 2024
Date accepted:	-

A comprehensive framework for multi-fidelity surrogate modeling with noisy data: a gray-box perspective

Katerina Giannoukou [*1], Stefano Marelli[†1], and Bruno Sudret[‡1]

[1] *Chair of Risk, Safety and Uncertainty Quantification, ETH Zürich, Switzerland*

January 12, 2024

Abstract

Computer simulations (a.k.a. white-box models) are more indispensable than ever to model intricate engineering systems. However, computational models alone often fail to fully capture the complexities of reality. When physical experiments are accessible though, it is of interest to enhance the incomplete information offered by computational models. Gray-box modeling is concerned with the problem of merging information from data-driven (a.k.a. black-box) models and white-box (i.e., physics-based) models. In this paper, we propose to perform this task by using multi-fidelity surrogate models (MFSMs). A MFSM integrates information from models with varying computational fidelities into a new surrogate model. The multi-fidelity surrogate modeling framework we propose handles noise-contaminated data and is able to estimate the underlying noise-free high-fidelity function. Our methodology emphasizes on delivering precise estimates of the uncertainty in its predictions in the form of confidence and prediction intervals, by quantitatively incorporating the different types of uncertainty that affect the problem, arising from measurement noise and from lack of knowledge due to the limited experimental design budget on both the high- and low-fidelity models. Applied to gray-box modeling, our MFSM framework treats noisy experimental data as the high-fidelity and the white-box computational models as their low-fidelity counterparts. The effectiveness of our methodology is showcased through synthetic examples and a wind turbine application.

1 Introduction

Predicting the behavior of complex systems and quantifying the corresponding uncertainty is a ubiquitous challenge in engineering and applied sciences. Two essential tools that engineers

[*]katerina.giannoukou@ibk.baug.ethz.ch

[†]marelli@ibk.baug.ethz.ch

[‡]sudret@ethz.ch

and scientists use to deal with this challenge are experiments and predictive models. Generally, experiments can be classified into two broad categories: physical and computer experiments. The latter, also known as computer simulations, are mathematical or computational models of a system which use pre-existing knowledge on the underlying physics of the system to predict its behaviour. Predictive models of this type can be referred to as *white-box models* (Rogers et al., 2017).

Although white-box models are interpretable, they sometimes fail to capture the entirety of the system they represent due to simplifying assumptions and approximations often needed to make them computationally tractable. A second class of predictive models are the so-called *black-box* models. These consist in data-driven models that act as global approximators of the response of a system, based on an available set of input-output observations. While black-box models can provide flexibility, they may not honor the underlying physics as their white-box counterparts do (Rogers et al., 2017).

Often, the response of a system can be predicted by one or more white-box computational models, while additional data can be obtained through physical experiments. In this context, it is natural to look for an approach that combines the white- and black-box modeling paradigms to benefit from the strengths of both. This combined approach is known as gray-box modeling (Tulleken, 1993). Traditionally, physical experiments have been used to improve computational models through model calibration, where parameters of a model are inferred by fitting the model predictions to available experimental data (Kennedy and O'Hagan, 2001; Higdon et al., 2004). Gray-box models take a more comprehensive approach by acknowledging that the computational model may not fully capture the system's complexity and thus, incorporating both knowledge-driven and data-driven elements. As an example, hybrid simulation, which combines physical and numerical substructures to create a hybrid model (Schellenberg et al., 2009; Abbiati et al., 2021), follows a gray-box modeling paradigm. Recently, machine learning approaches using physics-informed neural networks (Raissi et al., 2019) have been employed to perform gray-box modeling, e.g. in the work of Yan et al. (2022). Finally, gray-box models have been historically employed in the field of control theory and system identification, where experimental data are used to estimate the parameters and structure of a mathematical model that represents the underlying dynamical system (Ljung, 1998).

In many applications which require a large number of model evaluations, such as uncertainty quantification and optimization, an obstacle frequently encountered is the high computational cost associated with white-box simulations. For this reason, computational models are often replaced by *surrogate models* (SMs), also known as *metamodels*, or emulators. A SM acts as an inexpensive-to-evaluate approximator of an original model, and is constructed using a limited set of model evaluations, called the *experimental design* (ED), also known as *training set* in machine learning. Among the most widely used SMs for deterministic simulators are polynomial chaos expansions (PCE) (Xiu and Karniadakis, 2002; Blatman and Sudret, 2011), Gaussian processes (GPs) (Rasmussen and Williams, 2006), and support vector regression (Drucker et al., 1996).

A specific class of surrogate models that is particularly useful in scenarios where data or computational models of varying fidelities are accessible is *multi-fidelity surrogate models* (MFSMs). With model fidelity, we refer to the extent to which a model faithfully reflects the characteristics and behavior of the target system it intends to simulate. Generally, high-fidelity (HF) models produce accurate predictions, but are associated with high computational or financial costs. Low-fidelity models (LF) are instead less accurate, but also less expensive to run. Models of different fidelities can occur by, e.g., changing the mathematical or numerical model, or changing accuracy of the numerical solver using different levels of discretization (Fernández-Godino, 2023). Multi-fidelity (MF) surrogate modeling approaches combine multiple sources of different fidelity into a single surrogate model, usually augmenting a limited and expensive-to-obtain HF dataset with more extensive and less expensive lower-fidelity ones (Kennedy and O'Hagan, 2000; Le Gratiet and Garnier, 2014).

The choice of surrogate model is an integral part of the design and construction of a MFSM. Numerous MFSM techniques are based on Gaussian process modeling, following the autoregressive fusing scheme proposed by Kennedy and O'Hagan (2000). Such works include Forrester et al. (2007); Kuya et al. (2011), and Le Gratiet and Garnier (2014), among others, with the latter reformulating the approach from Kennedy and O'Hagan (2000) to have a recursive form, allowing for a reduced computational complexity. Polynomial chaos expansion is another SM that has gained popularity in the past two decades for the purpose of MF surrogate modeling (Ng and Eldred, 2012; Palar et al., 2016). The approaches mentioned so far use linear information fusion, which entail assuming that a higher-fidelity response can be expressed as a linear combination of a lower-fidelity model and a discrepancy function. Recently, multi-fidelity modeling approaches have been proposed in the machine learning community, for example the deep GP-based framework of Cutajar et al. (2018), the Bayesian neural network approaches from Meng et al. (2021); Kerleguer et al. (2024), and the generative adversarial network-based methodology proposed by Zhang et al. (2022). These approaches can capture the nonlinear relations between the different levels of fidelity. According to a comparison among different linear and nonlinear GPs-based MF techniques performed by Brevault et al. (2020), when the high- and low-fidelity models are weakly correlated, nonlinear techniques can outperform linear and less complex techniques, with the caveat of requiring a larger quantity of HF data.

In this paper, we propose to perform gray-box modeling using multi-fidelity surrogate models. For this purpose, we assume that the white-box computational models can capture the general behavior of the physics of the system, while the experiments can capture its entirety, but only on a very limited set of conditions, due to their associated costs. In particular, we consider the experimental data as noise-contaminated realizations of black-box HF models, whereas the available white-box computational models are considered as their LF counterparts. Because white-box computational models can in principle be expensive to evaluate, we also approximate them with surrogate models, constructed on a dataset of low-fidelity model evaluations. White-box models are approached within a nonintrusive context, eliminating the requirement for prior

knowledge of the underlying equations.

Most of the existing literature on MF surrogate modeling focuses on deterministic and noise-free high- and low-fidelity models. However, all real world measurement devices have a limited precision and resolution, and therefore, data resulting from measurements are generally contaminated by irriducible noise. This noise can be considered as a source of aleatory uncertainty, as it is inherent in the measurement process. Moreover, the available training data for the construction of all the elements of a MFSM is in principle relatively small due to computational budget constraints. Thus, we consider all MFSM predictions as affected by epistemic uncertainty. Recent studies that consider the presence of noise in a MF setting include the work of Raissi et al. (2017), who use GP regression to infer the solutions of differential equations when noisy data of different fidelities are available. Furthermore, Zhang et al. (2018) demonstrate that their linear regression-based MF surrogate modeling technique is robust to noise in the HF data, and is also able to estimate the noise level when enough HF data is available. Finally, Ficini et al. (2021) assess the robustness of a GP regression MFSM on problems affected by noisy objective function evaluations.

To the authors' best knowledge, however, no work has yet introduced a comprehensive approach to multi-fidelity surrogate modeling that considers the presence of both noise and/or epistemic uncertainty in the high- and/or low-fidelity data. This approach should also provide a way to quantify the accuracy of the MF model predictions with respect to a) the underlying noise-free HF model, and b) the noisy HF observations. The uncertainty about the mean prediction of a regression model is typically expressed via *confidence intervals* (CIs). On the other hand, the uncertainty on the prediction of an unseen noise-contaminated observation is shown via *prediction intervals* (PIs) (Kutner et al., 2005). Generally, PIs are wider than the corresponding CIs. Many of the existing MF surrogate modeling methods do not provide uncertainty estimations about the MF model predictions. In the cases when CIs are provided, their construction is linked to the particular methodology employed in constructing the MFSM, e.g., GP-based CIs (Raissi et al., 2017; Perdikaris et al., 2017). This can potentially make the frameworks less flexible when different MFSM architectures or SMs need to be explored. Moreover, no work has considered the distinction between confidence and prediction intervals in a MF setting; typically, the latter are disregarded.

The goal of the present paper is to introduce a novel general framework for multi-fidelity surrogate modeling that is able to deal with noisy data and epistemic uncertainty due to limited training information. We assume that the noisy HF data originate from deterministic models contaminated by unbiased stochastic noise. Therefore our MF framework aims at effectively emulating the underlying deterministic models, or in other words at denoising the black-box component. An essential and original feature of our methodology is its ability to provide estimates of the different kinds of uncertainty in its predictions in the form of both confidence and prediction intervals. Moreover, our framework can be applied in the field of gray-box modeling, where experimental data and computational models can be fused into a gray-box multi-fidelity surrogate model.

This paper is organized as follows: In Section 2, after recalling the relevant theory for MFSMs,

we present our framework for MF gray-box modeling, including the related confidence and prediction intervals, as well as the implementation details for each component of the framework. Subsequently, the proposed method is assessed in Section 3, where it is applied on two synthetic examples of increasing complexity and on a real-world application. Finally, in Section 4 we discuss concluding remarks and present prospects for future research.

2 Methods

In this section, we first formally state the multi-fidelity surrogate modeling problem, while establishing the notation we are adopting throughout the paper. Then, we describe our proposed methodology to construct a gray-box multi-fidelity surrogate model. Subsequently, we propose to express the uncertainty related to the MFSM predictions using confidence and prediction intervals, and we discuss the interpretation of these intervals in the MF setting. Lastly, we outline our proposed implementation for constructing a MF gray-box, as well as for estimating its confidence and prediction intervals.

2.1 Multi-fidelity surrogate modeling

We assume that we have s information sources which produce data of different fidelity levels. Without loss of generality, we initially focus on the case when two levels of fidelity are present. Let us consider a HF data set $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$ of size N_{H}, obtained, e.g., from an expensive experimental campaign or computational model. The input space is $\boldsymbol{X} \in \mathbb{R}^M$, while the output space is $Y \in \mathbb{R}$. We assume that observations from this HF information source are contaminated by additive noise, which can correspond to measurement noise in the case of experimental data, to numerical noise in the case of computer simulations, or in general, to unobserved sources of variability. Then, any observation y_{H} at an input point $\boldsymbol{x}$ can be expressed in the general form:

$$y_{\mathrm{H}}(\boldsymbol{x}, \varepsilon_{\mathrm{H}}) = \psi_{\mathrm{H}}(\boldsymbol{x}) + \varepsilon_{\mathrm{H}}, \tag{1}$$

where $\boldsymbol{x}$ is a realization of $\boldsymbol{X}$, $\psi_{\mathrm{H}}(\boldsymbol{x})$ is an unknown deterministic function, and ε_{H} is considered to be an additive noise term, independent of $\boldsymbol{x}$ and modeled as a random variable, following some prescribed zero-mean distribution:

$$\varepsilon_{\mathrm{H}} \sim f_{\varepsilon_{\mathrm{H}}}(\varepsilon_{\mathrm{H}}), \qquad \mathbb{E}\left[\varepsilon_{\mathrm{H}}\right] = 0. \tag{2}$$

Thus, each noise-contaminated observation in the available HF data set $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$ can be expressed as:

$$y_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}, \varepsilon_{\mathrm{H}}^{(i)}) = \psi_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}) + \varepsilon_{\mathrm{H}}^{(i)}, \quad i = 1, \ldots, N_{\mathrm{H}}, \tag{3}$$

where $\varepsilon_{\mathrm{H}}^{(i)}$ is a realization of the noise ε_H. In the rest of this paper, for the sake of notation conciseness we denote each observation $y_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}, \varepsilon_{\mathrm{H}}^{(i)})$ in the HF data set simply as $y_{\mathrm{H}}^{(i)}$.

Moreover, let us consider that for the same system there is another information source of lower fidelity, which provides us with the data set $(\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{Y}_{\mathrm{L}})$, where

$$y_{\mathrm{L}}(\boldsymbol{x}_{\mathrm{L}}^{(i)}, \varepsilon_{\mathrm{L}}^{(i)}) = \psi_{\mathrm{L}}(\boldsymbol{x}_{\mathrm{L}}^{(i)}) + \varepsilon_{\mathrm{L}}^{(i)}, \quad i = 1, \dots, N_{\mathrm{L}}, \tag{4}$$

with $\varepsilon_{\mathrm{L}} \sim f_{\varepsilon_{\mathrm{L}}}(\varepsilon_{\mathrm{L}})$ and $\mathbb{E}[\varepsilon_{\mathrm{L}}] = 0$. The size N_{L} of this lower-fidelity data set is generally larger than that of the corresponding HF data set.

A MFSM aims to directly estimate the underlying deterministic HF function $\psi_{\mathrm{H}}(\boldsymbol{x})$ with a function $\hat{\psi}_{\mathrm{H}}(\boldsymbol{x})$ by combining all the available variable-fidelity information. To this end, the LF response can be represented by a classical surrogate model $\hat{\psi}_{\mathrm{L}}(\boldsymbol{x}) \approx \psi_{\mathrm{L}}(\boldsymbol{x})$, since the cost of obtaining data from the associated source cannot in general be assumed negligible. Then, assuming that the LF model captures the general trend of the underlying HF function $\psi_{\mathrm{H}}(\boldsymbol{x})$, we can express the MFSM as a linear combination of the LF surrogate and a discrepancy function $\delta(\boldsymbol{x})$:

$$\hat{\psi}_{\mathrm{H}}(\boldsymbol{x}) = \rho(\boldsymbol{x}) \cdot \hat{\psi}_{\mathrm{L}}(\boldsymbol{x}) + \delta(\boldsymbol{x}), \tag{5}$$

where $\rho(\boldsymbol{x})$ is a scaling function. We can simplify Equation (5) by assuming the scaling function to be a constant ρ:

$$\hat{\psi}_{\mathrm{H}}(\boldsymbol{x}) = \rho \cdot \hat{\psi}_{\mathrm{L}}(\boldsymbol{x}) + \delta(\boldsymbol{x}). \tag{6}$$

The class of surrogates for the LF model and the discrepancy function can be chosen among a wide range of possibilities, including, among others, PCE, GP regression, or neural networks. However, the framework described in this section is *independent* of the particular choice of surrogate modeling methods, and hence, we will refrain from specifying the choice of SM for now.

Zhang et al. (2018) and Ficini et al. (2021) demonstrated that a number of multi-fidelity regression techniques are robust to noise, provided the number of available high-fidelity observations is large enough. In other words,

$$\lim_{N_H \to \infty} \hat{\psi}_{\mathrm{H}} = \psi_{\mathrm{H}}. \tag{7}$$

When information from more than two levels of fidelity is available, Equation (6) can be generalized in a recursive way (Kennedy and O'Hagan, 2000):

$$\hat{\psi}_s(\boldsymbol{x}) = \rho_s \cdot \hat{\psi}_{s-1}(\boldsymbol{x}) + \delta_s(\boldsymbol{x}), \tag{8}$$

where the predictor of a model at a particular fidelity can be used to construct the predictor for the immediately higher-fidelity model.

2.2 Gray-box modeling

Gray-box modeling consists in combining elements of both data-driven methods (black boxes), and physics-based computational models (white boxes) (Rogers et al., 2017). Gray-box models leverage existing physical knowledge about a system through their white-box elements, but due to

their black-box components, they are also able to capture complex relationships in experimental data that might not be explicitly captured by white-box models. By combining data-driven methods with physical insights, they can compensate for gaps in the understanding of a system's behavior. Hence, gray-box models can provide a more accurate and comprehensive representation of a system than either black- or white-box models alone.

Let us consider a set of observations $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$ obtained through an experimental campaign, each observation of which can be expressed as in Equation (3). The unknown underlying model ψ_{H} can be considered as an expensive-to-evaluate black box. Moreover, let us assume that the system response can also be predicted by a white-box computational model, such as a system of equations or a complex finite-elements simulator, denoted as $\mathcal{M}_{\mathrm{L}}(\boldsymbol{x})$.

Then, we can assume that experiments can accurately capture the behavior of the system under investigation, but the information they provide is incomplete, due to their scarcity. On the other hand, the available computational model can complement this information by providing a larger amount of data, but at the cost of lower accuracy, due either to inherent model simplifications (e.g. ignoring some physics), or to numerical limitations (e.g. discretization). Hence, it is possible to create a gray-box model using multi-fidelity surrogate modeling, wherein the experimental data is regarded as noise-affected realizations of a high-fidelity black-box model, and the white-box computational model is treated as its low-fidelity equivalent. Following the rationale described in the previous section, the computational model is replaced by a surrogate model $\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})$, constructed with a set of model evaluations $(\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{M}_{\mathrm{L}}(\boldsymbol{\mathcal{X}}_{\mathrm{L}}))$.

Adopting the MF surrogate modeling information fusion scheme introduced previously, a multi-fidelity gray-box model predictor aiming to emulate the underlying noise-free HF function output $\psi_{\mathrm{H}}(\boldsymbol{x})$ can be expressed as:

$$\hat{\psi}_{\mathrm{H}}(\boldsymbol{x}) = \rho \cdot \widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x}) + \delta(\boldsymbol{x}). \tag{9}$$

2.3 Confidence and prediction intervals

Confidence and prediction intervals are powerful means of conveying the uncertainty present in the predictions of a model, thereby significantly elevating the informative value of single-point estimates.

If we denote the unobservable error in the MFSM at a given input $\boldsymbol{x}_0$ with respect to the underlying HF model by $e_m(\boldsymbol{x}_0)$, then this reads:

$$e_{\mathrm{m}}(\boldsymbol{x}_0) = \psi_{\mathrm{H}}(\boldsymbol{x}_0) - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_0). \tag{10}$$

Confidence intervals (CIs) express the uncertainty about this model error or, in other words, about where the underlying HF function lies. More precisely, the $(1-2\alpha)$ confidence interval for the underlying HF function at $\boldsymbol{x}_0$ is an interval $[\psi_{\mathrm{lo},\alpha}(\boldsymbol{x}_0), \psi_{\mathrm{up},\alpha}(\boldsymbol{x}_0)]$, such that:

$$\mathbb{P}\left[\psi_{\mathrm{lo},\alpha}(\boldsymbol{x}_0) < \psi_{\mathrm{H}}(\boldsymbol{x}_0) < \psi_{\mathrm{up},\alpha}(\boldsymbol{x}_0)\right] = 1 - 2\alpha, \tag{11}$$

where α is typically set equal to 0.05 for a 90% CI.

This uncertainty is due to the incomplete information provided by the finite-size HF and LF experimental designs, and can be reduced when more data is available. Therefore, it can be interpreted as epistemic uncertainty. Assuming that our regression model is able to accurately represent the underlying HF function in the presence of unlimited data, we have $\lim_{N_H \to \infty}(\psi_{\mathrm{up},\alpha}(\boldsymbol{x}_0) - \psi_{\mathrm{lo},\alpha}(\boldsymbol{x}_0)) = 0$, meaning that as $N_{\mathrm{H}} \to \infty$, the confidence interval for the regression model predictor collapses to a single value at each point.

Moreover, according to Equation (1), a HF output y_{H} at an input $\boldsymbol{x}_0$ is expressed as the sum of the underlying HF function $\psi_{\mathrm{H}}(\boldsymbol{x}_0)$ and a realization of the noise random variable ε_{H}. Hence, from Equations (1) and (10), the error in the MF model with respect to a noise-contaminated HF observation can be expressed as follows:

$$\begin{aligned} y_{\mathrm{H}}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}) - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_0) &= \psi_{\mathrm{H}}(\boldsymbol{x}_0) + \varepsilon_{\mathrm{H}} - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_0) \\ &= (\psi_{\mathrm{H}}(\boldsymbol{x}_0) - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_0)) + \varepsilon_{\mathrm{H}} \\ &= e_{\mathrm{m}}(\boldsymbol{x}_0) + \varepsilon_{\mathrm{H}}. \end{aligned} \tag{12}$$

Thus, we can notice that this error is the sum of two independent components: the reducible model error $e_{\mathrm{m}}(\boldsymbol{x}_0)$ and the irreducible error ε_{H} due to the noise in the HF observations. This uncertainty regarding the value of an unseen HF observation is quantified by *prediction intervals* (PIs). Similarly to CIs, we can write that the $(1 - 2\alpha)$ prediction interval for a HF observation at $\boldsymbol{x}_0$ is an interval $[y_{\mathrm{lo},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}), y_{\mathrm{up},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}})]$, such that:

$$\mathbb{P}\left[y_{\mathrm{lo},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}) < y_{\mathrm{H}}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}) < y_{\mathrm{up},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}})\right] = 1 - 2\alpha. \tag{13}$$

From Equations (10) and (12), it is evident that the PI encloses the corresponding CI, and consequently, the former is wider than the latter. The difference in their width indicates how much we can improve predictions by increasing the amount of training data.

Figure 1 illustrates the difference between the CIs and PIs in a single-fidelity linear regression problem with noise-contaminated data. The area between the blue dashed lines is the 90% CI, and shows the uncertainty about where the regression line (blue line) should lie; alternative regression lines trained on different realizations of the same process that generated the current data are represented with thin blue lines. Moreover, the area between the gray dashed lines is the 90% PI, and expresses the uncertainty about where an unseen noise- contaminated observation is likely to fall.

2.4 Implementation

2.4.1 Construction of a multi-fidelity gray-box model

Starting from a set of high-fidelity experimental data $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$ and a lower-fidelity computational model $\mathcal{M}_{\mathrm{L}}(\boldsymbol{x})$, we provide here the methodology to combine the two to construct a gray box as a MFSM.

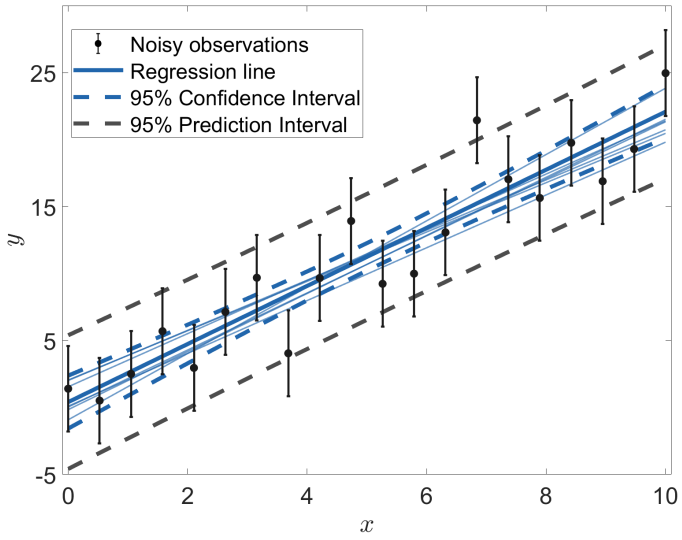


Figure 1: 90% confidence and prediction intervals for the linear regression trained on the illustrated noise-contaminated observations. The thin blue lines represent regression lines for alternative realizations of these observations.

Our approach to MF gray-box modeling uses polynomial chaos expansion as a surrogate model in the hybrid correction scheme introduced in Equation (9), extending the works of Ng and Eldred (2012); Palar et al. (2016); Berchier (2016). The main motivations behind our choice of PCE as a surrogate in our MFSM methodology include its robustness to noise, its efficiency in terms of training, and its applicability in uncertainty quantification problems.

PCE is a surrogate modeling technique which provides an approximation of a model with finite variance through its spectral representation on a polynomial basis (Xiu and Karniadakis, 2002; Ghanem and Spanos, 2003; Lüthen et al., 2021). Let $\boldsymbol{X} \in \mathbb{R}^M$ be a random vector with independent components and joint probability density function (PDF) $f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{M} f_{\boldsymbol{X}_i}(\boldsymbol{x}_i)$, where $f_{\boldsymbol{X}_i}$ is the marginal PDF of the random variable $\boldsymbol{X}_i$. In practice, the polynomial basis needs to be finite, and the truncated PCE of a computational model $\mathcal{M}(\boldsymbol{x})$ is defined as

$$\widetilde{\mathcal{M}}(\boldsymbol{x}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} c_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}), \tag{14}$$

where $c_{\boldsymbol{\alpha}} \in \mathbb{R}$ are the coefficients of the multivariate polynomials $\{\Psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathcal{A}\}$. Each polynomial $\Psi_{\boldsymbol{\alpha}}$ is the product of univariate polynomials orthogonal with respect to the PDF $f_{\boldsymbol{X}_i}$ of the input variable $\boldsymbol{X}_i$, and characterized by the multi-index $\boldsymbol{\alpha}$. $\mathcal{A} \subset \mathbb{N}^M$ is the finite set of multi-indices of the polynomials, and it can be obtained from different truncation schemes, such as total-degree, low-rank or hyperbolic truncation (Marelli et al., 2022).

For the calculation of the PCE coefficients, we adopt a regression-based strategy, as exhaustively reviewed in Lüthen et al. (2021, 2022), because of its applicability to data-driven problems and robustness to noise (Torre et al., 2019). Specifically, we opt for a sparse regression solver, least angle regression (LARS) (Blatman and Sudret, 2011). Moreover, for the choice of the PCE

basis we use degree adaptivity, as well as a total-degree truncation scheme for low-dimensional applications and hyperbolic truncation for higher-dimensional applications.

The first step in constructing our MF gray-box model as in Equation (9) entails creating a surrogate of the low-fidelity model using PCE. For this purpose, we first sample N_L realizations $\boldsymbol{\mathcal{X}}_\mathrm{L} = \{\boldsymbol{x}_\mathrm{L}^{(1)}, ..., \boldsymbol{x}_\mathrm{L}^{(N_\mathrm{L})}\}$ of the input random variables (e.g. through Latin hypercube sampling; McKay et al. (1979)) and obtain the corresponding model responses $\mathcal{Y}_\mathrm{L} = \{y_\mathrm{L}^{(1)}, ..., y_\mathrm{L}^{(N_\mathrm{L})}\}$. We then construct a PCE model

$$\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{A}_\mathrm{L}} c_{\alpha,\mathrm{L}} \Psi_\alpha(\boldsymbol{x}) \tag{15}$$

as discussed above. By using $\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x})$ in place of the original LF model $\mathcal{M}_\mathrm{L}(\boldsymbol{x})$, we eliminate the need for the HF training set to be a subset of the LF one, as from now on we are able to obtain evaluations of $\widehat{\mathcal{M}}_\mathrm{L}$ at a negligible cost. Moreover, we are able to remove any noise that may be present in the LF data in a general MFSM scenario.

Since our HF experimental dataset is given, we can now evaluate $\widehat{\mathcal{M}}_\mathrm{L}$ at the available corresponding input samples $\boldsymbol{\mathcal{X}}_\mathrm{H} = \{\boldsymbol{x}_\mathrm{H}^{(1)}, ..., \boldsymbol{x}_\mathrm{H}^{(N_\mathrm{H})}\}$ to obtain $\{\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}_\mathrm{H}^{(1)}), ..., \widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}_\mathrm{H}^{(N_\mathrm{H})})\}$. An estimator $\hat{\rho}$ of ρ in Equation (9) can be directly obtained as:

$$\hat{\rho} = \mathbb{E}_{\boldsymbol{x}}\left[\frac{y_\mathrm{H}(\boldsymbol{x}, \varepsilon_\mathrm{H})}{\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x})}\right] \approx \frac{1}{N_H}\sum_{i=1}^{N_H} \frac{y_\mathrm{H}^{(i)}}{\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}_\mathrm{H}^{(i)})}. \tag{16}$$

Moreover, the discrepancy term $\delta(\boldsymbol{x})$ in Equation (9) is given by

$$\delta(\boldsymbol{x}) = \hat{\psi}_\mathrm{H}(\boldsymbol{x}) - \rho \cdot \widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}). \tag{17}$$

Now, using as training data $(\boldsymbol{\mathcal{X}}_\delta, \mathcal{Y}_\delta) = (\boldsymbol{\mathcal{X}}_\mathrm{H}, \{y_\mathrm{H}^{(i)} - \hat{\rho}\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}_\mathrm{H}^{(i)}),\ i = 1, ..., N_\mathrm{H}\})$, we train a PCE model $\hat{\delta}(\boldsymbol{x})$ for the discrepancy $\delta(\boldsymbol{x})$,

$$\hat{\delta}(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{A}_\delta} c_{\alpha,\delta} \Psi_\alpha(\boldsymbol{x}). \tag{18}$$

Lastly, the LF and the discrepancy expansions can be merged into a single expansion, which can be expressed as follows:

$$\hat{\psi}_\mathrm{H}(\boldsymbol{x}) = \sum_{\alpha \in \mathcal{A}_\mathrm{L} \cap \mathcal{A}_\delta} (\hat{\rho} c_{\alpha,\mathrm{L}} + c_{\alpha,\delta}) \Psi_\alpha(\boldsymbol{x}) + \sum_{\alpha \in \mathcal{A}_\mathrm{L} \setminus \mathcal{A}_\delta} \hat{\rho} c_{\alpha,\mathrm{L}} \Psi_\alpha(\boldsymbol{x}) + \sum_{\alpha \in \mathcal{A}_\delta \setminus \mathcal{A}_\mathrm{L}} c_{\alpha,\delta} \Psi_\alpha(\boldsymbol{x}), \tag{19}$$

where $\mathcal{A}_\mathrm{L} \cap \mathcal{A}_\delta$ is the set of multi-indices present in both the LF and the discrepancy expansions, $\mathcal{A}_\mathrm{L} \setminus \mathcal{A}_\delta$ is the set of multi-indices present only in the LF expansion, and $\mathcal{A}_\delta \setminus \mathcal{A}_\mathrm{L}$ the set of multi-indices present only in the discrepancy expansion. This last step is optional, and it is only possible when both $\widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x})$ and $\hat{\delta}(\boldsymbol{x})$ are PCEs. In general, different SMs could be used for either of these two models, in which case the combined expression in Equation (19) is not available. Thus, the corresponding step is omitted, and the MF gray-box predictor is instead expressed as:

$$\hat{\psi}_\mathrm{H}(\boldsymbol{x}) = \hat{\rho} \cdot \widehat{\mathcal{M}}_\mathrm{L}(\boldsymbol{x}) + \hat{\delta}(\boldsymbol{x}). \tag{20}$$

In summary, the construction of the multi-fidelity gray-box model from a HF experimental data set $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$ and a LF computational model $\mathcal{M}_{\mathrm{L}}(\boldsymbol{x})$ according to Equation (9) involves the following main steps:

1. Use sampling to obtain an ED $(\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{Y}_{\mathrm{L}}) = (\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{M}_{\mathrm{L}}(\boldsymbol{\mathcal{X}}_{\mathrm{L}}))$ for the LF model;
2. Train a PCE model $\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})$ on $(\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{Y}_{\mathrm{L}})$;
3. Evaluate $\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})$ at the available HF parameter sets $\boldsymbol{\mathcal{X}}_{\mathrm{H}}$ to obtain $\{\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}),\, i = 1, ..., N_{\mathrm{H}}\}$;
4. Estimate $\hat{\rho} = \mathbb{E}\left[\frac{y_{\mathrm{H}}(\boldsymbol{x}, \varepsilon_{\mathrm{H}}))}{\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})}\right] \approx \frac{1}{N_H} \sum_{i=1}^{N_H} \frac{y_{\mathrm{H}}^{(i)}}{\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x}_{\mathrm{H}}^{(i)})}$;
5. Construct a PCE estimator $\hat{\delta}(\boldsymbol{x})$ for the discrepancy function, using the ED $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \{y_{\mathrm{H}}^{(i)} - \hat{\rho}\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}), i = 1, ..., N_{\mathrm{H}}\})$;
6. Use the computed $\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})$, $\hat{\rho}$, $\hat{\delta}(\boldsymbol{x})$ for the MF gray-box predictor as in Equation (20). Optionally, merge the LF and discrepancy expansions into one PCE.

In a broader multi-fidelity setting, a low-fidelity dataset, or even a pre-trained surrogate of the LF model (not necessarily a PCE model) can be available instead of a LF computational model. In this situation, one can follow the same procedure to construct the MFSM, by simply omitting Step 1, and in the second case also Step 2. Furthermore, when a high-fidelity computational model is available instead of a HF dataset, an additional step precedes Step 1. This consists in using sampling to obtain a HF experimental design $(\boldsymbol{\mathcal{X}}_{\mathrm{H}}, \mathcal{Y}_{\mathrm{H}})$. These adaptations accommodate the general multi-fidelity case, which need not strictly adhere to the grey-box modeling paradigm.

Please note that, in the methodology described above, ρ and $\delta(\boldsymbol{x})$ are estimated successively in two separate steps. However, in principle, estimating them jointly is also possible. One approach is to first determine the basis functions for the PCE $\hat{\delta}(\boldsymbol{x})$, defined by the truncation set $\mathcal{A}_\delta$, by following Steps 1-5. Then, considering $\widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x})$ as another basis function in the expression

$$\hat{\psi}_{\mathrm{H}}(\boldsymbol{x}) = \rho \cdot \widehat{\mathcal{M}}_{\mathrm{L}}(\boldsymbol{x}) + \sum_{\alpha \in \mathcal{A}_\delta} c_{\alpha,\delta,\mathrm{new}} \boldsymbol{\Psi}_\alpha(\boldsymbol{x}), \tag{21}$$

one can jointly estimate ρ and the coefficients $c_{\alpha,\delta,\mathrm{new}}$ using OLS.

Another approach is based on alternating least squares (Chevreuil et al., 2015), where the estimates for ρ and $\delta(\boldsymbol{x})$ are iteratively refined. The starting value for $\hat{\rho}$ in this joint optimization can be obtained from the 4th step of the algorithm described above.

In the applications discussed in Section 3, we opt for the method outlined in Steps 1-6 due to its simplicity. Notably, its performance closely paralleled that of the two variations discussed above.

2.4.2 Construction of confidence and prediction intervals

Our methodology for constructing confidence and prediction intervals is based on bootstrapping. The bootstrap estimator is used to determine measures of accuracy for statistical estimates, e.g., standard errors, biases, and confidence intervals, by creating multiple data sets from an original one using random re-sampling with replacement (Efron and Tibshirani, 1994). It is based on the notion that a bootstrap sample is drawn from the observed data in a way similar to how the observed data set is drawn from an unknown population probability distribution. Therefore, inference about a population from an observed data set can be performed by making inference about the latter from the resampled bootstrap data sets.

One of the applications of the bootstrap estimator lies in constructing confidence intervals for regression models (Freedman, 1981). Although the application of bootstrap to provide local error estimates to PCE model predictions within a single-fidelity context has been previously studied (see Marelli and Sudret (2018)), its usage in the context of MFSM has not yet been explored.

Our methodology for constructing CIs about the underlying HF function for a MFSM involves two main steps. The first step for a CI at an arbitrary given input $\boldsymbol{x}_0$ aims at obtaining N_B MF bootstrap model evaluations $\hat{\psi}^*_{\mathrm{H},j}(\boldsymbol{x}_0)$, $j = 1, ..., N_\mathrm{B}$. For this purpose, we need to construct N_B MFSMs from N_B MF bootstrap data sets, which we obtain by independently resampling pairs from the HF and the LF experimental designs. The second step consists in constructing the CI based on the available bootstrap model evaluations. To this end, we can apply one of several bootstrap variations, thoroughly described in Efron and Tibshirani (1994); Davison and Hinkley (1997); Carpenter and Bithell (2000). We choose the percentile method, due to its simplicity, its range-preserving property (i.e. by construction, the produced intervals always remain within the valid bounds of a system's response, as opposed to other bootstrap methods, e.g., the standard normal method), as well as the satisfactory performance it demonstrates in our setting.

The $(1-2\alpha)$-quantile CI is obtained from the α- and $(1-\alpha)$-quantile of the empirical quantile function of $\hat{\psi}^*_\mathrm{H}(\boldsymbol{x}_0)$:

$$[\psi_{\mathrm{lo},\alpha}(\boldsymbol{x}_0), \psi_{\mathrm{up},\alpha}(\boldsymbol{x}_0)] = \left[\hat{\psi}^{*[\alpha]}_\mathrm{H}(\boldsymbol{x}_0), \hat{\psi}^{*[1-\alpha]}_\mathrm{H}(\boldsymbol{x}_0)\right]. \tag{22}$$

More formally, the procedure for constructing a CI at $\boldsymbol{x}_0$ entails the following steps:

1. Obtain N_B MF bootstrap model evaluations $\hat{\psi}^*_{\mathrm{H},j}(\boldsymbol{x}_0)$, $j = 1, ..., N_\mathrm{B}$:

   (a) From the HF ED $(\boldsymbol{\mathcal{X}}_\mathrm{H}, \mathcal{Y}_\mathrm{H})$, create N_B HF bootstrap data sets $(\boldsymbol{\mathcal{X}}^*_{\mathrm{H},j}, \mathcal{Y}^*_{\mathrm{H},j})$. Each such data set contains N_H pairs $(\boldsymbol{x}^{*(b)}_{\mathrm{H},j}, y^{*(b)}_{\mathrm{H},j})$, $b = 1, ..., N_\mathrm{H}$, where $(\boldsymbol{x}^{*(b)}_{\mathrm{H},j}, y^{*(b)}_{\mathrm{H},j})$ is a random sample from $(\boldsymbol{\mathcal{X}}_\mathrm{H}, \mathcal{Y}_\mathrm{H})$, such that

$$P\left[(\boldsymbol{x}^{*(b)}_{\mathrm{H},j}, y^{*(b)}_{\mathrm{H},j}) = (\boldsymbol{x}^{(i)}_\mathrm{H}, y^{(i)}_\mathrm{H})\right] = \frac{1}{N_\mathrm{H}}, \text{ for } i = 1, ..., N_\mathrm{H}\,; \tag{23}$$

(b) Similarly, from the LF ED $(\boldsymbol{\mathcal{X}}_{\mathrm{L}}, \mathcal{Y}_{\mathrm{L}})$, create N_{B} LF bootstrap data sets $(\boldsymbol{\mathcal{X}}^{*}_{\mathrm{L},j}, \mathcal{Y}^{*}_{\mathrm{L},j})$, each one containing N_{L} elements. If $(\boldsymbol{x}^{*(b)}_{\mathrm{L},j}, y^{*(b)}_{\mathrm{L},j})$ is an element of the j-th LF bootstrap data set, then

$$P\left[(\boldsymbol{x}^{*(b)}_{\mathrm{L},j}, y^{*(b)}_{\mathrm{L},j}) = (\boldsymbol{x}^{(i)}_{\mathrm{L}}, y^{(i)}_{\mathrm{L}})\right] = \frac{1}{N_{\mathrm{L}}}, \text{ for } i = 1, ..., N_{\mathrm{L}}\,; \tag{24}$$

(c) Match one-to-one the HF and LF bootstrap data sets to construct N_{B} bootstrap MFSMs $\hat{\psi}^{*}_{\mathrm{H},j}(\boldsymbol{x})$, $j = 1, ..., N_{\mathrm{B}}$;

(d) Evaluate the bootstrap MFSMs $\hat{\psi}^{*}_{\mathrm{H},j}$ at $\boldsymbol{x}_0$ to obtain $\hat{\psi}^{*}_{\mathrm{H},j}(\boldsymbol{x}_0)$, $j = 1, ..., N_{\mathrm{B}}$;

2. Construct the $(1-2\alpha)$-percentile CI based on $\hat{\psi}^{*}_{\mathrm{H}}(\boldsymbol{x}_0)$:
Estimate $[\psi_{\mathrm{lo},\alpha}(\boldsymbol{x}_0), \psi_{\mathrm{up},\alpha}(\boldsymbol{x}_0)]$ as $\left[\hat{\psi}^{*[\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0), \hat{\psi}^{*[1-\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0)\right]$, where $\hat{\psi}^{*[\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0)$ and $\hat{\psi}^{*[1-\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0)$ are the α- and $(1-\alpha)$-empirical quantile of $\hat{\psi}^{*}_{\mathrm{H}}(\boldsymbol{x}_0)$.

Moving now to the construction of prediction intervals about an unseen noise-contaminated observation, we can follow the same procedure used for the confidence intervals about the underlying HF function, with the additional step of accounting for the noise inherent in the observations, as follows from Equation (12). More precisely, accounting for the noise comprises a two-step process. First, we need to infer the distribution of ε_{H} that characterizes the noise present in the HF data (see Equation (1)). We do this by obtaining realizations of this noise and then use classical inference to fit and select among a family of possible parametric univariate distributions.

In practice, we can expect that our MF predictor will exhibit some bias, which we denote as β. Then, we can obtain a realization of ε_{H} by computing the residual for each HF observation:

$$r^{(i)} = y^{(i)}_{\mathrm{H}} - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}^{(i)}_{\mathrm{H}}) - \beta^{(i)}. \tag{25}$$

Moreover, an estimate for the bias $\beta^{(i)}$ is obtained from bootstrap as follows (Efron and Tibshirani, 1994):

$$\hat{\beta}^{(i)} = \mathbb{E}\left[\hat{\psi}^{*}_{\mathrm{H}}(\boldsymbol{x}^{(i)}_{\mathrm{H}})\right] - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}^{(i)}_{\mathrm{H}}), \tag{26}$$

where $\mathbb{E}\left[\hat{\psi}^{*}_{\mathrm{H}}(\boldsymbol{x}^{(i)}_{\mathrm{H}})\right]$ is the bootstrap expectation, which can be approximated by the sample average

$$\mu^{*}(\boldsymbol{x}^{(i)}_{\mathrm{H}}) = \frac{1}{N_{\mathrm{B}}} \sum_{b=1}^{N_{\mathrm{B}}} \hat{\psi}^{*}_{\mathrm{H},b}(\boldsymbol{x}^{(i)}_{\mathrm{H}}). \tag{27}$$

Please note that the bootstrap estimate of bias does not consider biases arising from potential inaccuracies in our regression model, such as those introduced by the truncation of the PCE basis. However, it is capable of detecting other biases resulting from, e.g., the estimation of coefficients through sparse regression techniques.

Substituting $\mu^*(\boldsymbol{x}_{\mathrm{H}}^{(i)})$ for $\mathbb{E}\left[\hat{\psi}_{\mathrm{H}}^*(\boldsymbol{x}_{\mathrm{H}}^{(i)})\right]$ and $\hat{\beta}^{(i)}$ for $\beta^{(i)}$ in Equation (25), the residual computed at $\boldsymbol{x}_{\mathrm{H}}^{(i)}$ can be written as:

$$\begin{aligned} r^{(i)} &= y_{\mathrm{H}}^{(i)} - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}) - \hat{\beta}^{(i)} \\ &= y_{\mathrm{H}}^{(i)} - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)}) - (\mu^*(\boldsymbol{x}_{\mathrm{H}}^{(i)}) - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{H}}^{(i)})) \\ &= y_{\mathrm{H}}^{(i)} - \mu^*(\boldsymbol{x}_{\mathrm{H}}^{(i)}). \end{aligned} \tag{28}$$

This means that a realization of ε_{H} can be estimated as the difference between a HF observation and the bootstrap mean. Having N_{H} noise samples $r^{(i)}$, we use maximum likelihood estimation (MLE) to infer the parameters of a zero-mean distribution. Within the scope of this work, we consider zero-mean variants of the classical Gaussian, Laplace, and Uniform distributions, but in the general case any distribution could be considered. Finally, we use the Bayesian information criterion (BIC; Schwartz (1978)) to select the most appropriate distribution among those considered.

The second step for the PI construction at $\boldsymbol{x}_0$ consists in adding a new noise realization $\hat{\varepsilon}_{\mathrm{H},j}$ from the estimated noise to each of the bootstrap model evaluations $\hat{\psi}_{\mathrm{H},j}^*(\boldsymbol{x}_0)$ obtained in Step 1.d of the CI construction process to obtain a new noisy HF realization:

$$\hat{y}_{\mathrm{H},j}^*(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}}) = \hat{\psi}_{\mathrm{H},j}^*(\boldsymbol{x}_0) + \hat{\varepsilon}_{\mathrm{H},j}, \text{ for } j = 1, ..., N_{\mathrm{B}}. \tag{29}$$

Finally, similarly to the CI, the $(1-2\alpha)$-quantile PI is obtained from the α- and $(1-\alpha)$-quantile of the empirical quantile function of $\hat{y}_{\mathrm{H}}^*(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})$:

$$[y_{\mathrm{lo},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}), y_{\mathrm{up},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}})] = \left[\hat{y}_{\mathrm{H}}^{*[\alpha]}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}}), \hat{y}_{\mathrm{H}}^{*[1-\alpha]}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})\right]. \tag{30}$$

The process for the construction of a PI at $\boldsymbol{x}_0$ is summarized as follows:

1. Obtain N_{B} MF bootstrap model evaluations $\hat{\psi}_{\mathrm{H},j}^*(\boldsymbol{x}_0)$, $j = 1, ..., N_{\mathrm{B}}$:
   Steps a - d are the same as for the CI construction;

2. Estimate the irreducible noise ε_{H} present on the HF data:

   (a) Obtain N_{H} noise realizations from the residuals: $r^{(i)} = y_{\mathrm{H}}^{(i)} - \mu^*(\boldsymbol{x}_{\mathrm{H}}^{(i)})$, $i = 1, ..., N_{\mathrm{H}}$, where $\mu^*(\boldsymbol{x}_{\mathrm{H}}^{(i)})$ is the bootstrap mean at $\boldsymbol{x}_{\mathrm{H}}^{(i)}$;

   (b) Infer the noise distribution:

      i. Use MLE to fit a zero-mean Gaussian, Laplace, and Uniform distribution to the samples $r^{(i)}$;

      ii. Use BIC to choose the most suitable distribution $\hat{\varepsilon}_{\mathrm{H}}$;

3. Add a new realization of $\hat{\varepsilon}_{\mathrm{H}}$ to each of the bootstrap model evaluations $\hat{\psi}_{\mathrm{H},j}^*(\boldsymbol{x}_0)$ to obtain new noisy HF realizations:
   $\hat{y}_{\mathrm{H},j}^*(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}}) = \hat{\psi}_{\mathrm{H},j}^*(\boldsymbol{x}_0) + \hat{\varepsilon}_{\mathrm{H},j}$, $j = 1, ..., N_{\mathrm{B}}$;

4. Construct the $(1-2\alpha)$-percentile PI based on $\hat{y}^*_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})$:
Estimate $[y_{\mathrm{lo},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}}), y_{\mathrm{up},\alpha}(\boldsymbol{x}_0, \varepsilon_{\mathrm{H}})]$ as $\left[\hat{y}^{*[\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}}), \hat{y}^{*[1-\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})\right]$, where $\hat{y}^{*[\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})$ and $\hat{y}^{*[1-\alpha]}_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})$ are the α- and $(1-\alpha)$-quantile of $\hat{y}^*_{\mathrm{H}}(\boldsymbol{x}_0, \hat{\varepsilon}_{\mathrm{H}})$.

3 Validation and results

In this section, the performance of the proposed framework for multi-fidelity gray-box modeling is illustrated on three examples of increasing complexity: an analytical example with one-dimensional input (Section 3.3), a case study with a ten-dimensional finite-element model of a truss and its simply supported beam approximation (Section 3.4), and a real-world application involving wind turbine simulations (Section 3.5). In each application, the HF data contain noise, which is either naturally present (Example 3.5) or artificially introduced by us to replicate the gray-box scenario (Examples 3.3, 3.4). The validation of the proposed framework comprises two parts: assessing the performance of the MFSM constructed as described in Section 2.4.1, and appraising the confidence and prediction intervals constructed as outlined in Section 2.4.2.

For the implementation of the PCE models involved in the validation process, we use UQLab (Marelli and Sudret, 2014), a general-purpose uncertainty quantification software implemented in Matlab.

3.1 MFSM performance and convergence evaluation

We assess the predictive performance and convergence behaviour of our MFSM using the *normalized validation error*, computed on a test set consisting of N_{test} data points that were not used for training, as follows:

$$\epsilon_{\mathrm{val}} = \frac{\sum_{i=1}^{N_{\mathrm{test}}} (y_{\mathrm{t}}^{(i)} - \hat{\psi}_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{t}}^{(i)}))^2}{\sum_{i=1}^{N_{\mathrm{test}}} (y_{\mathrm{t}}^{(i)} - \mu_y)^2}, \tag{31}$$

where $y_{\mathrm{t}}^{(i)}$ equals the noise-free HF model response $\psi_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{t}}^{(i)})$ at the test point $\boldsymbol{x}_{\mathrm{t}}^{(i)}$, when this response is available (Examples 3.3, 3.4), or the noisy HF response $y_{\mathrm{H}}(\boldsymbol{x}_{\mathrm{t}}^{(i)})$ when the noise-free response is not known (Example 3.5), and μ_y is the mean value of the HF response.

The convergence of our MFSM with respect to the HF experimental design size can be investigated by performing simulations with increasing HF ED size, while keeping the LF ED fixed. Due to the statistical uncertainty associated with each HF random design, 50 replications are carried out, considering each time a different independent realization of this experimental design. Box plots are used to provide an aggregated view of the results obtained in all scenarios.

In the subsequent applications, our objective is to assess the added value of the MFSM in comparison to single-fidelity models, and specifically, we aim to determine whether the MFSM exhibits a faster convergence rate. For this purpose, for each HF experimental design and each

replication, alongside the MFSM, we construct a PCE surrogate model trained solely the HF data. The same test data is used to compute ϵ_{val}, for both the MFSM and the HF PCE model.

3.2 Performance measures for confidence and prediction intervals

Regarding the evaluation of confidence and prediction intervals, two well-established key indicators are given by the *confidence interval coverage probability* (CICP) and *prediction interval coverage probability* (PICP) respectively, as well as the *average coverage error* (ACE) (Wan et al., 2014).

If the nominal coverage of a CI is $1-2\alpha$ and the corresponding CI is $[\psi_{\text{lo},\alpha}(\boldsymbol{x}), \psi_{\text{up},\alpha}(\boldsymbol{x})]$, one can estimate the CICP associated with this nominal coverage using $\hat{C}_\alpha$, defined as

$$\hat{C}_\alpha = \frac{1}{N_{\text{t}}} \sum_{i=1}^{N_{\text{t}}} \mathbb{1}(\psi_{\text{H,t}}^{(i)} \in [\psi_{\text{lo},\alpha}(\boldsymbol{x}_{\text{H, t}}^{(i)}), \psi_{\text{up},\alpha}(\boldsymbol{x}_{\text{H,t}}^{(i)})]), \tag{32}$$

where $(\boldsymbol{\mathcal{X}}_{\text{H,t}}, \Psi_{\text{H,t}})$ is a test set of size N_{t} of HF inputs and the corresponding noise-free responses, and $\mathbb{1}(\cdot)$ is the indicator function, which returns 1 if the condition between parentheses is true, and 0 otherwise.

To account for the statistical uncertainty associated with the HF and LF random designs, as well as the bootstrap sampling, we perform $N_{rep} = 10$ replications with varying random seed, and compute the mean CICP (MCICP) over these replications:

$$\bar{C}_\alpha = \frac{1}{N_{\text{rep}}} \sum_{j=1}^{N_{\text{rep}}} \hat{C}_\alpha^{(j)}, \tag{33}$$

When the computed confidence intervals are reliable, the MCICP should be close to its nominal value, i.e., $\bar{C}_\alpha \approx 1-2\alpha$.

The CICP and MCICP can only be estimated when the underlying noise-free HF function is known, thus their computation is generally unfeasible in real-world applications when data contain noise.

Similarly, for a prediction interval with nominal coverage $1-2\alpha$, the mean prediction interval coverage probability (MPICP) can be estimated as follows:

$$\bar{P}_\alpha = \frac{1}{N_{\text{rep}}} \sum_{j=1}^{N_{\text{rep}}} \hat{P}_\alpha^{(j)}, \tag{34}$$

where

$$\hat{P}_\alpha = \frac{1}{N_{\text{t}}} \sum_{i=1}^{N_{\text{t}}} \mathbb{1}(y_{\text{H,t}}^{(i)} \in [y_{\text{lo},\alpha}(\boldsymbol{x}_{\text{H,t}}^{(i)}), y_{\text{up},\alpha}(\boldsymbol{x}_{\text{H,t}}^{(i)})]) \tag{35}$$

is the estimated PICP for each replication. Here $(\boldsymbol{\mathcal{X}}_{\text{H,t}}, \mathcal{Y}_{\text{H,t}})$ is a test set of size N_{t} of HF inputs and the corresponding noise-contaminated responses. Reliable PIs have $\bar{P}_\alpha \approx 1-2\alpha$.

The ACE metric aims instead to quantify the difference between the actual coverage of an interval and its designated nominal coverage. The ACE for a CI evaluation is defined as

$$\mathrm{ACE}_{\mathrm{CI},\alpha} = \mathrm{CICP} - (1 - 2\alpha). \tag{36}$$

Here, we use the MCICP instead of the CICP, and thus, for a CI of nominal coverage $1 - 2\alpha$ the ACE can be estimated by

$$\hat{E}_{\mathrm{CI},\alpha} = \bar{C}_\alpha - (1 - 2\alpha). \tag{37}$$

Likewise, the ACE for a PI can be estimated by

$$\hat{E}_{\mathrm{PI},\alpha} = \bar{P}_\alpha - (1 - 2\alpha). \tag{38}$$

An ACE value that is close to zero indicates reliable intervals. Moreover, a positive ACE denotes over-coverage, i.e., the interval actual coverage exceeds its nominal value, whereas a negative ACE indicates under-coverage, i.e., the interval actual coverage is lower than its nominal value.

3.3 Analytical 1-D example

Our first application is an analytical one-dimensional problem which serves well the purpose of visualisation of both the denoising performance and the confidence/prediction interval estimation. In this application, originally introduced in Brevault et al. (2020), the noise-free high- and low-fidelity models are given by:

$$f_{\mathrm{H}} = \left(\frac{x}{4} - \sqrt{2}\right) \sin(2\pi x + \pi) \tag{39}$$

$$f_{\mathrm{L}} = \sin(2\pi x), \tag{40}$$

where $x \sim \mathcal{U}[0, 2]$. The HF and LF functions are depicted in Figure 2.

To replicate the MF gray-box setting, we artificially contaminate the HF data with additive noise that follows a Gaussian distribution $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0, \sigma_{\varepsilon_{\mathrm{H}}})$.

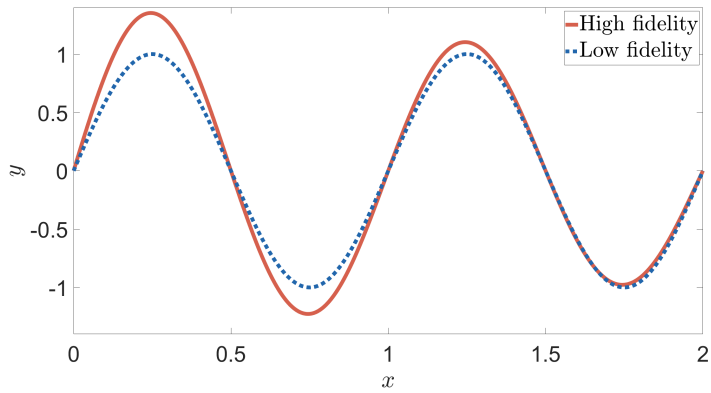
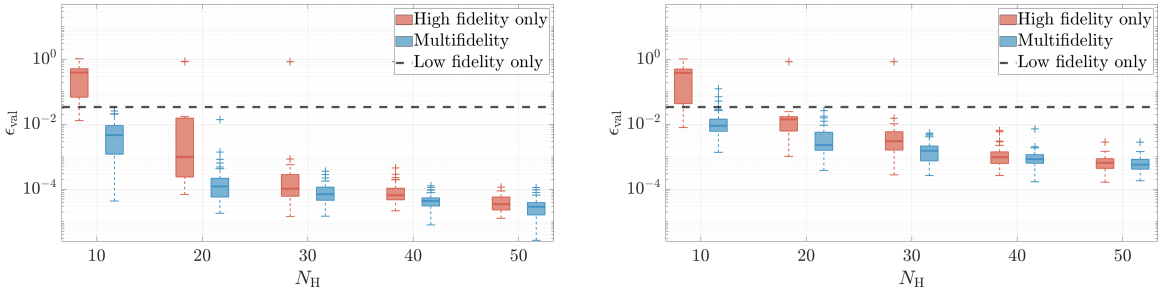


Figure 2: Analytical 1-D example – Noise-free high- and low-fidelity functions.

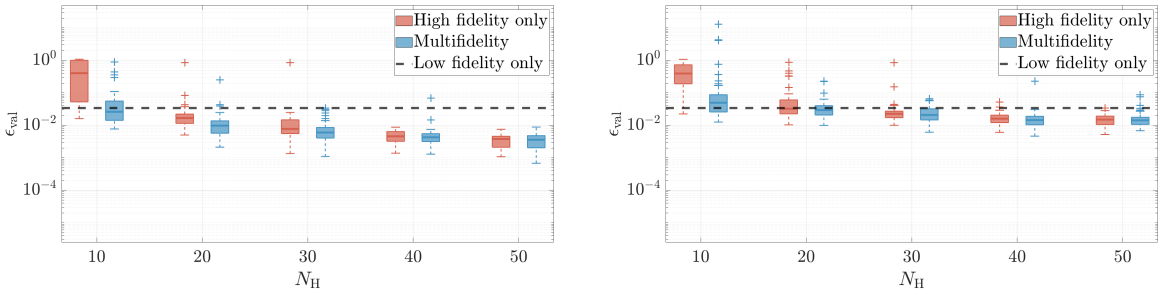
3.3.1 MFSM performance and convergence

We first assess the performance of the MFSM under varying levels of noise in the HF data. To this end, we compute the validation error ϵ_{val} for the cases where $\sigma_{\varepsilon_{\text{H}}}$ is set to 1%, 5%, 10%, and 20% of the standard deviation $\hat{\sigma}_{\text{H}}$ of the noise-free HF model, obtained from a PCE trained on $1,000$ noise-free HF data points, as described in, e.g., Blatman and Sudret (2011). As $\hat{\sigma}_{\text{H}} = 0.828$, the numerical values for $\sigma_{\varepsilon_{\text{H}}}$ read: (a) $\sigma_{\varepsilon_{\text{H}}} = 0.008$, (b) $\sigma_{\varepsilon_{\text{H}}} = 0.041$, (c) $\sigma_{\varepsilon_{\text{H}}} = 0.083$, (d) $\sigma_{\varepsilon_{\text{H}}} = 0.166$, respectively.

Here, for each noise level, the HF ED varies from 5 to 25 data points, while the LF ED is fixed in all experiments to 100 data points. For the computation of the validation errors shown in Figure 3, we use $N_{test} = 10^5$ noise-free HF data points generated using LHS in the input space.



(a) $\sigma_{\varepsilon_{\text{H}}} = 0.01\,\hat{\sigma}_{\text{H}}$ (b) $\sigma_{\varepsilon_{\text{H}}} = 0.05\,\hat{\sigma}_{\text{H}}$



(c) $\sigma_{\varepsilon_{\text{H}}} = 0.1\,\hat{\sigma}_{\text{H}}$ (d) $\sigma_{\varepsilon_{\text{H}}} = 0.2\,\hat{\sigma}_{\text{H}}$

Figure 3: Analytical 1-D example – Convergence of the validation error ϵ_{val} for increasing amount of HF training data under varying levels of noise in the HF data, $\varepsilon_{\text{H}} \sim \mathcal{N}(0, \sigma_{\varepsilon_{\text{H}}})$. Comparison of our MFSM (blue boxes) with a PCE model trained on HF data only (red). The dashed lines are the corresponding errors of a PCE model trained on LF data only.

The box plots in Figure 3 show the comparison between our MFSM and the PCE surrogate model trained solely on the corresponding HF data available in each case. In addition, the error of the PCE model trained solely on the available LF data is represented by dashed lines and serves as a baseline for comparison. Both single-fidelity PCE models have the same specifications as the PCE employed for the MFSM.

We observe that from as few as 5 HF training data points, our MFSM approach outperforms

surrogate models trained solely on either HF or LF data. This distinction becomes particularly evident for lower noise levels within the HF data. Hence, when HF data is scarce and data of different fidelities is present, the value of employing MFSMs as opposed to single-fidelity SMs becomes apparent. We notice that when sufficient HF data is available, e.g., here, approximately 20 data points for $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0,\, 0.01\hat{\sigma}_{\mathrm{H}})$, the MFSM and HF surrogate model performance is similar. In addition, as the level of noise in the HF data increases, the difference in performance of the MFSM and the HF SM diminishes. Indeed, Figure 3(d) shows a comparable performance between the two models, regardless of the size of the experimental design.

Moreover, we notice that for all noise levels, the multi-fidelity surrogate model error continuously decreases for increasing N_{H}, which indicates the convergence of the MFSM to the underlying noise-free HF model. However, the convergence rate is strongly influenced by the level of noise present in the HF data. As expected, the slowest convergence is observed for the strongest noise (Figure 3(d)).

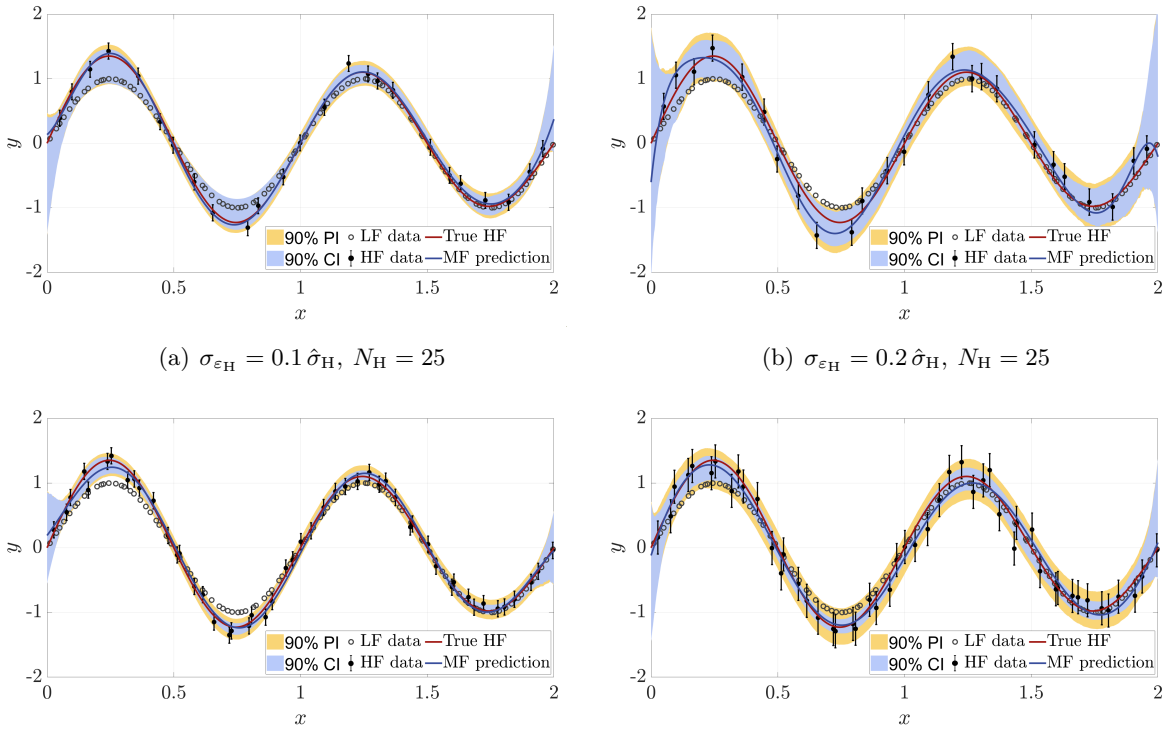
Finally, the scattering of ϵ_{val} of the MFSM, indicated by the box length, is generally smaller for larger HF experimental design sizes. This suggests more stable MFSM models that are less sensitive to the specific choice of the HF experimental design.

3.3.2 Confidence and prediction intervals

We now investigate the behaviour and performance of the confidence and prediction intervals for different HF experimental design sizes and different levels of noise in the HF observations. In the following, we use $N_{\mathrm{B}} = 1,000$ bootstrap replications to construct the CIs and PIs (Dubreuil et al., 2014).

Figure 4 shows the 90% CIs (blue area) and 90% PIs (yellow area) for the MFSM prediction (blue line) in four different cases occurring from all combinations among a realization of higher/lower noise in the HF observations and a realization of a larger/smaller HF ED. Here, again the LF ED is fixed to 100 samples. The HF and LF training data in each case is visualised by the black error bars and the gray circles, respectively. The error bars depict the 0.9-quantile of the estimated HF observation noise distribution. The plots appearing in the same row show that, for increasing noise and same EDs, both the CIs and the PIs become wider. This indicates that both the uncertainty about the underlying HF model and an unseen noise-contaminated HF observation increase. Moreover, the plots in the same columns reveal that, when the noise remains the same but the HF training data increases, the width of the CIs decreases. This means that our MFSM becomes more confident about where the regression model for the noise-free HF model lies. Also, the uncertainty about an unseen noise-contaminated HF observation, as shown by the PIs, decreases marginally. In this case, the PIs and CIs become more distinct, meaning that the uncertainty about a noise-contaminated HF observation is not anymore dominated by the regression model uncertainty.

Let us now proceed to the evaluation of the confidence and prediction intervals that our framework



(a) $\sigma_{\varepsilon_{\mathrm{H}}} = 0.1\,\hat{\sigma}_{\mathrm{H}},\ N_{\mathrm{H}} = 25$

(b) $\sigma_{\varepsilon_{\mathrm{H}}} = 0.2\,\hat{\sigma}_{\mathrm{H}},\ N_{\mathrm{H}} = 25$

(c) $\sigma_{\varepsilon_{\mathrm{H}}} = 0.1\,\hat{\sigma}_{\mathrm{H}},\ N_{\mathrm{H}} = 50$

(d) $\sigma_{\varepsilon_{\mathrm{H}}} = 0.2\,\hat{\sigma}_{\mathrm{H}},\ N_{\mathrm{H}} = 50$

Figure 4: Analytical 1-D example – 90% confidence and prediction intervals for the MFSMs trained on the illustrated HF and LF data sets. Plots in the same column exhibit the same noise level on the HF data, while the HF ED size increases. Plots in the same row use the same HF ED size and increasing level of noise on the HF data.

produces. Table 1 shows the detailed evaluation results including the MCICP and $\mathrm{ACE}_{\mathrm{CI}}$ for the CI evaluation, as well as the MPICP and $\mathrm{ACE}_{\mathrm{PI}}$, used for the PI evaluation. The MCICP and the $\mathrm{ACE}_{\mathrm{CI}}$ are estimated as in Equation (33) and Equation (37) respectively, where $N_{\mathrm{rep}} = 10$ replications with different seeds are performed. In each replication, the $\hat{C}_{\alpha}^{(j)}$ is computed as in Equation (32), using a test set consisting of $N_{\mathrm{t}} = 10,000$ data points $\psi_{\mathrm{H,t}}^{(i)}$ from the noise-free HF function, given by Equation (39). Similarly, to estimate the MPICP and the $\mathrm{ACE}_{\mathrm{PI}}$, in each of the N_{rep} replications, $\hat{P}_{\alpha}^{(j)}$ is computed as in Equation (35), using a test set with $N_{\mathrm{t}} = 10,000$ data points $y_{\mathrm{H,t}}^{(i)}$, where

$$y_{\mathrm{H,t}}^{(i)} = \psi_{\mathrm{H,t}}^{(i)} + \varepsilon_{\mathrm{H}}^{(i)}. \tag{41}$$

Here $\varepsilon_{\mathrm{H}}^{(i)}$ is a realization of the prescribed noise distribution, $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0, \sigma_{\varepsilon_{\mathrm{H}}})$ with $\sigma_{\varepsilon_{\mathrm{H}}}$ displayed in the second column of the table.

We can notice that, for all nominal coverage levels and every combination of $\sigma_{\varepsilon_{\mathrm{H}}}$ and N_{H}, the coverage of the PIs that our method provides is in excellent agreement with the corresponding nominal coverage. Specifically, the absolute value of the PI coverage error $\mathrm{ACE}_{\mathrm{PI}}$ rarely exceeds 1%. Moreover,the coverage of the constructed CIs is satisfactory, with the absolute value of

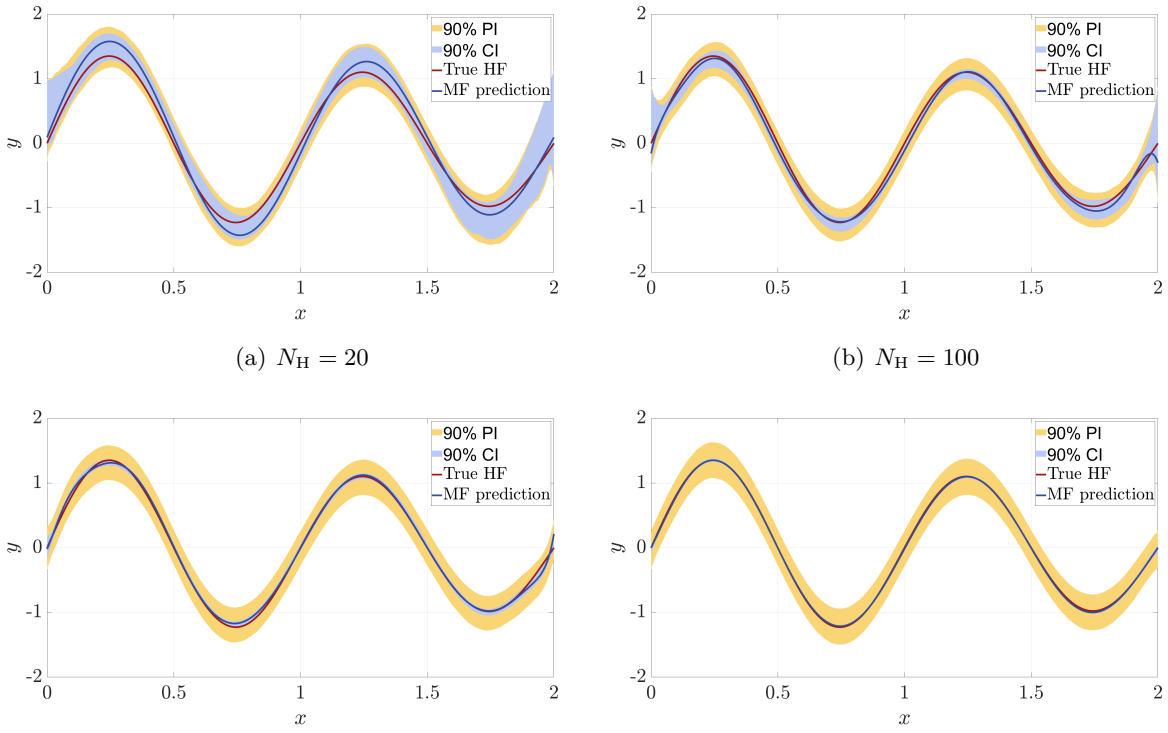
Table 1: Analytical 1-D example – Confidence and prediction intervals evaluation

$1-2\alpha$	$\sigma_{\varepsilon_{\mathrm{H}}}$	N_{H}	MCICP	$\mathrm{ACE}_{\mathrm{CI}}$	MPICP	$\mathrm{ACE}_{\mathrm{PI}}$
0.1	$0.1\,\hat{\sigma}_{\mathrm{H}}$	25	0.118	0.018	0.099	−0.001
	$0.1\,\hat{\sigma}_{\mathrm{H}}$	50	0.096	−0.004	0.110	0.010
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	25	0.125	0.025	0.104	0.004
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	50	0.132	0.032	0.104	0.004
0.5	$0.1\,\hat{\sigma}_{\mathrm{H}}$	25	0.548	0.048	0.494	−0.006
	$0.1\,\hat{\sigma}_{\mathrm{H}}$	50	0.544	0.044	0.530	0.030
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	25	0.569	0.069	0.509	0.009
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	50	0.561	0.061	0.526	0.026
0.9	$0.1\,\hat{\sigma}_{\mathrm{H}}$	25	0.929	0.029	0.888	−0.012
	$0.1\,\hat{\sigma}_{\mathrm{H}}$	50	0.992	0.092	0.905	0.005
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	25	0.960	0.060	0.906	0.006
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	50	0.991	0.091	0.919	0.019
0.95	$0.1\,\hat{\sigma}_{\mathrm{H}}$	25	0.988	0.038	0.941	−0.009
	$0.1\,\hat{\sigma}_{\mathrm{H}}$	50	1	0.050	0.950	0
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	25	1	0.050	0.952	0.002
	$0.2\,\hat{\sigma}_{\mathrm{H}}$	50	1	0.050	0.962	0.012

$\mathrm{ACE}_{\mathrm{CI}}$ being most of the times below 6%. The observed error in the CI coverage is almost exclusively due to over-coverage, indicated by a positive $\mathrm{ACE}_{\mathrm{CI}}$. We can attribute this to the presence of noise in the HF data. Indeed, we can notice that when the noise level is high ($\sigma_{\varepsilon_{\mathrm{H}}} = 0.2\,\hat{\sigma}_{\mathrm{H}}$), the CI coverage error increases consistently compared to instances where the noise is lower. Overall, the PIs achieve coverage much closer to the nominal level rather than the corresponding CIs.

Finally, we investigate the behaviour of the CIs and PIs asymptotically with respect to the HF experimental design size. Figure 5 shows the 90% CIs and PIs for four realizations of HF EDs of increasing size, from 20 up to 2500 data points. The level of noise is fixed, with $\sigma_{\varepsilon_{\mathrm{H}}} = 0.2\,\hat{\sigma}_{\mathrm{H}}$, and also fixed is the LF ED to 100 data points. Let us note that despite the LF ED typically being larger than the HF ED in practical applications, the last three out of the four cases do not align with this common scenario. In this study, we intentionally maintain this particular fixed LF ED across all cases to facilitate a focused investigation into the convergence behavior of the CIs and PIs of the MFSM with respect to the HF ED.

We observe that, as the HF ED increases, the CIs tend to converge to the regression model. This behavior aligns with our expectations, and reflects the fact that as more data becomes available, the epistemic uncertainty due to the lack of knowledge decreases and therefore, our MFSM exhibits increased confidence in its predictions. As regards the PIs, we notice that they tend



(a) $N_{\mathrm{H}} = 20$ (b) $N_{\mathrm{H}} = 100$

(c) $N_{\mathrm{H}} = 500$ (d) $N_{\mathrm{H}} = 2500$

Figure 5: Analytical 1-D example – Convergence of the confidence and prediction intervals for increasing HF ED size. In each plot, the blue area corresponds to the 90% CI, the yellow area to the 90% PI, while the red and blue lines depict the true noise-free HF response and the MF prediction respectively. To reduce the visual density of the plot, we did not include the HF and LF training data.

to converge to a non-zero width, indicative of the amount noise in the HF observations. This behavior is also expected, as the noise in the HF observations arises from aleatory uncertainty, and is thus irreducible regardless of the amount of the available training data. Consequently, predictions for unseen observations will inherently carry this uncertainty.

3.4 Truss model

In our second application, we aim to investigate the scalability of our method when applied to higher-dimensional problems. For this purpose, we consider a problem of engineering interest, and precisely, an ideal truss model with 23 bars and 6 upper cord nodes, as shown in Figure 6(a) (see Blatman and Sudret (2008)). This ten-dimensional model serves as the high fidelity.

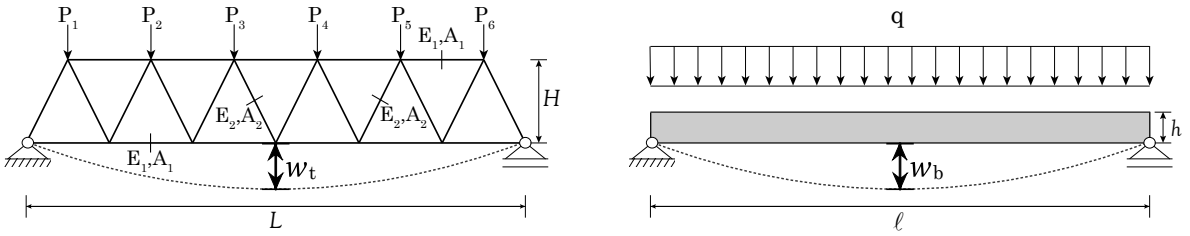
The HF truss structure has height H and length L, here considered constant, with $H = 2$m and $L = 24$m. The truss consists of two types of bars: horizontal bars with cross-sectional area A_1 and Young's modulus E_1, and oblique bars with cross-sectional area A_2 and Young's modulus E_2. The truss is loaded with six vertical loads P_i applied on the upper cord nodes. The quantity of

interest is the mid-span displacement of the truss, denoted as w_{t}. Here, w_{t} is calculated using an in-house finite element model programmed in Matlab. The geometrical and material properties of the truss members, as well as the loads, are modeled as random variables, with the distributions provided in Table 2.

Table 2: Truss model – Input variables and their distributions

Variable	Distribution	Mean	Standard deviation
E_1, E_2 [Pa]	Lognormal	$2.1 \cdot 10^{11}$	$2.1 \cdot 10^{10}$
A_1 [m^2]	Lognormal	$2 \cdot 10^{-3}$	$2 \cdot 10^{-4}$
A_2 [m^2]	Lognormal	$1 \cdot 10^{-3}$	$1 \cdot 10^{-4}$
P_1-P_6 [N]	Gumbel	$5 \cdot 10^{4}$	$7.5 \cdot 10^{3}$

In real engineering applications, the truss displacement w_{t} is commonly measured by laser measuring tools, and typically, such devices report a margin of error of 0.0015 m (1.5 mm). Therefore, in this example, we add artificial noise $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0,\, 0.0015)$ on the displacement w_{t}. To give a perspective on the level of the noise ε_{H} with respect to w_{t}, the standard deviation of w_{t}, as computed from a reference PCE trained on $1,000$ data points (see, e.g., Blatman and Sudret (2011)), is $\hat{\sigma}_{w_{\mathrm{t}}} \approx 0.0128$ m. This means that $\sigma_{\varepsilon_{\mathrm{H}}} \approx 0.12\, \hat{\sigma}_{w_{\mathrm{t}}}$.



(a) HF model: Ideal truss structure.

(b) LF model: Simply supported beam.

Figure 6: A truss structure with 23 bars and 6 upper cord nodes, used as the high fidelity (a), and its simply supported beam low-fidelity equivalent (b).

We now consider as the LF counterpart of the described HF truss model a homogeneous simply supported beam with length L and height h subjected to a uniform loading q, as shown in Figure 6(b). The mid-span deflection of the beam w_{b} can be computed as the sum of deflections due to bending and shear. For slender beams with $h \ll L$, we can neglect the shear contribution and approximate w_b as the deflection solely due to bending. Then,

$$w_{\mathrm{b}} = \frac{5\, q\, L^4}{384\, E\, I}. \tag{42}$$

Here, we consider $q = \frac{\sum_{i=1}^{6} P_i}{\ell}$. Moreover, the bending stiffness EI of the beam is determined by the Young's modulus E (material property) and the moment of inertia I (geometrical property).

We consider $E = E_1$, and assuming that only the cords contribute to I, we can approximate it as $I = 2A_1(\frac{H}{2})^2$. Thus, the mid-span deflection of the LF beam model can be computed as:

$$w_{\mathrm{b}} = \frac{5\,L^3\,\sum_{i=1}^{6} P_i}{384\,E_1\,2A_1\,\left(\frac{H}{2}\right)^2}, \tag{43}$$

where H is the height of the corresponding HF truss.

3.4.1 MFSM performance and convergence

Similarly to the previous analytical example, we now investigate the performance of our MFSM and its convergence with respect to the noise-free underlying HF truss model by computing the validation error ϵ_{val} for increasing HF experimental design size. More precisely, the HF ED size varies from 5 to 160 data points contaminated with the noise following the prescribed distribution, while the LF ED is fixed in all experiments to 300 data points. The larger amount of LF training data used compared to the previous application is due to the higher dimensionality and complexity of this application. Again, for the computation of ϵ_{val}, we use $N_{test} = 10^5$ noise-free HF data points generated using Latin Hypercube Sampling in the input space.

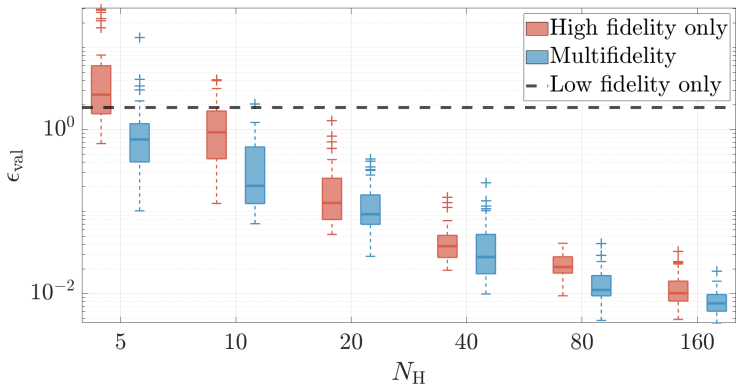


Figure 7: Truss model – Convergence of the validation error ϵ_{val} for increasing number of HF training data. Comparison of our MF gray-box with a PCE model trained on HF data only. The dashed lines are the corresponding errors of a PCE model trained on LF data only.

From Figure 7, we observe that our MFSM outperforms both the PCE model trained on HF data only, and the PCE model trained solely on LF data, across all the considered HF ED sizes. The performance difference between the MFSM and the HF PCE model is more pronounced when the available HF data comprises fewer than 20 data points.

3.4.2 Confidence and prediction intervals

We now proceed to the construction and evaluation of CIs and PIs for our MF truss model. We set the HF ED size equal to 80 samples, as the previous study demonstrated satisfactory performance at this sample size ($\epsilon_{\mathrm{val}} \approx 1\%$), with only marginal improvement observed when doubling the

sample amount. The HF data is contaminated with noise $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0, 0.0015)$. Furthermore, the LF ED consists of 300 samples. Similarly to the previous application, $N_{\mathrm{B}} = 1,000$ bootstrap replications are performed for the construction of the CIs and PIs. Also, for the computation of the evaluation metrics reported in Table 3, $N_{\mathrm{rep}} = 10$ replications are performed, and the test sets for the CI and PI evaluation consist of $N_{\mathrm{t}} = 10,000$ data points each, obtained as described in Section 3.3.2.

Table 3: Truss model – Confidence and prediction intervals evaluation, where $\varepsilon_{\mathrm{H}} \sim \mathcal{N}(0, 0.0015)$, $N_{\mathrm{H}} = 80$, and $N_{\mathrm{L}} = 300$

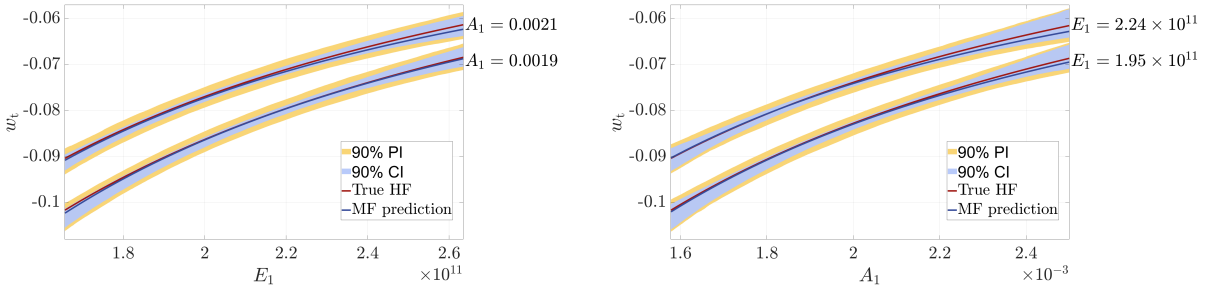
$1 - 2\alpha$	MCICP	$\mathrm{ACE}_{\mathrm{CI}}$	MPICP	$\mathrm{ACE}_{\mathrm{PI}}$
0.1	0.124	0.024	0.109	0.009
0.5	0.593	0.093	0.542	0.042
0.9	0.959	0.059	0.935	0.035
0.95	0.986	0.036	0.973	0.023

From Table 3, we observe that for all the nominal coverage levels examined, our method provides reliable CIs and PIs. More precisely, the PI average coverage error ranges from less than 1% to approximately 4%, while the corresponding error for the CIs varies from 2% to 9%. The observed error is always due to over-coverage, and though not ideal, is preferable to under-coverage and considered acceptable. Moreover, the PIs exhibit again closer coverage to the nominal levels compared to the CIs.

To illustrate the constructed CIs and PIs in this application, we select the random variables E_1 and A_1 as the most important ones, according to a sensitivity analysis on the HF truss model performed by Blatman and Sudret (2011). Figure 8 illustrates the 90% CIs (blue area) and PIs (yellow area) along slices in the two selected dimensions (with all the other parameters kept at their mean values), as well as the true HF model response and our MFSM response (red and blue line respectively) for the selected HF and LF experimental designs. Figure 8(a) shows these quantities as a function of E_1 for two different values of A_1 that correspond to its 0.25- and 0.75-quantiles, while the rest input random variables are fixed at their mean. Similarly, Figure 8(b) shows the same quantities as a function of A_1 for two different values of E_1.

3.5 Real-world application: aero-servo-elastic simulation of a wind turbine

In our last application, we aim to explore the applicability and performance of our framework in a real-world application involving real wind turbine simulations, performed by Abdallah et al. (2019). For this study, an onshore wind turbine standing on a 90 m tower is considered, with a rotor diameter of 110 m and a rated power of 2 MW. The goal is to investigate the impact of the wind speed, turbulence intensity, and shear profile on the variation of the extreme loads, and specifically the maximum flapwise bending moment at the wind turbine blade root. The



(a) Selected dimension: Young's modulus E_1

(b) Selected dimension: cross-sectional area A_1

Figure 8: Truss model – 90% confidence and prediction intervals along slices in the two selected dimensions for the MFSM trained on 80 HF and 300 LF data points. In each subplot, the blue area corresponds to the 90% CI, the yellow area to the 90% PI, while the red and blue lines depict the true noise-free HF response and the MFSM prediction respectively.

wind speed, turbulence intensity, and shear exponent are modeled as random variables, whose distributions are described in Table 4.

Table 4: Wind turbine simulations – Input variables and their distributions

Variable	Distribution	Parameters
Wind speed (U) [m/s]	Uniform	$[4, 25]$
Turbulence intensity (σ_U) [m/s]	Uniform	$[0.1, 6]$
Wind shear exponent (α) [-]	Uniform	$[-1, 1.5]$

Abdallah et al. (2019) used two different numerical aeroservo-elastic simulators, namely Bladed and FAST. Simulation data from Bladed are considered as the high-fidelity data, while the FAST simulation data are considered to be the low-fidelity data. Details on the technical characteristics of the Bladed and FAST simulators can be found in Abdallah et al. (2019); Bossanyi (2003); Jonkman and Buhl (2005). A 10-minute time series simulation in FAST takes about 5 minutes to run in real time, whereas the same simulation in Bladed takes approximately 30 minutes. Consequently, there are fewer Bladed simulations compared to FAST simulations. The experimental designs for the Bladed and FAST simulations can be found in Table 5.

Table 5: Experimental design for Bladed and FAST simulations

Simulator	Wind speed (U)	Turbulence intensity (σ_U)	Wind shear exponent (α)
Bladed	$4, 8, 10, 12, 15, 20, 25$	$0.1, 1, 2, 3, 4, 5, 6$	$\pm 1, \pm 0.6, \pm 0.2, \pm 0.1, 0, 1.5$
FAST	4, 5, 6, ..., 25	$0.1, 1, 2, 3, 4, 5, 6$	$\pm 1, \pm 0.6, \pm 0.2, \pm 0.1, 0, 1.5$

Each combination of wind speed, turbulence intensity, and shear exponent is used to generate

wind time series realizations with 12 different stochastic seeds for Bladed and 24 different seeds for FAST. Excluding certain combinations of input parameters that are not realistic results in $4,344$ and $33,480$ simulations for Bladed and FAST respectively. Our HF data set is the mean system response over the 12 time series from Bladed, thus 362 data points. Moreover, our LF data set is the mean system response over the 24 time series from FAST, resulting in a total of $1,395$ data points.

In this case study, both the HF and the LF data are considered noisy, due to the stochasticity in the wind turbine simulations. Our MF framework remains applicable as is.

3.5.1 MFSM performance and convergence

In this application, we explore the performance of our MFSM for increasing HF experimental design size, equal to $10\%, 20\%, ..., 70\%$ of the total HF data available, while keeping the LF ED fixed to all the available LF data. We compute the validation error ϵ_{val} of the MFSM on a test set consisting of the 30% of the HF data which was not used for training any of the MFSMs at each given replication: $N_{\text{test}} = 0.3 \times 362 = 109$ data points.

As shown in Figure 9, our MFSM outperforms the PCE model trained on HF data only, the difference being more evident especially for small HF EDs. We can notice that we achieve satisfactory performance ($\epsilon_{\text{val}} < 1\%$) already for $N_{\text{H}} = 144$, which corresponds to 40% of the available HF data.

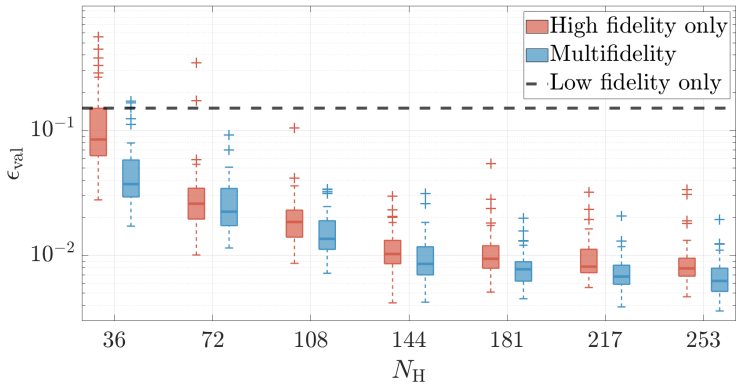


Figure 9: Wind turbine application – Convergence of the validation error ϵ_{val} for increasing number of HF training data. Comparison of our MF gray-box with a PCE model trained on HF data only. The dashed lines are the corresponding errors of a PCE model trained on LF data only.

Let us note that in this application, ϵ_{val} is not expected to approach zero as the HF ED size increases, because ϵ_{val} is now computed with respect to noisy HF data. Instead, ϵ_{val} is expected to converge to a value representative of the noise present in the HF data.

3.5.2 Confidence and prediction intervals

In this section, we provide the CIs and PIs for our MFSM prediction. In this real-world application, the true noise-free HF function is unknown, limiting our ability to assess the reliability of the constructed CIs. Thus, we are able to only appraise the constructed PIs on a test set from the HF data. We use 70% of the available HF data for training and the rest 30% for the PI evaluation, and we perform $N_{\text{rep}} = 10$ replications using different seeds to account for the statistical uncertainty in the HF random design and the bootstrap sampling.

From Table 6, we observe that the coverage of our PIs is close to the nominal, albeit generally overestimated. Once again in this application, the errors are due to over-coverage, which is more evident for nominal coverage 50% and 90%.

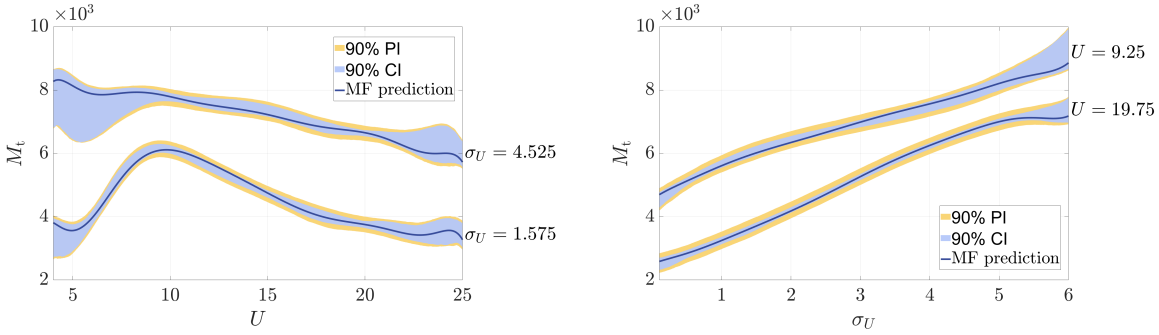
Table 6: Wind turbine application – Prediction intervals evaluation

$1-2\alpha$	MPICP	ACE_{PI}
0.1	0.119	0.019
0.5	0.583	0.083
0.9	0.950	0.050
0.95	0.981	0.031

The predicted extreme flapwise bending moment as well as the 90% CIs and PIs in the wind speed and turbulence intensity dimensions are illustrated in Figure 10. More precisely, Figure 10(a) shows the MFSM prediction and the corresponding intervals as a function of the wind speed U for two different values of the turbulence intensity σ_{U} that correspond to its 0.25- and 0.75-quantiles, while the wind shear exponent α is fixed at its mean. Similarly, Figure 10(b) depicts the MFSM prediction and the corresponding intervals as a function of σ_{U} for two different values of U.

4 Conclusions

In this paper, we presented a novel and comprehensive framework for multi-fidelity surrogate modeling that effectively handles noisy data and incorporates epistemic uncertainty arising from limited training information. Our regression-based approach aims to emulate the assumed underlying noise-free HF model, and provides accurate predictions and denoising capabilities. It also offers uncertainty estimations with respect to not only the underlying HF model, but also to unseen noise-contaminated HF observations in the form of confidence and prediction intervals respectively, constructed using the bootstrap methodology. The proposed framework is applied in the field of gray-box modeling, where noisy measurements, considered as the high fidelity, are combined with white-box computer simulations, treated as the low fidelity. However, our framework is not limited solely to the grey-box scenario of combining experimental data



(a) Selected dimension: wind speed U

(b) Selected dimension: turbulence intensity σ_U

Figure 10: Wind turbine application – 90% confidence and prediction intervals along slices in the two selected dimensions for the MFSM trained on 253 HF and 1,395 LF data points. In each subplot, the blue area corresponds to the 90% CI, the yellow area to the 90% PI, while the blue line depicts the MFSM prediction respectively.

and computer simulations. Its versatility extends to situations where both the HF and the LF components are experiments or simulations.

Our framework proves its efficacy in various scenarios, including a one-dimensional analytical example and a ten-dimensional application that incorporates a high-fidelity finite element model alongside a low-fidelity analytical approximation. In both scenarios, noise was artificially added to the HF data. In these synthetic examples, our multi-fidelity surrogate modeling method clearly outperforms both surrogate models trained on the available high- and low-fidelity data separately, and it shows convergence to the noise-free HF model with increasing number of HF training data. Moreover, the constructed confidence and prediction intervals exhibit remarkably high reliability, achieving coverage close to the nominal levels. Finally, our framework demonstrates its versatility and potential by being applied on a real-world example involving wind turbine simulations of different fidelity levels. In this application, our method provides again accurate predictions and reliable prediction intervals.

It should be noted that the reliability of the confidence and prediction intervals comes at the cost of computational time for their construction. For low-dimensional problems, this time can be considered negligible, but in higher dimensions (≥ 10) this is not the case anymore. Nonetheless, this time overhead is incurred only during the training phase and can subsequently be mitigated as the bootstrap results can be stored. Thus, any subsequent inference including predictions at unobserved points along with their associated confidence and prediction intervals, can be instantaneously accessed.

In our future work, we plan to extend our methodology in various directions. Firstly, different types of surrogate models and data fusing methodologies can be explored for our multi-fidelity surrogate model construction. Additionally, to improve the performance of confidence and prediction intervals, different techniques can be investigated, such as more advanced bootstrap

methods (Efron and Tibshirani, 1994). Finally, the provided uncertainty estimations for the multi-fidelity model predictions can be employed for different purposes, one being the adaptive design of sampling strategies to obtain new samples from the high- and the low-fidelity models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors express their sincere appreciation to Styfen Schär for his valuable contribution to the second of the three applications presented in this paper, particularly for formulating the low-fidelity simply supported beam approximation of the high-fidelity truss structure.

The research presented in this paper is part of the GREYDIENT project, funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 955393. GREYDIENT's support is gratefully acknowledged.

References

Abbiati, G., S. Marelli, N. Tsokanas, B. Sudret, and B. Stojadinović (2021). A global sensitivity analysis framework for hybrid simulation. *Mechanical Systems and Signal Processing 146*, 106997.

Abdallah, I., C. Lataniotis, and B. Sudret (2019). Parametric hierarchical Kriging for multi-fidelity aero-servo-elastic simulators – Application to extreme loads on wind turbines. *Probabilistic Engineering Mechanics 55*, 67–77.

Berchier, M. (2016). Multi-fidelity surrogate modelling with polynomial chaos expansions. Master's thesis, ETH Zürich.

Blatman, G. and B. Sudret (2008). Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *Comptes Rendus Mécanique 336*(6), 518–523.

Blatman, G. and B. Sudret (2011). Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *Journal of Computational Physics 230*(6), 2345–2367.

Bossanyi, E. A. (2003). GH Bladed theory manual. Technical Report 282/BR/009, Garrad Hassan & Partners Ltd.

Brevault, L., M. Balesdent, and A. Hebbal (2020). Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. *Aerospace Science and Technology 107*, 106339.

Carpenter, J. and J. Bithell (2000). Bootstrap confidence intervals: when, which, what? a practical guide for medical statisticians. *Statistics in Medicine 19*(9), 1141–1164.

Chevreuil, M., R. Lebrun, A. Nouy, and P. Rai (2015). A least-squares method for sparse low rank approximation of multivariate functions. *SIAM/ASA Journal on Uncertainty Quantification 3*(1), 897–921.

Cutajar, K., M. Pullin, A. Damianou, N. Lawrence, and J. González (2018). Deep Gaussian processes for multi-fidelity modeling. In *32nd Neural Information Processing Systems Conference*, Montreal, Canada.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola, and V. Vapnik (1996). Support vector regression machines. In M. Mozer, M. Jordan, and T. Petsche (Eds.), *Advances in Neural Information Processing Systems*, Volume 9. MIT Press.

Dubreuil, S., M. Berveiller, F. Petitjean, and M. Salaün (2014). Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. *Reliability Engineering & System Safety 121*, 263–275.

Efron, B. and R. J. Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Fernández-Godino, M. G. (2023). Review of multi-fidelity models. *Advances in Computational Science and Engineering 1*(4), 0–50.

Ficini, S., U. Iemma, R. Pellegrini, A. Serani, and M. Diez (2021). Assessing the performance of an adaptive multi-fidelity Gaussian process with noisy training data: A statistical analysis. In *AIAA AVIATION 2021 FORUM*. American Institute of Aeronautics and Astronautics.

Forrester, A. I., A. Sóbester, and A. J. Keane (2007). Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 463*(2088), 3251–3269.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics 9*(6), 1218–1228.

Ghanem, R. and P. Spanos (2003). *Stochastic finite elements: A spectral approach* (2nd ed.). Courier Dover Publications, Mineola.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne (2004). Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing 26*(2), 448–466.

Jonkman, J. and M. Buhl (2005). FAST user's guide. Technical Report NREL/EL-500-38230, National Renewable Energy Laboratory Golden, CO, USA.

Kennedy, M. C. and A. O'Hagan (2000). Predicting the output from a complex computer code

when fast approximations are available. *Biometrika 87*(1), 1–13.

Kennedy, M. C. and A. O'Hagan (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(3), 425–464.

Kerleguer, B., C. Cannamela, and J. Garnier (2024). A Bayesian neural network approach to multi-fidelity surrogate modeling. *International Journal for Uncertainty Quantification 14*(1), 43–60.

Kutner, M., C. J. Nachtsheim, J. Neter, and W. Li (2005). *Applied linear statistical models* (5th ed.). New York: McGraw-Hill.

Kuya, Y., K. Takeda, X. Zhang, and A. I. J. Forrester (2011). Multifidelity surrogate modeling of experimental and computational aerodynamic data sets. *AIAA Journal 49*(2), 289–298.

Le Gratiet, L. and J. Garnier (2014). Recursive co-Kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification 4*(5), 365–386.

Ljung, L. (1998). System identification. In A. Procházka, J. Uhlíř, P. W. J. Rayner, and N. G. Kingsbury (Eds.), *Signal Analysis and Prediction*, pp. 163–173. Boston, MA: Birkhäuser Boston.

Lüthen, N., S. Marelli, and B. Sudret (2021). Sparse polynomial chaos expansions: Literature survey and benchmark. *SIAM/ASA Journal on Uncertainty Quantification 9*(2), 593–649.

Lüthen, N., S. Marelli, and B. Sudret (2022). Automatic selection of basis-adaptive sparse polynomial chaos expansions for engineering applications. *International Journal for Uncertainty Quantification 12*(3), 49–74.

Marelli, S., N. Lüthen, and B. Sudret (2022). UQLab user manual – Polynomial chaos expansions. Technical report, Chair of Risk, Safety and Uncertainty Quantification, ETH Zurich, Switzerland. Report UQLab-V2.0-104.

Marelli, S. and B. Sudret (2014). UQLab: A framework for uncertainty quantification in Matlab. In *Vulnerability, Uncertainty, and Risk*. American Society of Civil Engineers.

Marelli, S. and B. Sudret (2018). An active-learning algorithm that combines sparse polynomial chaos expansions and bootstrap for structural reliability analysis. *Structural Safety 75*, 67–74.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 2*, 239–245.

Meng, X., H. Babaee, and G. E. Karniadakis (2021). Multi-fidelity Bayesian neural networks: Algorithms and applications. *Journal of Computational Physics 438*, 110361.

Ng, L. W.-T. and M. S. Eldred (2012). Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation. In *53rd AIAA/ASME/ASCE/AHS/ASC*

Structures, Structural Dynamics and Materials Conference, Honolulu, Hawaii, pp. 1852.

Palar, P. S., T. Tsuchiya, and G. T. Parks (2016). Multi-fidelity non-intrusive polynomial chaos based on regression. *Computer Methods in Applied Mechanics and Engineering 305*, 579–606.

Perdikaris, P., M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis (2017). Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 473*(2198), 20160751.

Raissi, M., P. Perdikaris, and G. Karniadakis (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics 378*, 686–707.

Raissi, M., P. Perdikaris, and G. E. Karniadakis (2017). Inferring solutions of differential equations using noisy multi-fidelity data. *Journal of Computational Physics 335*, 736–746.

Rasmussen, C. and C. Williams (2006). *Gaussian processes for machine learning* (Internet ed.). Adaptive computation and machine learning. Cambridge, Massachusetts: MIT Press.

Rogers, T. J., G. R. Holmes, E. J. Cross, and K. Worden (2017). On a grey box modelling framework for nonlinear system identification. In *Special Topics in Structural Dynamics, Volume 6*, pp. 167–178. Springer International Publishing.

Schellenberg, A., S. Mahin, and G. Fenves (2009). Advanced implementation of hybrid simulation. Technical report, University of California, Berkeley, Berkeley, CA.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics 6*(2), 461–464.

Torre, E., S. Marelli, P. Embrechts, and B. Sudret (2019). Data-driven polynomial chaos expansion for machine learning regression. *Journal of Computational Physics 388*, 601–623.

Tulleken, H. J. A. F. (1993). Grey-box modelling and identification using physical knowledge and Bayesian techniques. *Automatica 29*(2), 285–308.

Wan, C., Z. Xu, P. Pinson, Z. Y. Dong, and K. P. Wong (2014). Optimal prediction intervals of wind power generation. *IEEE Transactions on Power Systems 29*(3), 1166–1174.

Xiu, D. and G. E. Karniadakis (2002). The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing 24*(2), 619–644.

Yan, C., R. Vescovini, and L. Dozio (2022). A framework based on physics-informed neural networks and extreme learning for the analysis of composite structures. *Computers & Structures 265*, 106761.

Zhang, C., L. Liu, H. Wang, X. Song, and D. Tao (2022). SCGAN: stacking-based generative adversarial networks for multi-fidelity surrogate modeling. *Structural and Multidisciplinary Optimization 65*(6), 1–16.

Zhang, Y., N. H. Kim, C. Park, and R. T. Haftka (2018). Multifidelity surrogate based on single linear regression. *AIAA Journal 56*(12), 4944–4952.