Computationally efficient adaptive design of experiments for global metamodeling through integrated error approximation and multi-criteria search strategies

Sang-ri Yi[1*] and Alexandros A. Taflanidis[2], A.M.ASCE

[1]*Department of Civil & Environmental Engineering,*
University of California, Berkeley, CA 94720, USA

[2]*Department of Civil & Environmental Engineering & Earth Sciences*
University of Notre Dame, Notre Dame, IN 46556, USA

Abstract

Gaussian processes (GPs) are a popular technique for global metamodeling applications. Their objective in such settings is to establish an efficient and globally accurate approximation of the response surface of computationally expensive simulation models. When developing such GPs, the design of (simulation) experiments (DoE) plays an important role in reducing the required number of model runs for obtaining accurate approximations. Sequential (adaptive) selection of experiments can provide significant advantages, especially when the response surface is characterized by localized nonlinearities. Such adaptive DoE strategies for global metamodeling applications typically focus on minimizing the predictive GP variance, representing an exploration strategy, while recent developments have additionally considered the reduction of the GP bias obtained through cross validation, representing an exploitation strategy. While significant focus has been placed on the definition of appropriate adaptive DoE criteria, computational challenges still exist that limit the widespread adoption of adaptive DoE techniques, for example, related to the additional computational demand for identifying the optimal new experiment(s), or the necessity to establish proper schemes to combine exploration and exploitation strategies. To address these specific challenges, this research investigates two new adaptive DoE formulations. The first one focuses on the approximation of the popular integrated mean square error (IMSE) DoE criterion. The computationally demanding GP predictive variance update (after addition of each candidate experiment), required in the original IMSE formulation, is replaced by an approximation based on the current predictive variance and the domain of influence that surrounds each new experiment. The approximation is established through a parametric formulation that leverages the GP kernel to describe the

* Corresponding author. Department of Civil Engineering and Environmental Engineering, University of California, Berkeley. Email: yisangri@berkeley.edu

aforementioned domain, with characteristics that are progressively calibrated across the GP training stages, to minimize the discrepancy between the actual and the approximated IMSE. The second formulation establishes a multi-criteria search for simultaneously identifying multiple Pareto optimal experiments that balance exploration and exploitation objectives, replacing conventional strategies that establish a weighted combination of these objectives to promote a single DoE selection criterion.

Keywords: Adaptive design of experiments (DoE), Global Gaussian process, Multi-objective optimization, Leave-one-out cross-validation weights, Integrated mean square error (IMSE)

Introduction

The growing complexity of computational simulation models in the various engineering fields has increased the need for efficient surrogate modeling techniques (Razavi et al. 2012, McBride and Sundmacher 2019, Forrester and Keane 2009). These models, also referenced as metamodels, can offer an efficient data driven mapping between the input-output relationship of high-fidelity simulation models, formulated based on an observation set, frequently referenced as experiments or training points. Different data-driven surrogate modeling approaches exist, such as artificial neural networks (Jain et al. 1996), polynomial chaos expansion (Blatman and Sudret 2010), and Bayesian networks (Byun and Song 2021), and among them, Gaussian process (GP) metamodels, also referred to as Kriging in geostatistical contexts, have gained wide popularity for their flexibility arising from their interpolative property as well as the ability to quantify the predictive variance of the function estimates associated with the lack/abundance of training data (Rasmussen and Nickisch 2010, Gramacy 2020, Sacks, Welch et al. 1989, Kleijnen 2017). Furthermore, past research efforts have shown that GPs are relatively reliable compered to alternative formulations for a small to moderate size of training samples, i.e., up to a few thousand, (Deisenroth and Ng 2015, Forrester and Keane 2009), while recent advances in GP approaches, such as sparse training methods, are further pushing this limit to larger datasets (Csató and Opper 2002, Liu et al. 2020).

It is widely acknowledged that one of the most critical aspects in the GP development (and for many other surrogate modeling techniques) is the selection of the set of simulation experiments that serve as training

points, a process formally known as design of experiments (DoE). More efficient DoE strategies can accommodate development of accurate GPs using smaller number of training points, reducing the computational burden for performing simulations of the high-fidelity model. Space-filling DoE strategies perform well for many practical applications (McKay et al. 2000) by uniformly populating the domain of interest. However, to further improve efficiency, the experiments can be selected progressively in stages, leveraging in each new stage information provided by a GP metamodel that is developed using the experiments selected from previous stages (Provost et al. 1999). Such an adaptive, or *active learning*, strategy involves extra computations to find the optimal experiments as well as to perform the repeated GP parameter calibrations (across the different stages), but can eventually lead to large reductions of the number of high-fidelity model evaluations to achieve the same prediction accuracy, by adaptively identifying potential low accuracy domains where addition of new experiments can maximize their utility (value). Depending on the purpose for which the GP is trained, adaptive DoE strategies can be distinguished to application-oriented (Kleijnen and Beers 2004, Moustapha et al. 2022, Zhang and Taflanidis 2018, Kim, J. and Song 2020), when metamodel is intended to establish a specific task that indirectly defines a region of interest, or general-purposed, when metamodel is intended to serve as a universal approximation of the original high-fidelity model. The latter, which the focus of this paper lies on, is often referred to as *global* surrogate modeling and the DoE for this class aims to identify experiments that most effectively minimize the average error across the entire input-domain, without distinguishing any sub-domains of specialized interest (Liu et al. 2018, Kleijnen 2009, Pandita et al. 2021).

Most adaptive DoE approaches for global surrogate modeling applications focus on the GP prediction variance, assuming that locating an experiment at the domain of high variance will accommodate better exploration of the input domain, and ultimately reduce the variability of the metamodel estimates. Such variance-based adaptive DoE strategies share similar objectives to non-adaptive, space-filling approaches, in a way that they identify the experiment(s) that has furthest distance from the previously training points, however, the measure of distance is defined differently to account for the length of correlation in each input dimension (Picheny et al. 2010, Sacks, Welch et al. 1989, Welch 1983), leveraging the characteristics of the

GP kernel in this definition. Furthermore, more advanced formulations, such as the integrated mean square error (IMSE), discount the importance of experiments near the boundary of the input domain by considering their influence only *within* the specific input domain of interest. IMSE is often preferable to alternative variance-based DoE criteria, such as maximum mean squared error (MMSE) and mean squared error (MSE), in terms of numerical stability, optimality and the way it tackles near-boundary experiments (Beck and Guillas 2016, Krause et al. 2008). However, its implementation is computationally demanding because it requires integration over the domain of interest of the updated predictive variance, obtained by considering the addition of each candidate new experiment (and involving inversion of the updated GP covariance matrix), whereas simplified measures such as MSE (Jin et al. 2002) directly utilize the pre-update predictive variance without involving any significant computations.

To supplement the exploration established through variance-based DoE approaches, strategies that enforce exploitation principles in global surrogate modeling context can be additionally considered, which has shown to be particularly beneficial for approximating functions with localized nonlinearities. A common approach for achieving this objective is to define a bias measure by interpolating the GP cross-validation error (Liu et al. 2016, Kyprioti et al. 2020, Jin et al. 2002), and incorporate this measure in the DoE through combination with a variance-based exploration strategy, e.g., by weighted summation (Fuhg et al. 2021) or multiplication (Kyprioti et al. 2020). Furthermore, recent research efforts have shown that, like many other adaptive search algorithms, adaptive DoE strategies for global surrogate modeling can perform substantially better when the tradeoff between the two aforementioned search objectives, exploration and exploitation, is carefully considered (Garud et al. 2017, Liu et al. 2018, Fuhg et al. 2021). Unfortunately, existing efforts to accomplish the latter typically rely on user-selected weighting coefficients and tuning parameters. Furthermore, the computational challenges associated with variance-based exploration strategies (like IMSE) directly extend for such implementations, creating additional limitations for their use in practice.

To address these challenges, this paper develops two new adaptive DoE strategies. The primary contribution examines the approximation of the IMSE measure [or the weighted IMSE that extends IMSE to consider bias measure-based exploitation (Kyprioti et al. 2020)], focusing on reduction of the computational

burden for the IMSE estimation without lowering the quality of the identified experiments. The computationally demanding updating of the variance field given new candidate experiments requiring, as discussed above, inversion of the updated covariance matrices, is replaced by an approximation for the amount of variance reduction around each candidate new experiments through the introduction of a decaying shape function that describes its domain of influence. This approximation accommodates estimation of the reduction of the (weighted) IMSE with very small computational effort, by integrating the (weighted) shape function. The shape function approximation is established through the GP kernel and the introduction of an additional adjustment parameter. Proper selection of this parameter is critical for the approximation accuracy, and a progressive learning scheme is introduced to accommodate the selection, calculating the actual IMSE and comparing it to the approximated one to choose the optimal adjustment parameter value. To achieve the desired computational benefits, the implementation gradually switches to sole use of the IMSE approximation once sufficient confidence for it is achieved.

The secondary contribution examines the adaptive DoE selection as a multi-objective optimization problem, considering the variance (exploration) and bias (exploitation) criteria as separate objectives, and identifying Pareto optimal experiments that establish a balance between them. The sorted solutions (Pareto front) provide the batch selection of experiments, and are further truncated at a desired batch limit, for example related to computational resources within a parallel computing environment. The desired final batch of experiments is chosen based on the distance from the utopia point of the Pareto front, while also adopting an updating of the Pareto front GP variance to avoid selecting experiments in close proximity to one another. Through this implementation the two objectives are examined with no need to establish a preselected weighting scheme between them, naturally balancing the exploitation and exploration without introducing additional tuning parameters. Furthermore, the proposed method enables us to batch-select multiple experiments without additional effort, accommodating a formulation that is naturally aligned with modern parallel simulation computational environments.

The remaining of the paper is organized as follows. The next section provides a brief overview of GP models with discussions focusing on computational complexity of the variance and bias estimation, while the

section after that reviews adaptive DoE criteria. The following section introduces the approximation for the variance update after addition of each new experiment, accommodating a computationally efficient IMSE implementation, while in the section after that, the multi-criteria DoE scheme, that naturally balances between exploration and exploitation objectives, is presented. Finally, the following section presents illustrative examples to showcase the proposed methods, considering both benchmark functions as well as practical problems from the domain of natural hazards engineering, for which there has been a renewed interest in global surrogate modeling applications (Deierlein and Zsarnóczay 2019).

Overview of Gaussian Process (GP) regression

Let $z = \mathbb{F}(\mathbf{x})$ represent the high-fidelity (computationally expensive) simulator that is approximated through the surrogate model, with $\mathbf{x} \in \mathbb{R}^{n_x}$ representing the n_x-dimensional input and $z \in \mathbb{R}$ the scalar response output. The Gaussian process (GP) regression approximates $\mathbb{F}(\mathbf{x})$, as a GP realization utilizing a training set of simulations from the high-fidelity model. Different GP variants exist (Rasmussen and Nickisch 2010), but perhaps the most popular (Gramacy 2020, Kleijnen 2017), and the one adopted here, uses: (i) a mean function corresponding to a linear regression, $\mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}$, where $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{n_b}$ represents the n_b-dimensional vector of basis functions (for example, low order polynomials) and $\boldsymbol{\beta} \in \mathbb{R}^{n_b}$ the vector of regression coefficients; and (ii) a stationary covariance function $\text{cov}(\mathbf{x}, \mathbf{x}') = \tilde{\sigma}^2 R(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta})$ where $\tilde{\sigma}^2$ is a constant representing the process variance, and $R(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta})$ is the autocorrelation function between inputs $\mathbf{x}$ and $\mathbf{x}'$, having hyper-parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$, and frequently also referenced as GP correlation kernel. Examples of popular correlation functions include radial basis, Matérn, or exponential functions (Rasmussen and Nickisch 2010), with hyper-parameters representing the characteristic correlation length for each input dimension (dictating rate of correlation decay), and $R(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}) = R(d(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}))$ expressed as function of the normalized distance

$$d(\mathbf{x}, \mathbf{x}' | \boldsymbol{\theta}) = \left\| \frac{\mathbf{x} - \mathbf{x}'}{\boldsymbol{\theta}} \right\| \tag{1}$$

where $\|.\|$ is some chosen vector norm, and vector division is defined elementwise herein.

The GP model has three sets of unknown parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\}$ that need to be identified using the available training set. To formalize implementation, and all subsequent DoE discussions, assume that the training set corresponds to the input-output pair of n simulations (also referenced as observations or training points), $\{\mathbf{x}_i, z_i; i = 1,...,n\}$, and let us denote by $\mathbf{X} \in \mathbb{R}^{n \times n_x}$ and $\mathbf{Z} \in \mathbb{R}^n$ the corresponding input matrix and output vector, with ith row associated with the ith simulation. Also let $\mathbf{D} = \{\mathbf{X}, \mathbf{Z}\}$ denote the input-output observation set, and define the matrix of basis functions $\mathbf{F}(\mathbf{X}) \in \mathbb{R}^{n \times n_b}$, whose ith row corresponds to $\mathbf{f}(\mathbf{x}_i)^T$, the correlation matrix for the database $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta}) \in \mathbb{R}^{n \times n}$ whose $\{i, j\}$ element corresponds to $R(\mathbf{x}_i, \mathbf{x}_j | \boldsymbol{\theta})$. Note that the chosen notation emphasizes the dependence of all quantities on $\mathbf{X}$ [or $\mathbf{D}$ and $\boldsymbol{\theta}$] to better frame the adaptive DoE schemes. Finally, let $\mathbf{r}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) \in \mathbb{R}^n$ define the correlation vector between an input $\mathbf{x}$ and each training set component, with ith element of the vector corresponding to $R(\mathbf{x}, \mathbf{x}_i | \boldsymbol{\theta})$. The optimum values of Θ are commonly established using maximum likelihood estimation (MLE), a process detailed in Appendix A. These values will be denoted as $\boldsymbol{\beta}^*$, $\boldsymbol{\theta}^*(\mathbf{D})$, and $(\tilde{\sigma}^*(\mathbf{D}, \boldsymbol{\theta}))^2$, with dependencies on $\mathbf{D}$ and $\boldsymbol{\theta}$ explicitly noted for the latter two quantities (omitted for $\boldsymbol{\beta}^*$ for brevity).

Given the observations $\mathbf{D}$ and the calibrated parameters Θ, the GP approximates the response at an arbitrary point $\mathbf{x}$ as Gaussian with mean and variance given, respectively, by (Sacks, Welch et al. 1989):

$$\tilde{z}(\mathbf{x}|\mathbf{D}, \Theta) = \mathbf{f}(\mathbf{x})^T \boldsymbol{\beta}^* + \mathbf{r}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta})^T \mathbf{R}(\mathbf{X}|\boldsymbol{\theta})^{-1} (\mathbf{Z} - \mathbf{F}(\mathbf{X})\boldsymbol{\beta}^*) \tag{2}$$

$$\sigma^2(\mathbf{x}|\mathbf{D}, \Theta) = \left(\tilde{\sigma}^*(\mathbf{D}, \boldsymbol{\theta})\right)^2 \underline{\sigma}^2(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) \tag{3}$$

where the normalized variance has been defined as:

$$\underline{\sigma}^2(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) = 1 - \mathbf{r}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta})^T \mathbf{R}(\mathbf{X}|\boldsymbol{\theta})^{-1} \mathbf{r}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) + \mathbf{u}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta})^T (\mathbf{F}(\mathbf{X})^T \mathbf{R}(\mathbf{X}|\boldsymbol{\theta})^{-1} \mathbf{F}(\mathbf{X}))^{-1} \mathbf{u}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) \tag{4}$$

with $\mathbf{u}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) = \mathbf{F}(\mathbf{X})^T \mathbf{R}(\mathbf{X}|\boldsymbol{\theta})^{-1} \mathbf{r}(\mathbf{x}|\mathbf{X}, \boldsymbol{\theta}) - \mathbf{f}(\mathbf{x})$. Eqs. (2) and (3) are also frequently referred to as predictive mean and predictive variance, respectively. When purpose of the surrogate modeling is to establish deterministic predictions, then the predictive mean of Eq. (2) is adopted as the response approximation [e.g.

(Prebeg et al. 2014, Kyprioti et al. 2021, Contreras et al. 2020)], whereas when probabilistic predictions are preferred, the full Gaussian distribution of the GP, combining Eqs. (2) and (3), is utilized [e.g. (Kim, J. and Song 2020, Bodenmann et al. 2021, Jia and Taflanidis 2013)]. Note that the normalized variance in Eq. (4) does not depend on the response **Z**. This property motivates the introduction of adaptive design of experiments because it allows us to foresee how much variance can be reduced by adding a new experiment $\mathbf{x}_{new}$ into the existing training set.

With respect to computational complexity, the numerically intensive component in Eqs. (2) and (3) is the inversion of the correlation matrix, $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$ which has complexity $O(n^3)$ but needs to be performed only once, as it is independent of input **x**. Therefore, once $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$ is inverted for a calibrated GP, predictions for a new input **x** can be established with negligible computational cost, providing an efficient surrogate model. For numerical stability, the inversion of $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$ within the GP formulation is typically replaced by forward or backward substitution operations utilizing its Cholesky factorization (Roustant et al. 2012, Lophaven et al. 2002). Appendix B discusses specifics in the context of estimation of the predictive variance, which is the task that is required within the DoE formulation discussed in the next section.

Finally, the predictive capabilities of the calibrated GP can be quantified using cross validation statistics. Typically, this is accomplished using leave-one-out cross validation (LOOCV) (Kyprioti et al. 2020, Fuhg et al. 2021, Kleijnen 2009), with details provided in Appendix A. This entails estimation for each training point of the leave-one-out (LOO) error e_i^{cv}, corresponding to the predictions established for this specific point using the remaining observations, excluding this specific point from the database. Once this error is estimated, any desired validation metric can be utilized to assess the global metamodel accuracy. In this study, the normalized root mean square error (NRMSE), denoted $NRMSE_{cv}$ and reviewed in Appendix A will be used as such metric.

Adaptive design of experiments (DoE) for global surrogate modeling

The sequential (adaptive) design of experiments (DoE) obtains the training points iteratively in stages: the new experiment(s) are determined as a function of the previous experiments and the GP that is calibrated using these experiments, with an objective to maximize the expected utility of the new experiment(s). Let $X^d \subset \mathbb{R}^{n_x}$

represent the domain set of interest for the metamodel development, and denote by $\mathbf{x}_{new}$ each examined new simulation input (feasible experiment). The adaptive DoE is formulated as an optimization for the selection of the experiment that maximize a measure of information acquisition:

$$\mathbf{x}_{new}^{*} = \arg\max_{\mathbf{x}_{new} \in X^{d}} \overline{\Phi}(\mathbf{x}_{new}) \tag{5}$$

where $\overline{\Phi}(\bullet)$ is called acquisition or merit function (Koehler and Owen 1996, Johnson et al. 1990). Eq. (5) can be alternatively formulated using an equivalent minimization criterion

$$\mathbf{x}_{new}^{*} = \arg\min_{\mathbf{x}_{new} \in X^{d}} \Phi(\mathbf{x}_{new}) \tag{6}$$

where $\Phi(\bullet)$ is a measure that quantifies the remaining uncertainty (variance) after adding the new experiment $\mathbf{x}_{new}$ (Sacks, Welch et al. 1989). Throughout this paper, DoE objective functions without the bar notation (Φ) represent those for the minimization problem whereas those with the bar ($\overline{\Phi}$) are for maximization problems. The adaptive DoE requires, furthermore, convergence criteria to terminate iterations when the metamodel performance, quantified for example using LOOCV statistics, improves beyond a desired threshold, or when the available computational budget is exceeded.

Figure 1 demonstrates this adaptive DoE procedure, examining also the formulation of batch selection of experiments, discussed later. The computationally intensive steps within each iteration are: (i) the high-fidelity simulation(s); (ii) the GP parameters calibration; and (iii) the adaptive DoE selection requiring an optimization problem [Eq. (5) or (6)]. It is important to note that if the evaluation of the DoE objective function requires involved computations, the selection of the next experiment itself becomes a non-trivial task in terms of total computation demand. Of course, in most applications of interest, the computational burden associated with the high-fidelity simulation model is substantially larger than the cost of the adaptive DoE selection, providing value for implementation of advanced DoE schemes that can accommodate a reduction of high-fidelity simulations to achieve the same level of metamodel accuracy. Still, reduction of the computational burden of these DoE schemes is an important secondary objective, as long as the quality of the identified experiments is not compromised (primary objective).

A final remark is warranted about batch selection of experiments, something that can be particularly useful when parallel computing environments are utilized for the high-fidelity simulations. Most of the existing DoE methods, including those discussed in this paper, can accommodate a batch selection of multiple experiments through appropriate small modifications. This is commonly done by selecting one experiments at a time using the merit functions defined in Eqs. (5) or (6), until a predetermined batch number of experiments is acquired as shown in Figure 1. This one-at-a-time selection of experiments is suboptimal when compared to the alternative of simultaneous selection of the entire batch, since the latter better incorporates the correlations between the candidate experiments in assessing their expected total utility, but nevertheless, is very practical and widely advocated (Ginsbourger 2014, Vazquez and Bect 2011). To assess the information infused in the training set by each individual experiment, and ultimately avoid choosing close proximity experiments, the merit function(s) should be approximately updated after each selection. As illustrated in Figure 1, since the response at the location of the new experiment is not yet known, this is done without performing GP recalibration. For example, the predictive variance, which is ingredient of most merit functions, depends only on the training sample locations at each experiments [$\mathbf{X}$ in Eq.(4)] and not on the response [$\mathbf{Z}$] as long as $\boldsymbol{\theta}$ is known. Therefore, after adding a new experiment, this variance and the corresponding merit function can be updated without evaluation of the high-fidelity model. In contrast, in a purely sequential approach, $\boldsymbol{\theta}$ needs to be updated each time before selecting a new experiment.

Variance-based adaptive DoE

One of the widely used merit functions of adaptive DoE in global GP metamodeling is the integrated mean squared error (IMSE), which seeks the experiment that minimizes the average updated variance within the domain of interest (Sacks, Welch et al. 1989, Asher et al. 2015). The corresponding DoE objective function is:

$$\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) = \int_{\mathbf{x} \in X^d} w(\mathbf{x}) \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) d\mathbf{x} \tag{7}$$

where $w(\mathbf{x})$ is a weight function to prioritize any desired sub-domain within X^d , and $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$ is the updated normalized predictive variance after addition of $\mathbf{x}_{new}$, which is estimated as follows. Obtain first the updated correlation vector $\mathbf{r}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$, basis function matrix $\mathbf{F}(\mathbf{X}, \mathbf{x}_{new})$ and correlation matrix

$\mathbf{R}(\mathbf{X}, \mathbf{x}_{new} \mid \boldsymbol{\theta})$, by adding a row and, when needed, a column into $\mathbf{r}(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})$, $\mathbf{F}(\mathbf{X})$ and $\mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})$, respectively, as:

$$\mathbf{r}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta}) \\ R(\mathbf{x}_{new}, \mathbf{x} \mid \boldsymbol{\theta}) \end{bmatrix} \tag{8}$$

$$\mathbf{F}(\mathbf{X}, \mathbf{x}_{new}) = \begin{bmatrix} \mathbf{F}(\mathbf{X}) \\ \mathbf{f}(\mathbf{x}_{new})^T \end{bmatrix} \tag{9}$$

$$\mathbf{R}(\mathbf{X}, \mathbf{x}_{new} \mid \boldsymbol{\theta}) = \begin{bmatrix} \mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta}) & \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta}) \\ \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})^{\mathrm{T}} & R(\mathbf{x}_{new}, \mathbf{x}_{new} \mid \boldsymbol{\theta}) \end{bmatrix} \tag{10}$$

Then $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$ is given by:

$$\begin{aligned} \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{\text{new}}, \boldsymbol{\theta}) = 1 - \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{\text{new}}, \boldsymbol{\theta})^T \mathbf{R}(\mathbf{X}, \mathbf{x}_{\text{new}} \mid \boldsymbol{\theta})^{-1} \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{\text{new}}, \boldsymbol{\theta}) + \\ \mathbf{u}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{\text{new}}, \boldsymbol{\theta})^T (\mathbf{F}(\mathbf{X}, \mathbf{x}_{\text{new}})^T \mathbf{R}(\mathbf{X}, \mathbf{x}_{\text{new}} \mid \boldsymbol{\theta})^{-1} \mathbf{F}(\mathbf{X}, \mathbf{x}_{\text{new}}))^{-1} \mathbf{u}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{\text{new}}, \boldsymbol{\theta}) \end{aligned} \tag{11}$$

with $\mathbf{u}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) = \mathbf{F}(\mathbf{X}, \mathbf{x}_{new})^T \mathbf{R}(\mathbf{X}, \mathbf{x}_{new} \mid \boldsymbol{\theta})^{-1} \mathbf{r}(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) - \mathbf{f}(\mathbf{x})$. The efficient estimation of this updated variance is discussed in Appendix B. Note that the IMSE criterion is commonly expressed with respect to the variance of Eq. (3), which includes the process variance, but for illustration clarity the normalized variance is used herein, since the process variance is independent of $\mathbf{x}$. Also the weight function $w(\mathbf{x})$ is typically ignored in the IMSE formulation, with $w(\mathbf{x})=1$, though the representation of Eq. (7) is preferred here as it accommodates a unified description with the formulation incorporating GP bias as weights, discussed in the next section. For estimating the IMSE integral, efficient approximate formulations can be found in the literature under some regularity conditions for the GP and the X^d domain boundary (Ankenman et al. 2008, Cole et al. 2021). In this study, Monte Carlo Integration (MCI) is preferred, since it can accommodate arbitrary weight functions, correlation kernel shapes, and domains. Alternatively, Quasi-Monte Carlo could had been used. Using n_q samples $\{\mathbf{x}^{(q)}\}_{q=1,\ldots,n_q}$ following uniform distribution in X^d, MCI establishes the following approximation for the objective function:

$$\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \approx \frac{1}{n_q} \sum_{q=1}^{n_q} w(\mathbf{x}^{(q)}) \underline{\sigma}^2(\mathbf{x}^{(q)} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) \tag{12}$$

Based on the computational details presented in Appendix B, the computational cost for estimating $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ for each $\mathbf{x}_{new}$ is $O(n^2)$ for updating quantities that are common across all MCI samples and then additional $O(n^2)$ for each of the n_q integration points, though parallel computations can reduce the burden of the latter estimation. The computational workflow to identify the optimal $\mathbf{x}_{new}$ will be discussed later on.

Beyond IMSE, other criteria can be established that involve some form of integration or maximization of the updated variance within X^d (Picheny et al. 2010, Sacks, Schiller et al. 1989). All these formulations correspond to one-step-lookahead approaches, requiring the evaluation of the updated predictive variance $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$ and estimation of some function involving it to express the metamodel accuracy. As shown above, the computational cost of these approaches is significant for each candidate experiments $\mathbf{x}_{new}$ that is examined within the DoE optimization. To reduce this computational complexity, alternative formulations have been examined (Ginsbourger 2014, Jin et al. 2002), with the most popular one adopting as merit function the mean squared error (MSE) at the current training stage, leading to:

$$\overline{\Phi}_{\text{MSE}}(\mathbf{x}_{new}) = w(\mathbf{x}_{new})\underline{\sigma}^2(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{13}$$

This objective represents a special case of maximum entropy criterion (Jin et al. 2002, Liu et al. 2017). The MSE measure in Eq. (13) approximates the expected reduction of the error, corresponding, therefore, to an objective function targeted for maximization. MSE simply involves estimation of the variance at $\mathbf{x}_{new}$, establishing significant computational savings compared to IMSE. At the same time, it has two important shortcomings: (i) it cannot guarantee optimality after adding each experiment [no variance updating]; (ii) it does not consider a domain of influence of each experiment [no integration part]. The latter is manifested as tendency to place experiments around boundary region, something frequently argued (Beck and Guillas 2016, Krause et al. 2008) to represent an inefficient DoE scheme, especially if n_x is large (relative proportion of domain representing boundary increases). Therefore, in terms of the quality of the identified experiments, the IMSE criterion is preferable to the MSE criterion for global metamodeling applications, despite its larger computational burden. These characteristics motivate the developments established later in this manuscript,

aiming to reduce the computational burden to levels similar to MSE while keeping the quality of new experiments of IMSE.

Adaptive DoE with bias measure

Recent efforts have examined adaptive strategies that exploit the surrogate model response predictions to self-identify important regions in the DoE, something expressed in global metamodeling applications by utilizing the prediction bias (Xu et al. 2014, Le Gratiet and Cannamela 2015). The LOO error can be used to quantify this bias and be leveraged to guide the selection of $w(\mathbf{x})$, promoting exploitation strategies within the global metamodeling DoE. Since the LOO error is only known for the discrete locations corresponding to the training set, some form of interpolation (or kernel smoothing) needs to be introduced (Jin et al. 2002, Kyprioti et al. 2020), providing a continuous approximation of the error as a function of $\mathbf{x}$. Establishing such an approximation for the squared error leads to:

$$(\tilde{e}^{cv}(\mathbf{x}))^2 = \frac{\sum_{i=1}^{n} \gamma(\mathbf{x},\mathbf{x}_i)(e_i^{cv})^2}{\sum_{i=1}^{n} \gamma(\mathbf{x},\mathbf{x}_i)} \tag{14}$$

where $\tilde{e}^{cv}(\mathbf{x})$ is the LOO error approximation and $\gamma(\mathbf{x},\mathbf{x}_i)$ is the interpolation/smoothing function, expressing proximity between the error values at $\mathbf{x}$ and $\mathbf{x}_i$. In (Kyprioti et al. 2020) different such functions were explored and the one promoted was the nearest neighbor (NN) interpolation

$$\gamma(\mathbf{x},\mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{x} \in V_i \\ 0 & \text{else} \end{cases} \tag{15}$$

where V_i represents a Voronoi cell associated with each training point $\mathbf{x}_i$ and defined using the normalized distance of Eq. (1), so that $V_i = \left\{ \mathbf{x} \,\middle|\, d(\mathbf{x},\mathbf{x}_i \,|\, \boldsymbol{\theta}) \le d(\mathbf{x},\mathbf{x}_j \,|\, \boldsymbol{\theta}),\ \forall\ j \ne i,\ (i,j = 1,...,n) \right\}$. An alternative choice will be examined here, adopting an exponentially decaying smoothing function

$$\gamma(\mathbf{x},\mathbf{x}_i) = \exp(-d(\mathbf{x},\mathbf{x}_i \,|\, \boldsymbol{\theta})^2) \tag{16}$$

Consideration of the LOO error leads to the weighted IMSE and MSE criteria, denoted as IMSE_w and MSE_w herein. Adopting the formulation in (Kyprioti et al. 2020) this is established by choosing weight function $w(\mathbf{x})$ as

$$w(\mathbf{x}) = [(\tilde{e}^{cv}(\mathbf{x}))^2]^{\rho} \quad (17)$$

in the formulations of Eq. (7) (for $IMSE_w$) and Eq. (13) (for MSE_w), where $\rho \in [0, \infty)$ is a tuning parameter that is used to control the balance between the bias and variance indicators (Kyprioti et al. 2020). Note that for $\rho \to 0$, the aforementioned measures reduce to the classical IMSE and MSE measures, respectively. Ultimately, these classical measures have only exploration attributes, whereas their weighted counterparts combine both exploration and exploitation features. The balancing between these two features is established through ρ whose selection is non-trivial (Kyprioti et al. 2020), something that motivates the multi-objective DoE discussed later in this manuscript. Figure 2 illustrates some of the aforementioned concepts related to the bias and variance adaptive DoE components using a two-dimensional cosine weighted Gaussian mixture function. More details about this function are included in the illustrative examples section. Specifically, this figure includes the original function [part (a)], as well as GP-based predictions established utilizing 21 training points (depicted with red circles in some of the subplots): the predictive mean of Eq. (2) [part(b)]; the LOO error shown in Eq. (45) (red circles in parts [c] and [d]) as well as the corresponding weight function $w(\mathbf{x})$ obtained using either the NN interpolation of Eq. (15) [part (c)] or the exponentially decaying smoothing function of Eq. (16) [part (d)]; the predictive variance of Eq. (4) [part (e)], as well as its counterpart using the bias weights [part (f)]. Comparison of parts (c) and (d) depicts the difference of the alternative approaches for choosing bias weights: the NN interpolation provides a discontinuous (at the Voronoi cell boundaries) weighting function, whereas the exponential decaying kernel smoothing accommodates a smoother function, that, though, no longer interpolates the available LOO estimates. Comparison of parts (e) and (f) shows clearly how the suggested weights incorporate in the DoE bias information about the GP predictions, altering the domains of importance within X^d. Use of the predictive variance [part (e)] instead of its weighted counterpart [part (f)] naively places importance in domains in which the established GP already has high accuracy (small bias).

Optimization scheme for identification of new experiments

Optimization of Eq. (5) [or (6)] is known to have multiple local minima and for this reason a random search approach is recommended for its solution (Kyprioti et al. 2020). For the IMSE_w this combines the following steps for identification of each $\mathbf{x}_{new}^{*}$, denoted herein as IMSE_w-*SE* (Sequential Exact) algorithm.

Step 0 [Initialization]: Given the GP training set $\mathbf{X}$, correlation Kernel, $R(\mathbf{x},\mathbf{x}'|\boldsymbol{\theta})$, basis vector, $\mathbf{f}(\mathbf{x})$, and hyper-parameters, $\boldsymbol{\theta}$, calculate correlation matrix, $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$, and basic function matrix, $\mathbf{F}(\mathbf{X})$, as well as the lower Cholesky factorization of the former $\mathbf{L}(\mathbf{X}|\boldsymbol{\theta})$ and matrix $\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta})$ related to the QR decomposition of $\mathbf{L}^{-1}(\mathbf{X}|\boldsymbol{\theta})\mathbf{F}(\mathbf{X})$ [see Appendix B for details]. If DoE incorporates bias information, estimate the LOO error through Eq. (45) using, additionally, the response output $\mathbf{Z}$ and the MLE regression vector $\boldsymbol{\beta}^{*}$ of Eq. (41). Note that many (perhaps all) of these quantities will be available from the GP calibration stage. Choose interpolation function $\gamma(\mathbf{x},\mathbf{x}_i)$ for the weight estimation and, if needed, ρ.

Step 1 [Candidate experiments]: Generate n_c candidate experiments $\mathbf{X}^{candi}=\{\mathbf{x}_{new}^{(c)}\}_{c=1,\ldots,n_c}$ following a uniform distribution in X^d.

Step 2 [(Optional) Preliminary screening of candidate experiments using MSE_w]: For all $\mathbf{X}^{candi}$ evaluate $\underline{\sigma}^2(\mathbf{x}_{new}^{(c)}|\mathbf{X},\boldsymbol{\theta})$ utilizing Eq. (47) and $w(\mathbf{x}_{new}^{(c)})$ combining Eqs. (17) and (14). Retain only the $n_r=a_r n_c$ candidate experiments that correspond to the highest values of MSE_w, $\overline{\Phi}_{\text{MSE}}(\mathbf{x}_{new}^{(c)}|\mathbf{X},\boldsymbol{\theta})=w(\mathbf{x}_{new}^{(c)})\underline{\sigma}^2(\mathbf{x}_{new}^{(c)}|\mathbf{X},\boldsymbol{\theta})$, with a_r being the desired percentage of candidate experiments, e.g. 10%, that have larger weighted-variance values and so are more likely to correspond to the final optimal solution.

Step 3 [Integration points]: Generate n_q samples $\{\mathbf{x}^{(q)}\}_{q=1,\ldots,n_q}$ following a uniform distribution in X^d to be used for the MCI. Estimate $w(\mathbf{x}^{(q)})$ for all these samples combining Eqs. (17) and (14).

Step 4 [Calculation of objective function]: For each candidate (or retained from Step 2) experiment, define the updated basis function matrix $\mathbf{F}(\mathbf{X},\mathbf{x}_{new}^{(c)})$ as in Eq. (9), and then estimate the updated Cholesky

matrix $L(\mathbf{X},\mathbf{x}_{new}^{(c)}\,|\,\boldsymbol{\theta})$ using Eq. (49) and then the updated matrix $\mathbf{M}_R(\mathbf{X},\mathbf{x}_{new}^{(c)}\,|\,\boldsymbol{\theta})$ [see Appendix B for details]. Then for each of the integration points, calculate $\underline{\sigma}^2(\mathbf{x}^{(q)}\,|\,\mathbf{X},\mathbf{x}_{new}^{(c)},\boldsymbol{\theta})$ using Eq. (48). Finally obtain the IMSE$_w$ objective function $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)}\,|\,\mathbf{X},\boldsymbol{\theta})$ through MCI using Eq. (12).

Step 5 [Final selection]: Select as new experiment $\mathbf{x}_{new}^{*}$ the one that provides the minimum value for $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)}\,|\,\mathbf{X},\boldsymbol{\theta})$ among the n_c (or n_r if Step 2 was performed) candidate experiments evaluated in Step 4.

Based on the previous discussions on computational complexity, the most computationally demanding step of this process is Step 4, which requires updating of the variance and conducting MCI for each of the candidate experiments considered. The prescreening of experiments in Step 2 accommodates a reduction of this burden, as it removes candidate experiments that are not expected to correspond to the optimum, ignoring experiments in sub-domains of X^d with low current prediction variability (Kyprioti et al. 2020). Since Step 2 utilized an MSE-based objective function, its computational cost is minor. Of course the desire is to avoid impacting the final solution, i.e. experiment identified being same as if Step 2 were not utilized, something that evidently depends on the value of a_r. In this paper a_r=0.1 is used, as this value was shown to establish a reasonable compromise between two conflicting objectives: efficiency (smaller a_r desired) and robustness (larger a_r desired) (Zhang et al. 2018). For the MSE$_w$, the above optimization procedure is drastically simplified: only Steps 1 and 2 need to be implemented, with a single experiment, $n_r = 1$, identified at Step 2.

For identifying batch experiments, after Step 5, the identified $\mathbf{x}_{new}^{*}$ is augmented in $\mathbf{X}$, the updated correlation vector $\mathbf{r}(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new}^{*},\boldsymbol{\theta})$, basis function matrix $\mathbf{F}(\mathbf{X},\mathbf{x}_{new}^{*})$ and correlation matrix $\mathbf{R}(\mathbf{X},\mathbf{x}_{new}^{*}\,|\,\boldsymbol{\theta})$ are estimated [according to Eqs. (8)-(10)], then the $\mathbf{L}(\mathbf{X},\mathbf{x}_{new}^{*}\,|\,\boldsymbol{\theta})$ is updated according to Eq. (49), and the matrix $\mathbf{M}_R(\mathbf{X},\mathbf{x}_{new}^{*}\,|\,\boldsymbol{\theta})$ related to the QR decomposition of $\mathbf{L}^{-1}(\mathbf{X},\mathbf{x}_{new}^{*}\,|\,\boldsymbol{\theta})\mathbf{F}(\mathbf{X},\mathbf{x}_{new}^{*})$ is estimated. After these calculations, Steps 1-5 of the IMSE$_w$-*SE* algorithm are repeated to obtain the next experiment within the batch selection.

Approximation of integrated mean square error measure

Approximation of updated variance

This section proposes an approximation of the updated predictive variance $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$. Objective is to approximate $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$ using information about the original variance $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\boldsymbol{\theta})$, avoiding therefore the requirement to update the GP characteristics for each $\mathbf{x}_{new}$ examined. Though this approximation is couched here within the IMSE/IMSE$_w$ adaptive DoE formulation, it can be utilized within any setting that requires use of $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$. Foundation of the approximation (Le Gratiet and Cannamela 2015) is the concept that when adding a new experient $\mathbf{x}_{new}$, there is a volume of influence where the prediction variance is reduced, i.e. the reduction is a concentrated around the location of the point $\mathbf{x}_{new}$, and it decays as the distance from $\mathbf{x}_{new}$ increases. To formalize this concept, define the variance reduction as:

$$V_d(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) = \underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\boldsymbol{\theta}) - \underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) \tag{18}$$

Le Gratiet & Cannamela (2015) proposed the proportionality approximation:

$$\int_{\mathbf{x}\in X^d} V_d(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})d\mathbf{x} \propto \underline{\sigma}^2(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})\prod_{j=1}^{n_x}\theta_j \tag{19}$$

where θ_j is the hyper-paremeter in the covariance kernel associated with the correlation length of *j*-th input dimension, $\prod_{j=1}^{n_x}\theta_j$ corresponds to the volume of influence of the new experiment, and $\underline{\sigma}^2(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})$ represents the scale of variance reduction. Equivalently Eq. (19) may be viewed to represent an approximation of $V_d(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$ that corresponds to a uniform reduction (no dependance on $\mathbf{x}$) within a domaim of influence defined through representative lengths θ_j in each input dimension and centered around $\mathbf{x}_{new}$ for which the variance reduction is $\underline{\sigma}^2(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})$ [i.e. the previous variance] after the addition of the new experiment. It should be noted that the developments in Le Gratiet & Cannamela (2015) were couched within a multi-fidelity modeling context, while the presentation of Eq. (19) is simplified to an equivalent single-fidelity impementaiton. If used to approximate the IMSE criteria, the approximation of Eq. (19) drastically reduces

the computaitonal burden, making it comparable to the computaitonal burden of the MSE DoE formulation (only evaluation of $\underline{\sigma}^2(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ needed). It has, though, two signifciant shortcomings: (i) since the volume of influence is constant across the domain, it becomes exactly equivalent to the MSE measure unless some boundary correction is employed; (ii) for better accomodating extension to the weighted measures (for example IMSE_w), an approximation of $V_d(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$ is needed, instead of its integral, and in this case, a non-uniform expression (as function of $\mathbf{x}$) seems more appropriate (may accommodate higher accuracy), since the actual $V_d(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$ has a strong dependance on $\mathbf{x}$.

To address these shortocomings, the variance reduction around the new experiment $\mathbf{x}_{new}$ is approximated to be proportional to the original variance for each point $\mathbf{x}$, $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})$, and to a shape function that incorporates the influence of $\mathbf{x}_{new}$ on $\mathbf{x}$, i.e.

$$V_d(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) \simeq \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta}) A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda) \tag{20}$$

where $A(\mathbf{x}, \mathbf{x}_{\text{new}} \mid \lambda)$ a distance-decaying shape function centered at $\mathbf{x}_{new}$ with hyper-parameter λ. The rate of reduction is assumed be proportational to the that of the correlation kernel with λ representing a flexible exponent (selection discussed next), leading to

$$A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda) = R(\mathbf{x}, \mathbf{x}_{new} \mid \boldsymbol{\theta})^{\lambda} \tag{21}$$

Approximation of Eq. (20) using a shape function like the one in Eq. (21) greatly simplifies calculations since the only computationally complex component is $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})$, which, once calculated, can be reused for all $\mathbf{x}_{new}$ examined. Using in the DoE formulation the approximation of the updated variance :

$$\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta}) \simeq \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})[1 - A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda)] = \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})[1 - R(\mathbf{x}, \mathbf{x}_{new} \mid \boldsymbol{\theta})^{\lambda}] \tag{22}$$

provides an easily computable approximate merit function. For example, for the IMSE, this provides the approximation

$$\Phi^{\lambda}_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) = \int_{\mathbf{x} \in X^d} w(\mathbf{x}) \underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})[1 - R(\mathbf{x}, \mathbf{x}_{new} \mid \boldsymbol{\theta})^{\lambda}] d\mathbf{x} \tag{23}$$

which using the MCI setting examined previously in Eq. (12), can be estimated as

$$\Phi^{\lambda}_{\mathrm{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \approx \frac{1}{n_q} \sum_{q=1}^{n_q} w(\mathbf{x}^{(q)}) \underline{\sigma}^2(\mathbf{x}^{(q)} \mid \mathbf{X}, \boldsymbol{\theta}) [1 - R(\mathbf{x}^{(q)}, \mathbf{x}_{new} \mid \boldsymbol{\theta})^{\lambda}] \tag{24}$$

Since the computaitonally expensive component of $\Phi^{\lambda}_{\mathrm{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ in Eq. (24) [the $\underline{\sigma}^2(\mathbf{x}^{(q)} \mid \mathbf{X}, \boldsymbol{\theta})$ contribution] is same for all candidate experimente examined, use of this approximaiton reduces the IMSE_w/IMSE burden to levels similar to those of MSE_w/MSE, which was the intended target. The corresponding DoE formulations will be distinguished by a superscript $^{\lambda}$.

The remaining question is, how good is the established approximation and if it can create significant vulnerabilities in erroneously identifying suboptimal new experiments? To provide an answer to this question, let's consider first the asymptotic case of $\lambda \to \infty$, where $A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda)$ becomes a Dirac delta function $\delta(\mathbf{x} - \mathbf{x}_{new})$. The approximated volume of influence is then equivalent to $\underline{\sigma}^2(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ which makes the approximated IMSE measure to be equivalent to the MSE. On the other hand, when λ is zero, $A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda) = 1$ and the approximated volume of influence is $\underline{\sigma}^2(\mathbf{x} \mid \mathbf{X}, \boldsymbol{\theta})$, therefore, is independent to $\mathbf{x}_{new}$. Performing IMSE DoE using zero λ becomes equivalent to a pure random sampling. Meanwhile, when $\lambda = 2$, the proposed approximation of Eq. (22) is exact under the assumption that no other training point exists in close proximity to $\mathbf{x}$ and $\mathbf{x}_{new}$ (correlation between $\mathbf{X}$ and either of these two points is numerically zero). A proof of this is included in Appendix C, where it is additionally shown that values of $\lambda \geq 2$ are guaranteed to provide higher accuracy approximations, and that as the number of experiments increases, larger values of λ are expected to yield higher accuracy. This discussion shows that, depending on the value of λ, the quality of the approximation of Eq. (22) will change, and that the appropriate value is impacted by the characteristics of the problem, such as (as discussed in Appendix C) relative proximity to other training points, GP hyper-parameters, and dimensionality of input. For this reason, rather than an a-priori selection, an adaptive selection of λ is proposed here and integrated within the adaptive DoE process, while $\lambda \geq 2$ is established as a lower bound constraint. This adaptive selection, detailed in the next section, incorporates different mitigate strategies, to avoid adoption of λ values that might lead to identification of lower quality experiments.

Figure 3 revisits the example shown in Figure 2 to illustrate some concepts related to the variance approximation discussed here. In all subplots of this figure, contours of different functions are presented and, across these subplots, the addition of two different experiments (i.e., two selections for $\mathbf{x}_{new}$, denoted as Points A or B) are examined to illustrate their respective impact on the variance reduction $V_d(\mathbf{x} \mid \mathbf{X}, \mathbf{x}_{new}, \boldsymbol{\theta})$. Part (a) of the figure presents contours for the weighted predictive variance [equivalent to part (f) in Figure 2]; part (b) shows the shape function $A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda)$ (using λ=2) for each of the candidate points; parts (c) and (d) show, respectively, the exact and approximate variance reductions after adding point A, whereas parts € and (f) replicate the presentation for point B. Comparing the quality of the approximation across the different candidate points, is it evident that when the new experiment is located close to multiple of existing experiments (Point B), the quality of the approximation is reduced. The contours for the actual variance reduction around Point B show greater concentration around it, indicating that λ>2 would have been a better option for the approximation. These trends verify the previous discussions, illustrating that the optimal selection of λ depends on the distribution of the existing experiments, stressing the importance of an adaptive selection for it. It is important to note, nevertheless, that the approximation in both instances (both points) examined in Figure 3 seems to be qualitatively consistent in terms of variation patterns, demonstrating the potential accuracy of the variance approximation for guiding the DoE.

Adaptive selection of λ

The adaptive selection of λ is accommodated by comparing the actual $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ and approximate $\Phi^{\lambda}_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ merit functions for some initial DoE iterations, till confidence on the chosen λ value is ascertained. Inevitably, this requires calculation of the exact $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ for these initial iterations. The adaptive selection is couched within the IMSE$_w$-*SE* algorithm, and involves consideration of the following issues/criteria:

(i) **[Optimization]** The optimization of λ is repeated at each DoE iteration step when a *credible* value has not yet been identified, or when a critical deviation in the λ value is expected due to the increased sample size.

(ii) **[Sampling variability and running average]** Differences of λ across the iterations of the DoE need to be considered, originating from the fact that the number of training points for defining $\mathbf{X}$ changes or that a stochastic search is utilized in the experiment identification. Some weighted averaging across the iterations should be established for the promoted λ to address the stochastic search features, whereas this averaging should give higher priority to recent iterations, to accommodate the natural variation of λ as more experiments are added.

(iii) **[Objective function]** Selection of an appropriate λ at each iteration requires definition of an appropriate objective function based on comparisons between $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ and $\Phi^{\lambda}_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$. Target here is to promote the same optimal experiment within the DoE, and not necessarily match the two merit functions. As such, the selection needs to focus on top rank candidate experiments. To accommodate (ii), the focus cannot be solely on the best experiment – setting as objective to facilitate only the same optimal $\mathbf{x}_{new}$ using $\Phi^{\lambda}_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ - since such an approach will lead to a lack of robustness, with λ optimally chosen only for the specific candidate samples utilized within the stochastic search.

(iv) **[Credibility check]** Convergence to assess credibility of λ requires criteria that examine both the variability of the promoted λ and, more importantly, the performance of the established approximation. For the latter, similar issues as identified in (iii) need to be considered for enhanced robustness.

(v) **[Re-optimization]** Even after convergence is achieved, the credibility of λ needs to be re-evaluated at some point, to account for the fact that the appropriate value of λ is dependent, as explained earlier, on the training point distribution and the GP characteristics. If reduced performance is identified, then calibration needs to be repeated.

Note that topics (ii) and (v) are the primary measures protecting against the use of a λ value that leads to suboptimal experiments. All these topics are discussed next, before a comprehensive algorithm is presented in the next section.

Initially, the objective function for the selection of λ at each iteration is discussed. For guiding this selection we focus on some subset of the best performing experiments within candidate set $\mathbf{X}^{candi} = \{\mathbf{x}_{new}^{(c)}\}_{c=1,\ldots,n_c}$ of the IMSE$_w$-*SE* algorithm. Note that if Step 2 of the algorithm is implemented then the subset is selected within the candidate set of the retained experiments with n_c replaced by n_r in all subsequent discussions. For the reasons explained in topic (iii) above, we are interested in not only matching the optimal solution, but in obtaining a consistent correspondence between the actual $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ or approximate $\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ criteria through a larger subset of them. This subset corresponds to the lowest p-percentile values of each of the objective functions and includes a total of n_p top-ranking candidate experiments, with $n_p = \lfloor n_c p \rfloor$ and $\lfloor . \rfloor$ representing the floor function. Let $\min^{n_p}[.]$ denote the n_pth smallest value of the set included within the brackets. Then, the subsets that includes the lowest p-percentile value of candidate experiments according to $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ corresponds to:

$$\mathbf{X}_p = \{\mathbf{x}_{new} \in \mathbf{X}^{candi} \mid \Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \leq \min^{n_p}[\{\Phi_{IMSE}(\mathbf{x}_{new}^{(c)} \mid \mathbf{X}, \boldsymbol{\theta})\}_{c=1,\ldots,n_c}\}] \tag{25}$$

The respective subset utilizing approximate objective function for a given λ value is:

$$\mathbf{X}_p^{\lambda} = \{\mathbf{x}_{new} \in \mathbf{X}^{candi} \mid \Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \leq \min^{n_p}[\{\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new}^{(c)} \mid \mathbf{X}, \boldsymbol{\theta})\}_{c=1,\ldots,n_c}\}]\} \tag{26}$$

If the approximation $\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ yields consistent ranking of experiments as $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$, which is the ideal scenario, then $\mathbf{X}_p$ and $\mathbf{X}_p^{\lambda}$ include identical experiments for any p value. The discrepancy between the two sets needs to be quantified with respect to their performance for the actual objective function $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$, whereas to establish a comparison across the entire set, the corresponding empirical cumulative distribution function (CDF) is utilized. For each of the sets, $\mathbf{X}_p$ and $\mathbf{X}_p^{\lambda}$, this CDF describes the distribution of the actual objective function value $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ within the set and is given, respectively, by:

$$F_p(\varphi) = \frac{1}{n_p} \sum_{\mathbf{x}_{new} \in \mathbf{X}_p} \mathbf{1}[\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) < \varphi] \tag{27}$$

$$F_p^{\lambda}(\varphi) = \frac{1}{n_p} \sum_{\mathbf{x}_{new} \in \mathbf{X}_p^{\lambda}} \mathbf{1}[\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) < \varphi] \tag{28}$$

where $\mathbf{1}[.]$ is the indicator function, which is 1 if the quantity inside the brackets holds, else it is zero. Note that $F_p^{\lambda}(\varphi)$ may be viewed as projection of the $\mathbf{X}_p^{\lambda}$ set performance on the space of $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$, whereas $F_p(\varphi) \geq F_p^{\lambda}(\varphi)$ always holds since set $\mathbf{X}_p$ by design includes the lowest possible objective function values. The selection of λ can be based on the minimization of the gap between the two CDFs:

$$\lambda^* = \arg\min_{\lambda} \int_{\varphi_{\min}}^{\varphi_{\max}} \left(F_p(\varphi) - F_p^{\lambda}(\varphi) \right) d\varphi \tag{29}$$

where $\varphi_{\min}$ corresponds to the smallest $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ value within set $\mathbf{X}_p$ and $\varphi_{\max}$ to the largest $\Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ value within set $\mathbf{X}_p^{\lambda}$. Figure 4 illustrates these concepts. Note that optimization of Eq. (29) can be performed by a discrete search, as will be detailed later.

As discussed in topic (ii) above, the optimal λ values identified through Eq. (29) will have variability across the DoE iterations, and some averaging needs to be performed to address the influence of the stochastic search features, while also incorporating weights to give higher priority to recent estimates, to better capture the underlying trends as the number of experiments increases. Let λ_i^*. denote the optimal λ value obtained at *i*-th iteration, then using linearly decreasing weights within a window of *L* iterations, the weighted average value of λ at *k*-th iteration is:

$$\overline{\lambda}_k = \frac{\sum_{i=k-L+1}^{k} \gamma_i \lambda_i^*}{\sum_{i=k-L+1}^{k} \gamma_i} \tag{30}$$

with weights given by:

$$\gamma_i = \frac{i - (k - L)}{L} \tag{31}$$

The λ given by Eq. (30) represents the estimate for the variance approximation exponent at the current DoE iteration. The quality of this approximation, to assess convergence according to topic (iv) above, is

expressed through two different *credibility* criteria. The first one examines the coefficient of variation of $\overline{\lambda}_k$ within the window L, given by:

$$cv_{\lambda}^{k} = \frac{\sqrt{\frac{1}{L}\sum_{i=k-L+1}^{k}\left(\overline{\lambda}_i - \frac{1}{L}\sum_{j=k-L+1}^{k}\overline{\lambda}_j\right)^2}}{\frac{1}{L}\sum_{i=k-L+1}^{k}\overline{\lambda}_i} \tag{32}$$

If this coefficient of variation is below some target threshold ε_{conv} then the estimate of λ may be considered stable, indicating convergence. Convergence of λ is not assessed till at least L estimates of it are available to accommodate the weighting average calculation. The second credibility criterion focuses on the performance of the approximated $IMSE_w$ offered through $\overline{\lambda}_k$. Similarly to the definition of the objective function for selection of λ, the comparison is established across some subset of top-ranked candidate experiments. If the top s-percentile experiments are utilized, corresponding to total of $n_s = \lfloor n_c s \rfloor$ experiments, then the subsets of interest $\mathbf{X}_s$ and $\mathbf{X}_s^{\overline{\lambda}_k}$ are given by Eqs. (25) and (26), respectively, using n_s instead of n_p. The performance of $\overline{\lambda}_k$ for assessing convergence examines how suboptimal the solutions within set $\mathbf{X}_s^{\overline{\lambda}_k}$ are compared to the solutions in set $\mathbf{X}_s$. The worst performance within each set is considered. The gap between the two sets, since values in $\mathbf{X}_s$ are guaranteed to outperform values in $\mathbf{X}_s^{\overline{\lambda}_k}$, is normalized by the average performance in set $\mathbf{X}_s$, leading to convergence criterion:

$$g(\overline{\lambda}_k) = \frac{\max\limits_{\mathbf{x}_{new}\in\mathbf{X}_s^{\overline{\lambda}_k}}[\Phi_{IMSE}(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})] - \max\limits_{\mathbf{x}_{new}\in\mathbf{X}_s}[\Phi_{IMSE}(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})]}{\frac{1}{n_s}\sum\limits_{\mathbf{x}_{new}\in\mathbf{X}_s}\Phi_{IMSE}(\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})} \tag{33}$$

When the performance criterion is below some target threshold ε_{perf} then the estimate of λ may be considered to provide an accurate approximation of the actual objective function.

Once the criteria related to the quantities in Eqs. (32) and (33) are satisfied, convergence has been established and the $\overline{\lambda}_k$ value may be considered to provide a credible approximation to the $IMSE_w$ estimation.

As discussed in topic (v) above, the most appropriate value of λ is expected to change across the DoE iterations, and for this reason, the identification of an optimal λ according to Eq. (29) and the estimation of performance according to Eq. (33) for assessing convergence should be repeated after N_λ iterations. If λ is no longer assessed as credible, then calibration of λ is repeated in each iteration till convergence is re-established.

Algorithm for adaptive selection of λ

Combining the concepts discussed in the previous section the DoE algorithm utilizing an adaptive approximation for IMSE$_w$ is established. This algorithm will be denoted IMSE$_w$-*SA* (Sequential Approximation) herein, and is summarized in Figure 5. A flag/counter i_f is used to assess convergence for the value of λ, with initial value set equal to 0. Algorithm also requires selection of tuning parameters ε_{conv}, ε_{perf}, L, N_λ, s (or equivalently n_s), p (or equivalently n_p) for the adaptive λ selection. Steps 0-2 of the algorithm are identical to IMSE$_w$-*SE* presented earlier, while the remaining are modified as follows:

Step 3.1 [MCI integration]: Generate n_q samples $\{\mathbf{x}^{(q)}\}_{q=1,\ldots,n_q}$ following a uniform distribution in X^d to be used for the MCI. Estimate $w(\mathbf{x}^{(q)})$ for all these samples combining Eqs. (17) and (14), and calculate variance $\underline{\sigma}^2(\mathbf{x}^{(q)} \mid \mathbf{X}, \boldsymbol{\theta})$ using Eq.(47).

Step 3.2 [Correlation function for candidate experiments and integration points]: For each candidate (or retained from Step 2) experiment, estimate the correlation between it and each of the integration points, $R(\mathbf{x}^{(q)}, \mathbf{x}_{new}^{(c)} \mid \boldsymbol{\theta})$, for the chosen GP kernel.

Step 4.1 [Calibrate λ]: If $i_f \in [1, N_\lambda]$ then proceed to Step 4.5. Else if $i_f = 0$, perform Step 4 of the original IMSE$_w$-*SE* algorithm to estimate $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)})$ for each $\mathbf{x}_{new}^{(c)}$, and then update the promoted value of λ as follows. Establish a range of candidate λ values (with λ>2 as discussed earlier) and estimate $\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new}^{(c)})$ for each $\mathbf{x}_{new}^{(c)}$ of the candidate (or retained experiments) using Eq. (24). Note that the only component impacted by λ in this equation is $R(\mathbf{x}^{(q)}, \mathbf{x}_{new}^{(c)} \mid \boldsymbol{\theta})^{\lambda}$, and so calculation even for large number

of candidates λ values can be efficiently performed. Use $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)})$ and $\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new}^{(c)})$ to define sets $\mathbf{X}_p$ [Eq. (25)] and $\mathbf{X}_p^{\lambda}$ [Eq. (26)] (for each λ) and CDF approximations $F_p(\varphi)$ [Eq. (27)] and $F_p^{\lambda}(\varphi)$ [Eq. (28)] (for each λ), and perform optimization of Eq. (29). The latter is established using a discrete search, selecting the candidate λ value that yields the smallest gap between $F_p(\varphi)$ and $F_p^{\lambda}(\varphi)$. Denote the identified λ value as λ_k^*.

Step 4.2 [Moving average λ estimate]: Calculate the moving average value of $\bar{\lambda}_k$ using Eq. (30)

Step 4.3 [Assessing stability of λ estimate]: Estimate the coefficient of variation cv_{λ}^{k} using Eq. (32). If $cv_{\lambda}^{k} > \varepsilon_{conv}$ then the estimate of λ is not yet stable. Proceed to Step 5.

Step 4.4 [Assessing quality of IMSE$_w$ approximation]: Estimate $\Phi_{\text{IMSE}}^{\bar{\lambda}_k}(\mathbf{x}_{new}^{(c)})$ for each $\mathbf{x}_{new}^{(c)}$ of the candidate (or retained experiments) using Eq. (24). Use $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)})$ (from Step 4.1) and $\Phi_{\text{IMSE}}^{\bar{\lambda}_k}(\mathbf{x}_{new}^{(c)})$ to define the sets $\mathbf{X}_s$ [modified Eq. (25) for n_p=n_s] and $\mathbf{X}_s^{\bar{\lambda}_k}$ [Eq. (26) for n_p=n_s]. Estimate the indicator of the quality of the IMSE$_w$ approximation, $g(\bar{\lambda}_k)$, using Eq. (33). If $g(\bar{\lambda}_k) > \varepsilon_{perf}$ then estimate of λ is not credible; set $i_f = 0$. If $g(\bar{\lambda}_k) \le \varepsilon_{perf}$ convergence to a credible $\bar{\lambda}_k$ has been established; set $i_f = 1$ and utilize $\bar{\lambda}_k$ for the IMSE$_w$ approximation that will be used in the next N_λ DoE iterations. Proceed to Step 5.

Step 4.4 [IMSE$_w$ approximation]: Using the $\bar{\lambda}_k$ identified as credible in the previous DoE iteration in which the λ calibration was updated, estimate $\Phi_{\text{IMSE}}^{\bar{\lambda}_k}(\mathbf{x}_{new}^{(c)})$ for each $\mathbf{x}_{new}^{(c)}$ of the candidate (or retained experiments) using Eq. (24). Set $i_f = i_f + 1$ to keep track of the number of iterations since the previous λ calibration.

Step 5 [Final selection] If Step 4.1 was performed, select as new experiment the one that provides the minimum value for $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)})$. Else, select as new experiment the one that provides the minimum value for $\Phi_{\text{IMSE}}^{\lambda}(\mathbf{x}_{new}^{(c)})$.

Note that dependence on $\mathbf{X},\boldsymbol{\theta}$ were omitted in the notations of Φ_{IMSE}, $\Phi_{\text{IMSE}}^{\lambda}$ and $\Phi_{\text{IMSE}}^{\bar{\lambda}_k}$ within the algorithm description for brevity. As also shown in Figure 5, the updated Step 4 has two alternative paths. If convergence to a credible λ value has not yet been established, or if the maximum number of steps N_λ for the use of convergent values are exceeded, then Steps 4.1-4.4 are performed, with Step 4.1 representing the only computational demanding one within the IMSE$_w$-*SA* algorithm, since it requires the calculation of the exact IMSE$_w$. Alternatively, the IMSE$_w$ approximation is directly used in Step 4.5, adopting the previously converged λ value, providing a dramatic computational reduction, as discussed earlier, for the DoE. Note that if λ is chosen a priori, rather than being estimated adaptively, then Step 4 of the IMSE$_w$-*SA* algorithm would always correspond to Step 4.5. The algorithm involves the following tuning parameters as summarized earlier: ε_{conv}, ε_{perf}, L, N_λ, s, p. The values used in the case studies discussed in this paper are ε_{conv}=10%, ε_{perf}=1%, L=10, N_λ=25, s=10% and p=5%, with additional constraints that value of n_p and n_s correspond to at least 10 samples. Only small sensitivity was identified to these parameter values in terms of performance, though, as will be discussed in the illustrative examples section these parameters have an effect on convergence to a credible λ and therefore on computational efficiency.

Adaptive DoE using multi-criteria search

The second advancement established in this paper for the adaptive DoE is multi-criteria search strategy that established an alternative balance between the variance (exploration) and bias (exploitation) selection criteria while simultaneously promoting a seamless identification of batch experiments. Foundation of the strategy is to consider the bias and variance criteria as separate objectives and formulate a bi-objective DoE. This problem does not necessarily have a single solution, and so the identification provides the Pareto optimal experiments. Let $\Phi_{\text{var}}(\mathbf{x}_{new})$ denote the merit function related to the GP variance and $\Phi_{\text{bias}}(\mathbf{x}_{new})$ the merit

function related to the GP bias and assume without a loss of generality that optimality is expressed with respect to minimization. Note that both these functions have dependence on $\mathbf{X}$ and $\boldsymbol{\theta}$ but for notational simplicity this dependence will not be explicitly denoted. The set of Pareto optimal experiments, denoted herein as $\mathbf{X}_{PF}$, is obtained though the multi-objective optimization problem:

$$\mathbf{X}_{PF} = \underset{\mathbf{x}_{new} \in X^d}{\arg\min} \{\Phi_{\text{var}}(\mathbf{x}_{new}), \Phi_{\text{bias}}(\mathbf{x}_{new})\} \tag{34}$$

A candidate experiment is termed dominant and belongs in Pareto set $\mathbf{X}_{PF}$ if there is no experiment that simultaneously improves both objectives. The representation of the set $\mathbf{X}_{PF}$ in the merit function space $\{\Phi_{\text{var}}(\mathbf{x}_{new}), \Phi_{\text{bias}}(\mathbf{x}_{new})\}$ corresponds to the Pareto front. An illustration is shown in Figure 6(b), The numerical details and computational complexity for the identification of the Pareto set through Eq. (34) will be discussed later, when the overall multi-objective algorithm is discussed.

Though any bias and variance–based functions can be used in the bi-objective identification, the discussions here are couched to utilizing the IMSE, given by Eq. (7), and LOOCV squared error, given by Eq. (14), leading to:

$$\Phi_{\text{var}}(\mathbf{x}_{new}) = \Phi_{\text{IMSE}}(\mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \tag{35}$$

$$\Phi_{\text{bias}}(\mathbf{x}_{new}) = -(\tilde{e}^{cv}(\mathbf{x}_{new}))^2 \tag{36}$$

where, note, that for the IMSE function the weights $w(\mathbf{x})$ do not incorporate any information for the metamodel bias since this is explicitly considered in the $\Phi_{\text{bias}}(\mathbf{x}_{new})$ objective.

The set $\mathbf{X}_{PF}$ identified through Eq. (34) represents the candidate set for a batch selection of the next experiments. The direct use of this entire set as new experiments is not recommended, though, for two reasons: (i) first because the number of experiments in this set, denoted by N_{PF} herein, might be larger than the number of batch experiments desired, denoted N_B herein; (ii) second because the $\Phi_{\text{var}}(\mathbf{x}_{new})$ has not considered the simultaneous addition of all the new experiments, since it separately examined the impact on the GP variance of each experiment. Issue (ii) is especially important. It means that Pareto set may contain experiments in close-proximity to one another, each substantially benefiting the other(s) if/when added in the pool of available

experiments, and making the addition of all of them potentially redundant. To address this vulnerability, while also accommodating topic (i) above, a sequential search within the Pareto set is promoted: a single experiment is chosen from this set, the GP variance is updated considering the addition of the new experiment, the variance-based merit function for all remaining experiments in the Pareto set is adjusted based on this update, and then the next experiment is identified. Unfortunately, a similar update cannot be established for the bias merit function. To mathematically describe this recursive addition of experiments, we will herein denote by $\mathbf{x}_{new}^{(j)}$ the experiment (member of $\mathbf{X}_{PF}$) identified in the *j*th iteration, by $\mathbf{X}_{PF}^{\sim j}$ the remaining Pareto set excluding all experiments identified up to the *j*th iteration, by $\mathbf{X}^{j}$ the set of available experiments obtained by adding to $\mathbf{X}$ the selected experiments up to the *j*th iteration, and by $\Phi_{\text{var}}^{(j)}$ the updated variance-based objective using $\mathbf{X}^{j}$.

For selecting the single experiment that best balances the two objectives at each iteration alternative criteria exist (Kim, I. Y. and De Weck 2005, Gunantara 2018), with one of the most popular ones, and the one chosen here, being the selection of the point with minimum distance to the Utopia point in the Pareto front. The latter corresponds to the unattainable point beyond the Pareto front that yields the minimum of both objective functions, as also shown in Figure 6(b). Establishing an appropriate normalization for each merit function with respect to its scale within the Pareto front, the objective function for selecting the single Pareto optimal design at the *j*+1 iteration is:

$$F_{PF}^{(j)}(\mathbf{x}_{\text{new}}) = \left[\frac{\Phi_{\text{var}}^{(j)}(\mathbf{x}_{\text{new}}) - \min_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{var}}^{(j)}(\mathbf{x}_{\text{new}})\}}{\max_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{var}}^{(j)}(\mathbf{x}_{\text{new}})\} - \min_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{var}}^{(j)}(\mathbf{x}_{\text{new}})\}}\right]^2 + \left[\frac{\Phi_{\text{bias}}(\mathbf{x}_{\text{new}}) - \min_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{bias}}(\mathbf{x}_{\text{new}})\}}{\max_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{bias}}(\mathbf{x}_{\text{new}})\} - \min_{\mathbf{X}_{PF}^{\sim j}}\{\Phi_{\text{bias}}(\mathbf{x}_{\text{new}})\}}\right]^2 \tag{37}$$

where $\max_{\mathbf{X}_{PF}^{\sim j}}\{.\}$ and $\min_{\mathbf{X}_{PF}^{\sim j}}\{.\}$ correspond to the maximum and minimum, respectively, of the quantity in the brackets across the remaining Pareto set $\mathbf{X}_{PF}^{\sim j}$. The next experiment is identified by:

$$\mathbf{x}_{new}^{(j+1)} = \underset{\mathbf{x}_{\text{new}} \in \mathbf{X}_{PF}^{\sim j}}{\arg\min}\, F_{PF}^{(j)}(\mathbf{x}_{\text{new}}) \tag{38}$$

Once this experiment is identified, the updated GP variance $\sigma^2(\mathbf{x} \mid \mathbf{X}^{j}, \mathbf{x}_{new}^{(j+1)}, \boldsymbol{\theta})$ by the addition of this single new experiment can be obtained by Eq. (11) and the efficient procedure discussed in Appendix B. Use of $\sigma^2(\mathbf{x} \mid \mathbf{X}^{j}, \mathbf{x}_{new}^{(j+1)}, \boldsymbol{\theta})$ then leads to the updated estimate of $\Phi_{\text{var}}^{(j+1)}(\mathbf{x}_{\text{new}})$ for the next iteration.

To accommodate the recursive identification, if the number of Pareto points is smaller than the target ($N_{PF} < N_B$) then the Pareto set can be expanded by considering dominant designs of higher rank (Deb et al. 2002). This is established by identifying dominant designs among the remaining candidate experiments, removing the ones already identified as belonging to the Pareto set.

Figure 7 revisits the example used previously in Figure 2 and illustrates aspects of the multi-objective DoE for selection of a batch of 5 experiments, focusing on how the specifics of the bias interpolation approach affect the formulation. The two different options for defining $w(\mathbf{x})$ discussed previously are examined, the NN interpolation given by Eq. (15) and the exponential decay kernel smoothing given by Eq. (16), abbreviated as E in this figure. The dominated solutions, the identified Pareto front and the batch of 5 experiments are shown in the objective function space in parts (a) and (b) of this figure for the NN and E implementations, respectively. Part (c) and (d) show the contours of $\Phi_{\text{bias}}(\mathbf{x})$ as well as the Pareto solutions and chosen batch of experiments, for the same cases. Note that for NN interpolation, these contours represent, equivalently, the Voronoi cells for the training points. The NN interpolation leads to identification of a single dominant solution in each of the Voronoi cells, leading to a sparsely populated Pareto front and, ultimately, to a final batch of experiments that is more scattered across the entire domain, and is not concentrated in regions with higher bias (compare to the accuracy of the GP predictions depicted in Figure 2 earlier). Equivalently, this means that the DoE will place greater emphasis on exploration rather than exploitation. In contrast, the kernel smoothing using the exponential decay function identifies a larger number of dominant solutions, clustered at high bias and/or variance locations, as evident in part (d) of Figure 7. The final batch of experiments are located close to domains with higher bias, depicting a stronger exploitation tendency.

Finally, the multi-objective DoE using IMSE as variance measure, is implemented through the following algorithm denoted herein as IMSE-*MB* (*Multi-objective Batch*). To circumvent the well-known higher computational burden associated with solving multi-objective optimization problems, algorithm adopts a random search approach, similar to the one established for IMSE$_w$-SE. Specifically, Steps 0-1 and 3 of the

algorithm are identical to IMSE$_w$-SE, while Step 2 is removed since the behavior across the entire Pareto set is warranted. The remaining steps are modified as follows:

Step 4.1 [Calculation of variance-based merit function]: For each candidate experiment perform Step 4 of the original IMSE$_w$-*SE* utilizing $w(\mathbf{x})=1$ to obtain $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)}\,|\,\mathbf{X},\boldsymbol{\theta})$. Set this equal to variance objective $\Phi_{\text{var}}^{0}(\mathbf{x}_{new}^{(c)})$ and set experiment counter j=0.

Step 4.2 [Calculation of bias objective function]: For each candidate estimate $(\tilde{e}^{cv}(\mathbf{x}))^2$ using Eq. (14) to obtain corresponding bias objective $\Phi_{\text{bias}}(\mathbf{x}_{new}^{(c)})$ as in Eq. (36).

Step 4.3 [Pareto set selection]: Identify the Pareto set $\mathbf{X}_{PF}$ of dominant designs within $\mathbf{X}^{candi}$ using merit functions from Steps 4.1. and 4.2 . If number of experiments in this set N_{PF} is not sufficient (does not exceed target N_B) expand selection to consider higher rank dominance till $\mathbf{X}_{PF}$ includes a sufficient number of experiments. This defined ultimately set $\mathbf{X}_{PF}^{\sim 0}$

Step 5.1 [Selection of next experiment]: For all points $\mathbf{x}_{\text{new}}$ in set $\mathbf{X}_{PF}^{\sim j}$ estimate distance from utopia point $F_{PF}^{(j)}(\mathbf{x}_{\text{new}})$ using Eq. (37) and merit functions $\Phi_{\text{bias}}(\mathbf{x}_{new})$ and $\Phi_{\text{var}}^{j}(\mathbf{x}_{new}^{(c)})$. Select the next experiment $\mathbf{x}_{new}^{(j+1)}$ though Eq. (38).

Step 5.2 [Stopping criteria for batch selection]: If j+1=N_B stop, else proceed to Step 5.3.

Step 5.3 [Updating of sets and matrices]: Remove $\mathbf{x}_{new}^{(j+1)}$ from $\mathbf{X}_{PF}^{\sim j}$ to obtain $\mathbf{X}_{PF}^{\sim j+1}$ and add it to $\mathbf{X}^{j}$ to obtain $\mathbf{X}^{j+1}$.Use $\mathbf{X}=\mathbf{X}^{j+1}$ and $\mathbf{x}_{new}=\mathbf{x}_{new}^{(j+1)}$ in Eqs. (8)-(10) and Eq. (49) to obtain, respectively, the updated correlation vector $\mathbf{r}(\mathbf{x}\,|\,\mathbf{X}^{j+1},\boldsymbol{\theta})$, basis function matrix $\mathbf{F}(\mathbf{X}^{j+1})$, correlation matrix $\mathbf{R}(\mathbf{X}^{j+1}\,|\,\boldsymbol{\theta})$ and Cholesky factorization $\mathbf{L}(\mathbf{X}^{j+1}\,|\,\boldsymbol{\theta})$ and estimate matrix $\mathbf{M}_R(\mathbf{X}^{j+1}\,|\,\boldsymbol{\theta})$ related to the QR decomposition of $\mathbf{L}^{-1}(\mathbf{X}^{j+1}\,|\,\boldsymbol{\theta})\mathbf{F}(\mathbf{X}^{j+1})$.

Step 5.4 [Updating of variance based merit function]: Use the updated quantities from Step 5.3 to calculate $\underline{\sigma}^2(\mathbf{x}^{(q)}\,|\,\mathbf{X}^{j+1},\mathbf{x}_{new}^{(c)},\boldsymbol{\theta})$ using Eq. (48) for each of the integration points and each $\mathbf{x}_{new}^{(c)}$ in $\mathbf{X}_{PF}^{\sim j+1}$

. Using this variance obtain the IMSE objective function $\Phi_{\text{IMSE}}(\mathbf{x}_{new}^{(c)} \mid \mathbf{X}^{j+1}, \boldsymbol{\theta})$ through MCI using Eq. (12) with . Set this equal to variance objective and set experiment counter $j=j+1$. Proceed to Step 5.1.

The algorithm is summarized in Figure 8. Note that since the updating and search in Step 5 is constrained always within the previously retained Pareto set , the computational demanding step of this algorithm is the original estimation of IMSE for candidate experiments (Step 4.1). This reduces the overall computational burden to be similar to the single-objective optimization implementations discussed earlier.

Illustrative examples

Case study examples description

The proposed DoE advancements from the previous two sections are showcased using four numerical examples. The input dimension in these examples ranges from n_x=2 to n_x=9, and each example poses different challenges with respect to the nonlinearity of the input-to-output mapping. Three of the examples are taken directly from (Kyprioti et al. 2020), while the other one is first introduced here. The first two examples correspond to analytic benchmark functions: the two-dimensional cosine weighted Gaussian mixture function (n_x=2) inspired by (Jiang et al. 2015), referenced herein as 'WtGMix', and the six-dimensional Harman function (n_x=6) proposed by (Dixon and Szegö 1978), referenced herein as 'Hartman 6'. Details for these examples and for the input domain considered X^d are included in (Kyprioti et al. 2020). Note that the WtGMix corresponds to the function that was used earlier (Figures 2, 3, and 7), to illustrate some of the DoE concepts. The remaining two examples represent practical applications from the earthquake engineering field. The first one is four-dimensional function (n_x=4) corresponding to the standard deviation of the normalized base shear of a single degree-of-freedom (SDoF) oscillator exposed to stationary stochastic seismic excitation modeled by the Kanai-Tajimi power spectrum (Kanai 1957, Tajimi 1960). The output function in this case is expressed as (Lutes and Sarkani 2004):

$$z = \omega_s^{\,2}\left(S_o \int_{-\infty}^{\infty} \frac{1+4\zeta_g^{\,2}\left(\omega/\omega_g\right)^2}{\left(1-(\omega/\omega_g)^2\right)^2+4\zeta_g^{\,2}\left(\omega/\omega_g\right)^2} \frac{(1/\omega_s)^4}{\left(1-(\omega/\omega_g)^2\right)^2+4\zeta_s^{\,2}(\omega/\omega_s)^2} d\omega\right)^{1/2} \tag{39}$$

where the four input parameters represent the frequency (ω_s) and damping (ζ_s) of the SDoF oscillator, and the frequency (ω_g) and damping (ζ_g) of the Kanai-Tajimi spectrum used to describe the earthquake acceleration input. The range examined for these values, defining X^d, is $\omega_g \in [2,6]\pi$ rad/sec , $\zeta_g \in [0.1,0.6]$, $\omega_s \in [1,13.5]\pi$ rad/sec and $\zeta_s \in [0.02,0.06]$. This example will be referenced herein as 'SDoF-KT'. The last example, termed 'Isolation', corresponds to the displacement of the base of a three-story base isolated structure exposed to non-stationary near-fault seismic excitation (Kyprioti et al. 2020). Numerical details for the isolated structure and the excitation are included in (Jia and Taflanidis 2014). This case corresponds to a to nine-dimensional example (n_x=9) with inputs including five structural and isolation parameters and four additional parameters related to the earthquake excitation. Details for the definition of the input, including ranges considered, are included in (Kyprioti et al. 2020).

Surrogate modeling and DoE details

For the GP, a constant basis function $f(\mathbf{x})$=[1] and a generalized exponential correlation function (Kyprioti et al. 2020, Rasmussen and Nickisch 2010) are chosen, while hyper-parameter optimization is performed using maximum likelihood estimation. The metamodel accuracy is evaluated using a test-sample validation. This is preferred to the cross-validation setting discussed earlier, as it avoids dependence of the DoE validation on the specific set of simulation experiments chosen (Kleijnen and Van Beers 2022, Zhang and Taflanidis 2018). Since objective of the validation is to compare across the different DoE strategies, the test-sample implementation is necessary for accommodating consistency. Using a set of N_t points uniformly distributed in X^d, the normalized root mean squared error validation metric in this case is given by:

$$NRMSE_{test} = \frac{\sqrt{\sum_{k=1}^{N_t} (z^{(k)} - \tilde{z}(\mathbf{x}^{(k)} \mid \mathbf{D}, \Theta))^2 / N_t}}{\max_k \{z^{(k)}\} - \min_k \{z^{(k)}\}} \tag{40}$$

where $\{\mathbf{x}^{(k)}, z^{(k)}; k = 1,..., N_t\}$ represents the input-output test-sample set, $\max_k \{z^{(k)}\}$ and $\min_k \{z^{(k)}\}$ denote the maximum and minimum values of the response test samples, and the value of N_t is chosen here as 5000.

All examined DoE strategies start with n_{init} experiments, obtained through Latin Hypercube space-filing sampling in X^d, and sequentially add experiments, either one at-a-time or in batches of n_b experiments, till the

771 desired number of experiments n_{fin} is reached. Following the recommendations in (Kyprioti et al. 2020), the
772 initial n_{init} and final n_{fin} experiment sizes are chosen so that the metamodel starts with low accuracy and becomes
773 highly accurate at the end. The details for each example are: WtGMix n_{init}= 15 and n_{fin}= 125, SDoF-KT n_{init}=
774 30 and n_{fin}= 220, Hartman6 n_{init}= 30 and n_{fin}= 200, Isolation n_{init}= 25 and n_{fin}= 250. Detailed rational for these
775 choices is included in (Kyprioti et al. 2020).

776 The N_q for the MCI is selected to establish a small coefficient of variation for each example examined,
777 chosen as N_q=5000 for all problems, while the number of candidate experiments n_c is chosen proportional to
778 the dimensionality of each example, with goal to balance between computational efficiency and adequate
779 exploration of the X^d domain. The details for each example are: WtMixG n_c= 1000, SDoF-KT n_c=3000,
780 Hartman6 n_c= 5000, and Isolation n_c= 7000. The DoE implementation is repeated 30 times for each case, using
781 different samples for the initialization and the DoE identification. Results will be reported typically for the
782 average DoE accuracy across these trials, though some variability trends will be also briefly discussed. For
783 each trial, the same samples have been utilized across all DoE strategies to facilitate a consistent comparison.

784 The DoE variants examined are reviewed in Table 1. These correspond to: (i) the standard IMSE and MSE
785 implementations; (ii) previous formulations to incorporate bias weights (Kyprioti et al. 2020) that serve as the
786 foundation of the proposed here advances,; (iii) as well as the advanced implementations established in this
787 manuscript. For accommodating the incorporation of the bias two different interpolation functions are
788 considered, the NN interpolation given by Eq. (15) and the exponential decay kernel smoothing given by Eq.
789 (16). These will be distinguished, respectively, by abbreviation NN and E, and will be denoted in parenthesis
790 after the name of the DoE variant, when needed. For example, $IMSE_w$-(V) denotes the $IMSE_w$ implementation
791 with bias estimated using NN interpolation (i.e., Voronoi tessellation) and $IMSE_w$-(E) denotes the alternative
792 implementation using exponential decay smoothing. For $IMSE_w$, implementation for batch selection of
793 experiments is also considered. In this case the adjustment using different ρ values to balance between
794 exploration and exploitation is also examined, to accommodate a comparison to the multi-criteria DoE
795 implementation. Following the recommendations in (Kyprioti et al. 2020) the value of ρ is modified across
796 each batch of 5 experiments, with the first three corresponding to ρ=1 (combination of exploration and

797 exploitation) and the remaining two to ρ=0 (pure exploration). The DoE variants corresponding to batch IMSE$_w$
798 will be denoted as *B*-IMSE$_w$ and B^{ρ}-IMSE$_w$ for the implementation without or with the ρ adjustments,
799 respectively. For the IMSE$_w$ approximation employed through the IMSE$_w$-*SA* algorithm, two different
800 formulations are examined, corresponding to the adaptive selection of λ or the use of a default value of $\lambda = 2n_x$.
801 This default value for λ has been chosen after examining performance for multiple case studies, including the
802 ones presented in this manuscript. These variants will be denoted as A_d-IMSE$_w$ and A_c-IMSE$_w$ for the
803 implementations with and without adaptive selection of λ, respectively. Finally, the multi-objective IMSE$_w$
804 implementation, employed through the IMSE$_w$-*MB* algorithm, will be denoted as *MB*-IMSE$_w$.

805 Note that variants (MSE$_w$, IMSE$_w$, *B*-IMSE$_w$ and B^{ρ}-IMSE$_w$) represent the aforementioned previously
806 established formulations (Kyprioti et al. 2020), and should serve as the baseline reference against which the
807 efficiency of the proposed here advances should be compared to. Within this context, the non-weighted variants
808 (MSE, IMSE) represent classical baseline approaches for assessing the benefits of incorporating the bias
809 weights.

810 ***Results and discussion***

811 Results are separately presented for the DoE variants corresponding to sequential experiment selection and
812 batch experiment selection. These two sets of results accommodate, respectively, evaluation of the proposed
813 advancements related to the IMSE$_w$ approximation and the multi-criteria DoE formulation. Results are
814 primarily presented through the variation of the metamodel accuracy with respect to the number of experiments
815 *n*. Emphasis is placed on the average performance across the 30 trials.

816 Figures 9 and 10 present results for the variants corresponding to sequential experiment selection for the
817 two different choices for the bias interpolation function, (NN) in Figure 9 and (E) in Figure 10. Figure 11
818 presents details for the computational savings established through the A_d-IMSE$_w$ formulation, defined as the
819 average (across the 30 trials) acceptance ratio of the λ approximation, i.e. the frequency of omitting the exact
820 IMSE$_w$ evaluation in the IMSE$_w$-*SA* algorithm (omitting Step 4.1). This is presented as function of the DoE
821 iteration, which is equal to n-n_{init} (total experiments minus initial experiments). The overall efficiency,
822 corresponding to the mean value of the efficiency across the DoE iterations is also reported in this figure. The

first 10 iterations are omitted in estimating the overall efficiency, since convergence to a credible λ is not possible before that point based on the tuning parameter selections presented previously. Figure 12 investigates further the impact of tuning parameters for the for A_d-IMSE$_w$ performance, Figure 13 shows the variation of the average (over the 30 trials) value of $\bar{\lambda}_k$ for the SDoF-KT example. Figure 14 presents results for the DoE variants corresponding to batch-selection of experiments. Figure 15 facilitates a comparison of the different interpolation functions, gathering some of the results presented earlier (in Figures 9, 10 and 14) in one figure. Finally, the variability of the DoE performance across the 30 trials is investigated in Figure 16, showing boxplots of the metamodel accuracy when n_{fin} has been reached, for specific DoE variants of interest.

Comparing first the standard implementations of MSE, MSE$_w$, IMSE, and IMSE$_w$ in Figure 9 and Figure 10, earlier discussions and past literature results are easily verified: IMSE outperforms MSE, in some applications by a very large margin, whereas inclusion of bias weight can improve performance. The benefit of introducing the bias weight is more evident in the examples with significant localized nonlinearities (WtGMix and Hartman6) (Kyprioti et al. 2020). One exception is Figure 9(c) for which IMSE$_w$ and IMSE have similar performance. This is possibly due to the mild local nonlinearity of the SDoF-KT application.

The comparisons across Figures 9 and 10 showcase clearly the preference of IMSE-based DoE schemes over the MSE-based ones. As also discussed earlier, this performance improvement comes at the expense of a larger computational burden. The approximations for IMSE establish a balance between the DoE quality of IMSE-based schemes (Figures 9 and 10) and the computational efficiency of MSE-based schemes (Figure 11). Both A_d-IMSE$_w$ and A_c-IMSE$_w$ outperform the baseline MSE$_w$ by a significant margin and provide similar DoE quality as IMSE. Identical patterns hold for the implementations without the bias weight, i.e. comparing A_d-IMSE and A_c-IMSE to MSE in these figures, though the emphasis in the discussions herein is on the bias-weighted schemes since they represent the preferred adaptive DoE formulation.

Especially the improvement of A_c-IMSE$_w$ over MSE$_w$ is very noteworthy since, recall, these DoE formulations have practically identical computational burden. This clearly showcases the advantages the proposed A_c-IMSE$_w$ scheme can offer over existing alternative formulations. As expected, the adaptive λ selection (A_d-IMSE$_w$) outperforms the formulation with a prescribed λ (A_c-IMSE$_w$), in some instances with a

noticeable margin. This comes of course at the expense of a larger computational burden, as shown in Figure 11. Results in this figure show that in earlier DoE iterations the value of λ is not deemed credible (acceptance ratio of the λ approximation has lower values) whereas even in later iterations re-evaluation of the approximation quality or divergence from a credible λ are observed (acceptance ratio of the λ approximation lower than 1). Recall in all such instances estimation of the exact IMSE_w estimation is warranted. Details are, as expected, different across the case study examples, with computational savings in range of 0.45~0.90. This corresponds to a significant improvement (10~55% of the original computations needed only) when compared to the IMSE_w formulation, but do represent an additional computational burden when compared against the A_c-IMSE_w formulation which tends to give, as mentioned earlier, similar performance. Note than in some (few) instances A_c-IMSE_w even outperforms A_d-IMSE_w. Though this is not expected behavior, it may occur due to overfitting, since the adaptive selection for λ is the optimal for the current DoE iteration but is implemented for future iterations, and is not guaranteed to be necessarily optimal for those.

Of course, it needs to be emphasized that A_d-IMSE_w enjoys significant robustness over A_c-IMSE_w, since, as discussed earlier, the promoted choice of $\lambda=2n_x$ is not guaranteed to be near-optimal for every application, and can potentially lead to vulnerabilities within the DoE by promoting low quality experiments. Though to the experience of the authors, this selection has performed very well in all examples we have tested so far, it might be a poor choice for some other applications. This robustness advantage that A_d-IMSE_w offers is significant. This discussion shows that, despite the similar performance of A_d-IMSE_w and A_c-IMSE_w and the higher computational burden of the former, the A_d-IMSE_w is the recommended choice, with a modification, though, of its tuning parameters to promote higher computational efficiency (reduce tolerances for convergence). This is further investigated in Figure 12 which shows results for three different target threshold ε_{perf} for the convergence criterion of Eq. (33), which represents the most influential parameter impacting the A_d-IMSE_w convergence. The resultant accuracy is only marginally impacted by the selection of ε_{perf}, but the impact on computational efficiency is significant: the mean efficiency for ε_{perf} threshold values of 5% and 10% are increased, respectively, to 0.87 and 0.91 for WtGMix, 0.95 and 0.95 for Hartman 6, 0.45 and 0.92 for

SDoF-KT, and 0.67 and 0.94 for Isolation examples. For complementing these comparisons note that the mean efficiency for the 1% threshold values was already presented in Figure 11. Similarly, no significant impact on accuracy is observed in the case of exponential decay bias interpolation (not reported here due to space limitation), while the mean efficiency, similarly, increased for the ε_{perf} threshold values of 5% and 10%, respectively, to 0.92 and 0.94 for WtGMix, and 0.96 for all other cases. Note that computational efficiency of 0.96 is the theoretical upper bound for the implementation considered here due to the selection to examine the quality of the approximation every N_λ=25 iterations.

Comparing across the two bias-interpolation function results in Figure 9, Figure 10, and Figure 15, show that the NN implementation outperforms the exponentially decaying function for the MSE_w criteria, while the exponentially decaying function performs equal or better for the $IMSE_w$ criteria. This is explained by the relatively heavier exploitive tendency of the NN interpolation function, as shown earlier in Figure 2 (prioritizing disproportionally domains with large bias). This tendency informs better the selection of the next experiment using a MSE metric - that focuses on the worst-case scenario performance- but does not have the same impact on the IMSE metric -that investigates average performance. Less sensitivity is observed for the performance of the IMSE approximation formulations, with both A_d-$IMSE_w$ and A_c-$IMSE_w$ showing a good agreement with the respective $IMSE_w$ implementation. It is important to note, though, that for the adaptive case, the trends of the acceptance ratio of λ, significantly differ (Figure 11) with the exponentially decaying function accommodating a much faster convergence to a credible λ value and overall to significant higher computational efficiency. This should be attributed to the discontinuous behavior, and therefore objectives in the λ^* optimization, introduced by the NN interpolation, which impacts the stability of the stochastic search-based identification of the optimal λ across the DoE iterations. This is clearly shown in Figure 13 with behavior of $\bar{\lambda}_k$ significantly affected by the choice of bias interpolation function in terms of both the mean underlying trend (convergence to different values) as well as the variability around this trend (greater shifts for the NN selection of interpolation function). Results for the mean trend in this figure also show that as the training sample size increases, the optimal value of λ may also increase, verifying arguments made earlier and also

showcasing the importance of re-evaluating appropriateness of the converged value for λ. The differences between the optimal λ values across the two interpolating functions showcase the importance of an adaptive selection (A_d-IMSE$_w$ formulation) since recommendation of an appropriate value for any desired application (as needed in the A_c-IMSE$_w$ formulation) seems to be impossible.

Next, the discussions move to the comparisons for the batch DoE selection. Results in Figure 14 and Figure 15 verify first the trends reported in (Kyprioti et al. 2020): though balancing between exploration and exploitation might provide some utility in earlier DoE iterations, especially when the metamodel accuracy is poor depending on the original DoE selection (Kyprioti et al. 2020), as this accuracy improves no such utility can be identified. This is evident in the results presented here by the fact that B-IMSE$_w$ outperforms all other variant implementations. Comparing across the different formulations that attempt to balance between exploration and exploitation, B^{ρ}-IMSE$_w$ and MB-IMSE$_w$, the multi-objective formulation proposed in this study (MB-IMSE$_w$) suffers from reduced performance compared to the baseline alternative (B^{ρ}-IMSE$_w$). Though trends are heavily dependent on the type of bias interpolation function used (topic discussed later), there is an undeniable performance reduction when the MB-IMSE$_w$ formulation is utilized. This reduction might be attributed to the choice for the promoted solution across the Pareto front, but alternative approaches that the authors examined have not yielded better results. Despite its more natural exploration and exploitation balance (no need to select a-priori an ρ) the reduction of performance for MB-IMSE$_w$ is an important constraint for its promotion as definite preferred solution over B^{ρ}-IMSE$_w$.

Comparing across the multi-objective DoE implementations for different bias interpolation functions [MB-IMSE$_w$-(NN) and MB-IMSE$_w$-(E)], significant sensitivity is observed for the MB-IMSE$_w$ formulation, much higher to sensitivity observed for any other variant. As demonstrated earlier in Figure 7 the NN implementation restricts the identification of a single candidate experiment within each Voronoi cell for the batch selection which can be very restrictive choice. This is the reason the exponentially decaying interpolation tends to outperform the NN interpolation by a significant margin in the majority of the case study examples.

Finally, the performance variability across the 30 trials, showcased through the boxplots in Figure 16, allows us to investigate some additional trends. As also reported in (Kyprioti et al. 2020) the implementations

without the bias weight (IMSE and MSE) demonstrate significant variability in the results, clearly showcasing, once more, the importance of introducing the bias weights in the DoE formulation. By comparison, the variability of the performance across all other variants is small. Examining the trends for the advancements introduced in this manuscript, we can observe a larger variability of A_d-IMSE$_w$ when compared to A_c-IMSE$_w$, as expected due to the addition of the adaptive λ selection within the DoE stochastic search implementation. The exponential decay interpolation function leads to reduced variability over its Voronoi counterpart, again as expected due to the discontinuous interpolation features of the latter, while the multi-objective implementation shows similar performance as the other batch selection DoE variants, with the exception of some outliers. This undesirable feature creates additional concerns for the promotion, at least at the current stage, of this DoE formulation.

Conclusions

This paper investigated the adaptive design of experiments for Gaussian Process (GP) global metamodeling applications. Formulation considered both the minimization of the predictive GP variance, representing an exploration strategy, and the reduction of the GP bias obtained through cross validation, representing an exploitation strategy. With respect to the GP variance, two alternative, popular implementations were discussed, the integrated mean square error (IMSE$_w$) criterion, and the simplified maximum square error without variance updating criterion (MSE$_w$). Two different advances were examined. The primary one considered an IMSE approximation, reducing the computational burden to levels comparable to MSE but offering the same DoE quality. This was accommodated through introduction of a decaying shape function that describes the reduction of GP variance within a domain of influence for each new candidate experiment. Two different variants were examined corresponding to an adaptive selection of the exponent involved in the shape function definition (A_d-IMSE$_w$) or an a-priori selection (A_c-IMSE$_w$). The adaptive selection is formulated by developing a framework for identification of a credible value of this exponent by judicial comparisons between the exact and approximate IMSE$_w$ results across the DoE iterations. The secondary advancement introduced a multi-objective DoE selection (MB-IMSE$_w$) for identifying multiple Pareto optimal experiments that balance exploration and exploitation objectives, replacing conventional

strategies that weight these objectives to promote a single DoE selection criterion. This formulation was promoted for batch-selection of simulation experiments. After the initial Pareto-front identification, candidate experiments within this front are chosen based on some desired optimality criterion, while an updating of the GP variance after identification of each such experiment is suggested, to avoid the identification of close-proximity experiments within the promoted batch.

The performance of the proposed DoE variants was illustrated across four case study examples, including both analytic benchmark functions and earthquake engineering applications. Across all DoE variants, two interpolation (smoothing) approaches were examined for accommodating the incorporation of bias weights: nearest neighbor (NN) interpolation that has been promoted in past studies and an exponential decay kernel smoothing introduced first here. The exponential decay provided more consistent performance, addressing challenges associated with the discontinuities introduced through the NN implementation, though it did not necessarily outperform it. For this reason, both remain appropriate candidate choices for future applications. Considering the DoE variants, results showed clearly the benefits of the proposed IMSE approximation, with both A_d-IMSE$_w$ and A_c-IMSE$_w$ outperforming MSE$_w$ and providing similar accuracy as the IMSE$_w$, at substantially reduced computational burden. Though A_c-IMSE$_w$ was not substantially outperformed by A_d-IMSE$_w$, the additional robustness offered by the latter and the dependence of an appropriate value for the exponent involved in the shape function on the details of the application considered, promote the A_d-IMSE$_w$ implementation as the recommended variant, despite its higher computational cost. Similar benefits were not illustrated for the secondary proposed advancement, with the multi-objective implementation showing reduced performance and greater variability of the results compared to alternative variants for batch selection of experiments. Despite its more natural exploration and exploitation balance, these characteristics mean that further advancements are needed for this formulation to be promoted over the alternative ones.

Though, as discussed above, a range of options remain attractive candidates for future applications, the recommended implementation is A_d-IMSE$_w$ with an exponential decay kernel smoothing.

Data Availability Statement

The DoE formulations discussed in this paper can be implemented through quoFEM, an open-source research software for uncertainty quantification (UQ) in natural hazard engineering, developed by NHERI SimCenter: https://simcenter.designsafe-ci.org/research-tools/quofem-application/. Numerical codes for the DoE optimization can be obtained from the corresponding author upon reasonable request.

Acknowledgments

This research was financially supported by the National Science Foundation under Grant CMMI- 2131111. This support is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Ankenman B, Nelson BL, Staum J (2008) Stochastic kriging for simulation metamodeling. 2008 Winter Simulation Conference, Miami, FL, December 7-10, IEEE, p 362-370.

Asher MJ, Croke BF, Jakeman AJ, Peeters LJ (2015) A review of surrogate models and their application to groundwater modeling. Water Resour Res 51(8):5957-5973.

Beck J, Guillas S (2016) Sequential design with mutual information for computer experiments (MICE): Emulation of a tsunami model. SIAM/ASA Journal on Uncertainty Quantification 4(1):739-766.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is nearest neighbor meaningful?. International conference on database theory, Berlin, Germany, Springer, p 217-235.

Blatman G, Sudret B (2010) An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. Prob Eng Mech 25(2):183-197.

Bodenmann L, Reuland Y, Stojadinovic B (2021) Dynamic Updating of Building Loss Predictions Using Regional Risk Models and Conventional Post-Earthquake Data Sources. Proceedings of the 31st European Safety and Reliability Conference (ESREL 2021), Angers, France, September 19-23, Research Publishing Services.

Byun J, Song J (2021) Generalized matrix-based Bayesian network for multi-state systems. Reliab Eng Syst Saf 211:107468.

Cole DA, Christianson RB, Gramacy RB (2021) Locally induced Gaussian processes for large-scale simulation experiments. Statistics and Computing 31(3):1-21.

Contreras MT, Gironás J, Escauriaza C (2020) Forecasting flood hazards in real time: a surrogate model for hydrometeorological events in an Andean watershed. Natural Hazards and Earth System Sciences 20(12):3261-3277.

Csató L, Opper M (2002) Sparse on-line Gaussian processes. Neural Comput 14(3):641-668.

Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE transactions on evolutionary computation 6(2):182-197.

Deierlein GG, Zsarnóczay A (2019) State-of-Art in Computational Simulation for Natural Hazards Engineering. doi:10.5281/zenodo.2579582.

Deisenroth M, Ng JW (2015) Distributed gaussian processes. Proceedings of the 32nd International Conference on Machine Learning, vol 37. , PMLR, p 1481-1490.
Dixon L, Szegö GP (1978) Towards global optimisation, V2 edn. North-Holland, Amsterdam.
Dubrule O (1983) Cross validation of kriging in a unique neighborhood. Journal of the International Association for Mathematical Geology 15(6):687-699.
Forrester AI, Keane AJ (2009) Recent advances in surrogate-based optimization. Prog Aerospace Sci 45(1-3):50-79.
Fuhg JN, Fau A, Nackenhorst U (2021) State-of-the-art and comparative review of adaptive sampling methods for kriging. Archives of Computational Methods in Engineering 28(4):2689-2747.
Garud SS, Karimi IA, Kraft M (2017) Design of computer experiments: A review. Comput Chem Eng 106:71-95.
Ginsbourger D (2014) Sequential design of computer experiments. Wiley StatsRef: Statistics Reference Online:1-9.
Gramacy RB (2020) Surrogates: Gaussian process modeling, design, and optimization for the applied sciences, 1st Edition edn. Chapman and Hall/CRC, New York.
Gunantara N (2018) A review of multi-objective optimization: Methods and its applications. Cogent Engineering 5(1):1502242.
Jain AK, Jianchang Mao, Mohiuddin KM (1996) Artificial neural networks: a tutorial. MC 29(3):31-44. doi:10.1109/2.485891.
Jia G, Taflanidis AA (2013) Kriging metamodeling for approximation of high-dimensional wave and surge responses in real-time storm/hurricane risk assessment. Comput Methods Appl Mech Eng 261:24-38.
Jia G, Taflanidis AA (2014) Sample-based evaluation of global probabilistic sensitivity measures. Comput Struct 144:103-118.
Jiang P, Shu L, Zhou Q, Zhou H, Shao X, Xu J (2015) A novel sequential exploration-exploitation sampling strategy for global metamodeling. IFAC-PapersOnLine 48(28):532-537.
Jin R, Chen W, Sudjianto A (2002) On sequential sampling for global metamodeling in engineering design. Proceedings of the ASME 2002 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference. Volume 2: 28th Design Automation Conference, Montreal, Quebec, Canada., September 29–October 2, vol 36223., p 539-548.
Johnson ME, Moore LM, Ylvisaker D (1990) Minimax and maximin distance designs. Journal of statistical planning and inference 26(2):131-148.
Kanai K (1957) Semi-empirical formula for the seismic characteristics of the ground. 東京大學地震研究所彙報 (Bulletin of the Earthquake Research Institute, University of Tokyo) 35(2):309-325.
Kim IY, De Weck OL (2005) Adaptive weighted-sum method for bi-objective optimization: Pareto front generation. Structural and multidisciplinary optimization 29(2):149-158.
Kim J, Song J (2020) Probability-Adaptive Kriging in n-Ball (PAK-Bn) for reliability analysis. Struct Saf 85:101924.
Kleijnen JP (2009) Kriging metamodeling in simulation: A review. Eur J Oper Res 192(3):707-716.
Kleijnen JP (2017) Regression and Kriging metamodels with their experimental designs in simulation: a review. Eur J Oper Res 256(1):1-16.
Kleijnen JP, Beers WV (2004) Application-driven sequential designs for simulation experiments: Kriging metamodelling. J Oper Res Soc 55(8):876-883.
Kleijnen JP, Van Beers WC (2022) Statistical tests for cross-validation of Kriging models. INFORMS Journal on Computing 34(1):607-621.
Koehler JR, Owen AB (1996) 9 Computer experiments. In: Ghosh S, Rao CR (eds) Handbook of Statistics, vol 13. ,Elsevier Science & Technology, p 261-308.
Krause A, Singh A, Guestrin C (2008) Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. Journal of Machine Learning Research 9(2):235-284.
Kyprioti AP, Taflanidis AA, Nadal-Caraballo NC, Campbell M (2021) Storm hazard analysis over extended geospatial grids utilizing surrogate models. Coast Eng 168:103855.
Kyprioti AP, Zhang J, Taflanidis AA (2020) Adaptive design of experiments for global Kriging metamodeling through cross-validation information. Structural and Multidisciplinary Optimization 62(3):1135-1157.
Le Gratiet L, Cannamela C (2015) Cokriging-based sequential design strategies using fast cross-validation techniques for multi-fidelity computer codes. Technometrics 57(3):418-427.
Liu H, Cai J, Ong Y (2017) An adaptive sampling approach for Kriging metamodeling by maximizing expected prediction error. Comput Chem Eng 106:171-182.
Liu H, Ong Y, Cai J (2018) A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. Structural and Multidisciplinary Optimization 57(1):393-416.
Liu H, Ong Y, Shen X, Cai J (2020) When Gaussian process meets big data: A review of scalable GPs. IEEE transactions on neural networks and learning systems 31(11):4405-4423.

Liu H, Xu S, Ma Y, Chen X, Wang X (2016) An adaptive Bayesian sequential sampling approach for global metamodeling. Journal of Mechanical Design 138(1):011404.
Lophaven SN, Nielsen HB, Søndergaard J (2002) DACE - A Matlab Kriging Toolbox, Version 2.0 http://www.imm.dtu.dk/pubdb/p.php?3213.
Lutes LD, Sarkani S (2004) Random Vibrations. Elsevier Science & Technology, Oxford.
McBride K, Sundmacher K (2019) Overview of surrogate modeling in chemical process engineering. Chemie Ingenieur Technik 91(3):228-239.
McKay MD, Beckman RJ, Conover WJ (2000) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 42(1):55-61.
Moustapha M, Marelli S, Sudret B (2022) Active learning for structural reliability: Survey, general framework and benchmark. Structural safety 96(2):102174. doi:10.1016/j.strusafe.2021.102174.
Page TJ (1984) Multivariate Statistics: A Vector Space Approach. Journal of Marketing Research 21(2):236.
Pandita P, Tsilifis P, Awalgaonkar NM, Bilionis I, Panchal J (2021) Surrogate-based sequential Bayesian experimental design using non-stationary Gaussian Processes. Comput Methods Appl Mech Eng 385:114007.
Parlett BN (1981) LINPACK Users' Guide. SIAM Review 23(1):126-128.
Picheny V, Ginsbourger D, Roustant O, Haftka RT, Kim N (2010) Adaptive Designs of Experiments for Accurate Approximation of a Target Region. J Mech Des 132(7). doi:10.1115/1.4001873.
Prebeg P, Zanic V, Vazic B (2014) Application of a surrogate modeling to the ship structural design. Ocean Eng 84:259-272. doi:https://doi.org/10.1016/j.oceaneng.2014.03.032.
Provost F, Jensen D, Oates T (1999) Efficient progressive sampling. Conference on Knowledge Discovery in Data: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining; 15-18 Aug. 1999 , ACM, p 23-32.
Rasmussen CE, Nickisch H (2010) Gaussian processes for machine learning (GPML) toolbox. The Journal of Machine Learning Research 11:3011-3015.
Razavi S, Tolson BA, Burn DH (2012) Review of surrogate modeling in water resources. Water Resour Res 48(7).
Roustant O, Ginsbourger D, Deville Y (2012) DiceKriging, DiceOptim: two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization. Journal of statistical software 51(1). doi:10.18637/jss.v051.i01.
Sacks J, Schiller SB, Welch WJ (1989) Designs for computer experiments. Technometrics 31(1):41-47.
Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Statistical science 4(4):409-423.
Stewart GW (1998) Matrix algorithms. SIAM, Philadelphia, PA.
Strang G (2016) Introduction to linear algebra, Fifth edition edn. Wellesley-Cambridge Press, Wellesley, MA.
Sundararajan S, Keerthi SS (2001) Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. Neural Comput 13(5):1103-1118.
Tajimi H (1960) Statistical method of determining the maximum response of building structure during an earthquake. Proc.of the 2nd WCEE 2:781-798.
Vazquez E, Bect J (2011) Sequential search based on kriging: convergence analysis of some algorithms. arXiv preprint arXiv:1111.3866.
Welch WJ (1983) A mean squared error criterion for the design of experiments. Biometrika 70(1):205-213.
Xu S, Liu H, Wang X, Jiang X (2014) A robust error-pursuing sequential sampling approach for global metamodeling based on voronoi diagram and cross validation. Journal of Mechanical Design 136(7):071009.
Zhang J, Taflanidis AA (2018) Adaptive kriging stochastic sampling and density approximation and its application to rare-event estimation. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 4(3):04018021.
Zhang J, Taflanidis AA, Nadal-Caraballo NC, Melby JA, Diop F (2018) Advances in surrogate modeling for storm surge prediction: storm selection and addressing characteristics related to climate change. Nat Hazards 94(3):1225-1253.

Appendix A: Calibration and validation of Gaussian process

The three sets of unknown GP parameters $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2\}$ can be selected using maximum likelihood estimation (MLE). The MLE solution for the basis coefficients and process variance is provided, respectively, by the closed-form solution (Lophaven et al. 2002, Sacks, Welch et al. 1989):

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}^*(\mathbf{D}, \boldsymbol{\theta}) = (\mathbf{F}(\mathbf{X})^T \mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})^{-1} \mathbf{F}(\mathbf{X}))^{-1} \mathbf{F}(\mathbf{X})^T \mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})^{-1} \mathbf{Z} \tag{41}$$

$$\left(\tilde{\sigma}^*(\mathbf{D}, \boldsymbol{\theta})\right)^2 = \frac{(\mathbf{Z} - \mathbf{F}(\mathbf{X})^T \boldsymbol{\beta}^*)^T \mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})^{-1} (\mathbf{Z} - \mathbf{F}(\mathbf{X})^T \boldsymbol{\beta}^*)}{n} \tag{42}$$

whereas for the hyper-parameters, the MLE selection corresponds to optimization:

$$\boldsymbol{\theta}^*(\mathbf{D}) = \arg\min \left[\ln(|\mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})|) + n \ln\left((\tilde{\sigma}^*(\mathbf{D}, \boldsymbol{\theta}))^2\right) \right] \tag{43}$$

which can be numerically performed (Lophaven et al. 2002, Rasmussen and Nickisch 2010). Note that for notational simplicity, $\boldsymbol{\beta}^*(\mathbf{D}, \boldsymbol{\theta})$ is replaced by $\boldsymbol{\beta}^*$ in the remaining discussions. .

The predictive capabilities of the calibrated GP can be quantified using cross leave-one-out cross validation (LOOCV) (Kyprioti et al. 2020, Fuhg et al. 2021, Kleijnen 2009), established by removing the *i*th simulation from the training set, using the remaining set to estimate predictions for that specific simulation, and then defining the leave-one-out (LOO) error as the difference between the exact and prediction response:

$$e_i^{cv} = \tilde{z}(\mathbf{x}_i \mid \mathbf{D}_{\sim i}, \Theta) - \mathrm{z}_i \tag{44}$$

where $\mathbf{D}_{\sim i} = \{\mathbf{X}_{\sim i}, \mathbf{Z}_{\sim i}\}$ represents the training data without the *i*th simulation [remove the *i*th row from matrices $\{\mathbf{X}, \mathbf{Z}\}$]. If the LOO predictions are established, as is customary, without re-calibration of the GP parameters, closed-form solutions are readily available for e_i^{cv}, given by (Sundararajan and Keerthi 2001, Dubrule 1983, Rasmussen and Nickisch 2010):

$$e_i^{cv} = \frac{\left[\mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})^{-1} (\mathbf{Z} - \mathbf{F}(\mathbf{X}) \boldsymbol{\beta}^*)\right]_i}{\left[\mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})^{-1}\right]_{ii}} \tag{45}$$

where $[\bullet]_{ii}$ denotes *i*th diagonal component in a matrix, and $[\bullet]_i$ denotes *i*th element of a vector. These closed-form solutions greatly simplify LOOCV, avoiding the repeated matrix inversion operation to compute each $\tilde{z}(\mathbf{x}_i \mid \mathbf{D}_{\sim i}, \Theta)$, reducing the computational cost of the error evaluation to be $O(n^2)$ for a given $\mathbf{R}(\mathbf{X} \mid \boldsymbol{\theta})$. Once

the LOOCV error is estimated, any desired validation metric can be utilized to assess the global metamodel accuracy. In this study, the normalized root mean square error (NRMSE) is used, given by:

$$NRMSE_{cv} = \frac{\sqrt{\sum_{i=1}^{n} (e_i^{cv})^2 / n}}{\max_i \{z_i\} - \min_i \{z_i\}} \tag{46}$$

where $\max_i \{z_i\}$ and $\min_i \{z_i\}$ denote the maximum and minimum values of the response dataset. When the NRMSE is smaller than a desired tolerance level, the GP may be considered as well-constructed for global metamodeling purposes.

Appendix B: Computational details for variance estimation

This Appendix discusses computational details for the variance estimation, including its updating when new experiments are added. Specifically the focus is on the estimation of the normalized variance given by Eq. (4), the only component of the variance that depends on **x,** and therefore needs to be repeatedly estimated for each examined input. Also in all discussions on numerical complexity the assumption that $n>>n_b$ is made, expressing this complexity primarily as function of n.

Starting with the variance estimation, the computationally intensive part, as discussed previously, is the inversion of the correlation matrix $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$, which has burden $O(n^3)$ (Strang 2016). Typically, this inversion is replaced with operations that utilize the Cholesky decomposition for better numerical accuracy and stability. To formalize these operations, let $\mathbf{L}(\mathbf{X}|\boldsymbol{\theta}) \in \mathbb{R}^{n\times n}$ be the lower triangular matrix satisfying the Cholesky factorization $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta}) = \mathbf{L}(\mathbf{X}|\boldsymbol{\theta})\mathbf{L}(\mathbf{X}|\boldsymbol{\theta})^T$. Note that performing inversion or Cholesky decomposition have the same computational burden $O(n^3)$, and the operations based on the Cholesky decomposition are introduced primarily for numerical stability. Using $\mathbf{L}(\mathbf{X}|\boldsymbol{\theta})$, Eq. (4), is expressed as (Roustant et al. 2012):

$$\underline{\sigma}^2(\mathbf{x}|\mathbf{X},\boldsymbol{\theta}) = 1 - \left\|\mathbf{v}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta})\right\|_2^2 + \left\|\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta}) \setminus (\mathbf{M}(\mathbf{X}|\boldsymbol{\theta})^T \mathbf{v}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta}) - \mathbf{f}(\mathbf{x}))\right\|_2^2 \tag{47}$$

where $\|.\|_2$ is the vector two-norm, vector $\mathbf{v}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta}) \in \mathbb{R}^n$ and matrix $\mathbf{M}(\mathbf{X}|\boldsymbol{\theta}) \in \mathbb{R}^{n\times n_b}$ are given by $\mathbf{v}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta}) = \mathbf{L}(\mathbf{X}|\boldsymbol{\theta}) \setminus \mathbf{r}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta})$ and $\mathbf{M}(\mathbf{X}|\boldsymbol{\theta}) = \mathbf{L}(\mathbf{X}|\boldsymbol{\theta}) \setminus \mathbf{F}(\mathbf{X})$, respectively, with "\" denoting inversion by forward or backward substitution, and $\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta}) \in \mathbb{R}^{n_b\times n_b}$ is the upper triangular matrix originating from the

economy size orthogonal triangular decomposition (also known as QR decomposition) of $\mathbf{M}(\mathbf{X}|\boldsymbol{\theta})$ [so that $\mathbf{M}(\mathbf{X}|\boldsymbol{\theta}) = \mathbf{M}_Q(\mathbf{X}|\boldsymbol{\theta})\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta})$ with $\mathbf{M}_Q(\mathbf{X}|\boldsymbol{\theta}) \in \mathbb{R}^{n \times n_b}$ and $\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta})$ having the characteristics described above]. The computational intensive components of the estimate in Eq. (47), corresponding to the decompositions for estimation of primarily $\mathbf{L}(\mathbf{X}|\boldsymbol{\theta})$ [complexity $O(n^3)$] and secondarily $\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta})$ [complexity $O(n_b^2 n)$], are independent of $\mathbf{x}$ and therefore need to be performed only once. Actually, both quantities are readily available from the GP calibration (Roustant et al. 2012). Therefore, the estimation of Eq. (47) for each new $\mathbf{x}$ has small computational burden, as it only involves calculation of vector $\mathbf{v}(\mathbf{x}|\mathbf{X},\boldsymbol{\theta})$ and expression $\mathbf{M}_R(\mathbf{X}|\boldsymbol{\theta}) \setminus (\mathbf{M}(\mathbf{X}|\boldsymbol{\theta})^T \mathbf{v}(\mathbf{x}|\mathbf{X},,\boldsymbol{\theta}) - \mathbf{f}(\mathbf{x}))$, and the estimation of the two vector norms. This burden is $O(n^2)$ overall for each examined $\mathbf{x}$ (Roustant et al. 2012).

Moving, further, to the calculation of the updated variance $\underline{\sigma}^2(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$ after the addition of $\mathbf{x}_{new}$ in the existing training set given by Eq. (11), the estimate of Eq. (47) is updated to be:

$$\underline{\sigma}^2(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) = 1 - \left\|\mathbf{v}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})\right\|_2^2 + \left\|\mathbf{M}_R(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta}) \setminus (\mathbf{M}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})^T \mathbf{v}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) - \mathbf{f}(\mathbf{x}))\right\|_2^2 \quad (48)$$

where $\mathbf{v}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) = \mathbf{L}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta}) \setminus \mathbf{r}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$ and $\mathbf{M}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta}) = \mathbf{L}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta}) \setminus \mathbf{F}(\mathbf{X},\mathbf{x}_{new})$, with $\mathbf{L}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$ corresponding to the lower Cholesky factorization of $\mathbf{R}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$ and $\mathbf{M}_R(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$ to the upper triangular matrix from the orthogonal triangular decomposition of $\mathbf{M}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$, and the updated correlation vector $\mathbf{r}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})$, basis function matrix $\mathbf{F}(\mathbf{X},\mathbf{x}_{new})$ and correlation matrix $\mathbf{R}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$ are given, respectively, by Eqs. (8), (9) and (10). The computationally intensive component of this estimation is the calculation of the Cholesky decomposition of $\mathbf{R}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta})$ which has complexity $O(n^3)$. This can be simplified utilizing the readily available decomposition of $\mathbf{R}(\mathbf{X}|\boldsymbol{\theta})$ to be (Parlett 1981, Stewart 1998):

$$\mathbf{L}(\mathbf{X},\mathbf{x}_{new}|\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{L}(\mathbf{X}|\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{v}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})^{\mathrm{T}} & \sqrt{R(\mathbf{x}_{new},\mathbf{x}_{new}|\boldsymbol{\theta}) - \left\|\mathbf{v}(\mathbf{x}|\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta})\right\|_2^2} \end{bmatrix} \quad (49)$$

which has computational complexity $O(n^2)$. This reduces the computational burden for estimation of the updated predictive variance when a single experiment is added in the training set, to be $O(n^2)$ and $O(n_b^2 n)$ for

estimating the independent from $\mathbf{x}$ matrices, $\mathbf{L}(\mathbf{X},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})$ and $\mathbf{M}_R(\mathbf{X},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})$, respectively, and then additional $O(n^2)$, as discussed earlier, for each input $\mathbf{x}$ for which this variance needs to be estimated.

Appendix C: Examining limiting cases for λ

This section further discusses some limiting cases for the behavior of the variance reduction approximation prescribed by Eqs. (20) and (21). First, consider the case where all the previous selected experiments $\mathbf{X}$ are in a far distance from $\mathbf{x}$ and $\mathbf{x}_{new}$, so that following condition is satisfied:

$$\left\|\mathbf{x}-\mathbf{x}_{\text{new}}\right\| \ll \left\|\mathbf{x}-\mathbf{x}_i\right\|, \text{ for } \forall\ i=0,...,n \tag{50}$$

We can then approximate $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) \simeq \underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{x}_{new},\boldsymbol{\theta})$, indicating that as soon as $\mathbf{x}_{new}$ is added it dominates the variance estimation for $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{new}$ are jointly Gaussian with correlation coefficient $R(\mathbf{x},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})$ the conditional variance is further calculated as:

$$\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\mathbf{x}_{new},\boldsymbol{\theta}) \approx \underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{x}_{new},\boldsymbol{\theta}) = \underline{\sigma}^2(\mathbf{x}\,|\,\boldsymbol{\theta})\left(1-R(\mathbf{x},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})^2\right) \tag{51}$$

where $\underline{\sigma}^2(\mathbf{x}\,|\,\boldsymbol{\theta})$ represents the prior variance at $\mathbf{x}$ when $\mathbf{x}_{new}$ is removed, i.e. ultimately corresponds to $\underline{\sigma}^2(\mathbf{x}\,|\,\mathbf{X},\boldsymbol{\theta})$ within the GP formulation. Comparing Eqs. (51) and (22), we obtain $A(\mathbf{x},\mathbf{x}_{new}\,|\,\lambda) = R(\mathbf{x},\mathbf{x}_{new}\,|\,\lambda)^2$, which reduces to Eq. (21) with $\lambda = 2$.

However, the condition in Eq. (50) is easily violated when the number of training samples increases. More importantly, this condition is easily violated in higher dimensions for the input, since, as n_x increases, the mutual distances between any pairs of the training data points $\mathbf{X}$, the candidate points $\mathbf{x}_{new}$, and examined $\mathbf{x}$ rapidly becomes indiscernible to each other (Beyer et al. 1999). Moreover, it can be easily shown that Eq. (51) is an upper bound. If $R(\mathbf{x},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})$ is replaced with $R(\mathbf{x},\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta})$ in this equation, then formulation becomes exact. But since $\mathbf{x}$, $\mathbf{x}_{new}$ and $\mathbf{X}$ are jointly Gaussian, it can be proven that $R(\mathbf{x},\mathbf{x}_{new}\,|\,\mathbf{X},\boldsymbol{\theta}) \le R(\mathbf{x},\mathbf{x}_{new}\,|\,\boldsymbol{\theta})$ (Page 1984), implying that the relative impact of $\mathbf{x}_{new}$ on $\mathbf{x}$ will be reduced under presence of other close-proximity training points, making Eq. (51) an upper bound. Note that $R(\mathbf{x},\mathbf{x}_{new}\,|\,\lambda)^2 < A(\mathbf{x},\mathbf{x}_{new}\,|\,\lambda)$ for $\lambda < 2$

, indicating that in order to avoid further increasing the gap to $R(\mathbf{x}, \mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta})$ the exponent for $A(\mathbf{x}, \mathbf{x}_{new} \mid \lambda)$ should be greater than 2. Additionally, this is showcasing that a variation of λ is expected as more experiments are added and the condition $R(\mathbf{x}, \mathbf{x}_{new} \mid \mathbf{X}, \boldsymbol{\theta}) \leq R(\mathbf{x}, \mathbf{x}_{new} \mid \boldsymbol{\theta})$ holds with larger margins, indicating a trend towards use of larger values of λ in such conditions. Note that an explicit dependence on the hyper-parameters exists in these arguments (defining correlation length scales), indicating an even more complex relationship to not only the training points **X,** but also the GP features.

Figure Captions

Figure 1 Adaptive design of experiments workflow. Note that implementation also considers batch selection of experiments.

Figure 2 Illustration of components for variance and bias functions for a two-dimensional example.

Figure 3 Illustration of variance reduction approximation

Figure 4 Illustration of selection of objective function for calibration of λ

Figure 5 Workflow for the approximated IMSE algorithm ($IMSE_w$-SA)

Figure 6 Illustration of multi-objective DoE. Identification of Pareto-front and selection

Figure 7 Illustration of multi-objective DoE aspects for two different interpolation functions (NN: natural neighbor, E: exponential decay).

Figure 8 Workflow for the multi-objective DoE algorithm (IMSE-MB)

Figure 9 Comparison of metamodel accuracy (average over 30 trials) for different values of n for DoE variants corresponding to sequential experiment selection using NN for bias interpolation function.

Figure 10 Comparison of metamodel accuracy (average over 30 trials) for different values of n for DoE variants corresponding to sequential experiment selection using exponentially decaying bias interpolation function.

Figure 11 Acceptance rate of λ and overall efficiency of A_d-$IMSE_w$

Figure 12 Comparison of metamodel accuracy (average over 30 trials) for different values of n for A_c-$IMSE_w$ and A_d-$IMSE_w$ with different minimum performance thresholds using NN for bias interpolation function.

Figure 13 The averaged λ for A_d-IMSE$_w$ (SDoF-KT example)

Figure 14 Comparison of metamodel accuracy (average over 30 trials) for different values of n for DoE variants corresponding to batch (size of 5) experiment selection

Figure 15 Comparison of metamodel accuracy (average over 30 trials) for different values of n for selective DoE variants, emphasizing comparisons across the different bias interpolation functions.

Figure 16 Variability in the accuracy of the illustrated DoE methods across 30 independent trials.