On kernel functions for bi-fidelity Gaussian process regressions

Pramudita Satria Palar[1] · **Lucia Parussini**[2] · **Luigi Bregant**[2] · **Koji Shimoyama**[3] · **Lavi Rizki Zuhal**[1]

Abstract
This paper investigates the impact of kernel functions on the accuracy of bi-fidelity Gaussian process regressions (GPR) for engineering applications. The potential of composite kernel learning (CKL) and model selection is also studied, aiming to ease the process of manual kernel selection. Using the autoregressive Gaussian process as the base model, this paper studies four kernel functions and their combinations: Gaussian, Matern-3/2, Matern-5/2, and Cubic. Experiments on four engineering test problems show that the best kernel is problem dependent and sometimes might be counter-intuitive, even when a large amount of low-fidelity data already aids the model. In this regard, using CKL or automatic kernel selection via cross validation and maximum likelihood can reduce the tendency to select a poor-performing kernel. In addition, the CKL technique can create a slightly more accurate model than the best-performing individual kernel. The main drawback of CKL is its significantly expensive computational cost. The results also show that, given a sufficient amount of samples, tuning the regression term is important to improve the accuracy and robustness of bi-fidelity GPR, while decreasing the importance of the proper kernel selection.

Keywords Kernel function · Bi-fidelity · Gaussian process regression · Engineering design

1 Introduction

Surrogate models are versatile tools capable of accelerating scientific and engineering tasks in situations where computationally expensive computer simulations hinder the practical use of such simulations. Multi-fidelity (MF) simulation is one research topic currently under intense study in the field of surrogate modeling (Zhang et al. 2021; Park et al. 2017). The main principle of MF modeling is to fuse various information from simulations with different levels of fidelity. In general, the role of low-fidelity simulations is to capture the general trend of the data. In contrast, fewer high-fidelity simulations amend the low-fidelity data to create more accurate models. Among several applications of MF surrogate models in engineering include design optimization (Liu et al. 2016; Tao and Sun 2019), uncertainty quantification (de Baar et al. 2015; Jofre et al. 2018), sensitivity analysis (Palar et al. 2018), and also reliability analysis (Yoo et al. 2020). An MF surrogate model with two fidelity levels is commonly called a bi-fidelity model.

Kernel-based models are suitable for MF modeling due to their non-parametric nature, e.g., MF radial basis function (Song et al. 2019; Serani et al. 2019) and support vector regression (Maolin et al. 2020). In the field of uncertainty quantification, non-intrusive polynomial chaos expansion has also been extended into its MF counterparts (Ng and Eldred 2012; Palar et al. 2016; Bryson and Rumpfkeil 2017). Variants of MF deep neural networks have also been developed and applied to problems such as aerodynamic design (Zhang et al. 2021) and vortex-induced vibration (Meng et al. 2021). The low-fidelity simulation can be created by several means, e.g., by reducing the number of mesh elements, using looser convergence criteria, or using governing equations with lower fidelity (Fernández-Godino et al. 2016).

In this paper, we are interested in the Gaussian process regression (GPR) surrogate model, also known as Kriging, which proved to be versatile in many applications. There

Responsible Editor: Ramin Bostanabad

✉ Pramudita Satria Palar
pramsp@ftmd.itb.ac.id

[1] Faculty of Mechanical and Aerospace Engineering, Institut Teknologi Bandung, Bandung, Indonesia

[2] Department of Engineering and Architecture, University of Trieste, Piazzale Europa 1, 34127 Trieste, Italy

[3] Institute of Fluid Science, Tohoku University, Sendai, Japan

are several variants of MF GPR, in which the linear autoregressive co-Kriging is probably the first of such variants (Kennedy and O'Hagan 2000). Other variants include the hierarchical GPR (Han and Görtz 2012), which treats the low-fidelity GP as the trend for the high-fidelity GP, and non-linear autoregressive Gaussian Process (NARGP) which employs non-linear correlation between the high- and low-fidelity simulations (Perdikaris et al. 2017). The recursive co-Kriging model is equivalent to the original autoregressive co-Kriging model but allows the construction of separate GPR models from different levels of fidelity to be fused into a single model. An MF GP can also be constructed using a deep representation, in which the resulting model is called MF deep GP (Cutajar et al. 2019). Breault et al. compared various MF Kriging techniques and argued that the recursive co-Kriging is a robust technique for various aerospace engineering applications (Brevault et al. 2020). On the other hand, MF deep GP typically prevails in problems with the more complex relationship between the LF and HF simulation. Toal agreed that the correlation between the low- and high-fidelity samples should be high to ensure good quality MF co-Kriging model (Toal 2015). However, as mentioned earlier, specific variants of MF GP, such as NARGP and MF deep GP can deal with problems in which the correlation is low but exhibits some spatial correlations.

Despite extensive research in MF GPR, the selection of kernel function is an issue that still needs to be addressed. The squared exponential/Gaussian is arguably the most widely used kernel function in MF GPR, as used in several previous research (de Baar et al. 2015; Liu et al. 2022). Other popular alternatives are the Matern kernels, which have also been applied in modeling physics and engineering problems (Pang et al. 2017; Bonfiglio et al. 2017). The cubic function is another alternative; although its implementation is rare than Gaussian and Matern, some applications deploy such a kernel in practice, e.g., Bertram et al. (2018). Misspecification of the kernel can result in a poor-performing surrogate model, as observed in some previous publications that applied GPR on analytical functions (Bachoc 2013) and engineering problems (Satria Palar et al. 2020). To the best of our knowledge, no research intensively compared various kernel functions in the context of MF GPR. Moreover, there is also a potential to use a combination of kernel functions to build more accurate MF Kriging models. Combinations of kernel functions themselves have been studied in single-fidelity Gaussian Process, in which the combinations can be constructed via genetic programming (Kronberger and Kommenda 2013; Jin 2020), weighted-sum formulation (Satria Palar et al. 2020), or greedy search. A typical objective function used in finding the best combination is the likelihood function or Akaike/Bayesian information criterion, although the latter is typically used to compare several final models. Another approach is to combine multiple GPR models with various kernel functions, as proposed by Sundar and Shield for reliability analysis (Sundar and Shields 2019). The combination of kernel functions is attractive since it eases the process of manual kernel selection, with the potential for higher accuracy.

Recently, Satria Palar et al. (2020) proposed GPR with composite kernel learning (CKL) technique that constructs an ensemble of single kernels with primary applications in engineering design optimization. The CKL considers a weighted-sum combination of kernels, in which the weights of the mixture are treated as extra hyperparameters and simultaneously optimized with the length scales. It is worth noting that there exists a technique with a similar name called compositional kernel learning (Jin 2020), which uses genetic programming to build the kernels. The kernels selected for CKL are frequently used in engineering design (e.g., Gaussian and Matern kernels), although using other kernels is possible. It has been shown that GP with CKL yields surrogate models with better accuracy than a single kernel. In this respect, CKL deals with the issue of kernel selection and improving the model's accuracy simultaneously. The excellent performance of CKL on single-fidelity problems suggests its potential for bi-fidelity surrogate modeling with GP. A common practice in using a bi-fidelity GP is to use the same kernel in all fidelity levels. It is possible that the accuracy can be further improved if different kernels are used. However, it is not trivial to pick the most suitable kernels. The utilization of CKL in bi-fidelity GP then aims to reduce the burden on kernel selection while achieving better accuracy simultaneously, as shown in Palar et al. (2022). However, the experiment in Palar et al. (2022) was performed under a noiseless setting (i.e., the regression term is set to the smallest constant to keep stable computation). Therefore, the result interpretation might differ if the regression term is set as one hyperparameter.

This paper aims to shed light on the potential of kernel selection and combination of kernels for building a more accurate bi-fidelity GP model. Even more fundamentally, our work investigates the impact of kernel functions for bi-fidelity approximations using GP, with a specific context toward applications in engineering design. The model of interest is the recursive co-Kriging, although the result from this study will also be helpful for other variants of bi-fidelity GP. We pick the recursive co-Kriging as the method of choice primarily due to its robustness in various types of engineering problem (Brevault et al. 2020). Furthermore, the simplicity of the recursive co-Kriging, compared to more complex variants such as MF deep GP, also makes it a good choice for focusing on the impact of kernel functions. However, we are aware of the limitations of the recursive co-Kriging with a linear autoregressive model, which is unsuitable for bi-fidelity problems with complex correlations. For complex correlation, we hypothesized

that the type of GP might play a bigger role than the kernel function, which is why methods such as deep GP were developed in the first place. This paper serves as a first step toward understanding the impact of kernel functions on MF GP. In practice, when the correlation between the low- and high-fidelity function is small, it is difficult to guess whether it is due to a complex correlation or no correlation at all. Therefore, this paper focuses on problems with evident correlation and the relationship between the low- and high-fidelity function is not utterly complex.

We study two criteria for kernel selection, namely, cross-validation error and maximum likelihood criteria. As for the combined kernel, we study the potential of CKL for bi-fidelity GP, aiming to eliminate manual kernel selection while simultaneously achieving higher accuracy. In this paper, the number of fidelity levels is set to two, although extending it to three or more levels is certainly possible. The reason why the present study focuses on two fidelity levels is that most engineering applications of MF techniques deal with only two fidelity levels (Fernández-Godino et al. 2016). Furthermore, it is sufficient to draw the first important insight regarding the impact of kernel functions using only two fidelity levels, especially considering the scope and objectives of our study, as explained in the next paragraph. Although using more than two fidelity levels is certainly interesting, it will complicate the discussion and analysis of the result. Therefore, we leave the study on more than two fidelity levels as the topic for future works. We performed experiments on four engineering problems with various levels of fidelity and complexity to investigate the impact of kernel functions for bi-fidelity GPR. In addition, we are also interested in investigating the impact of tuning the regression term on kernel selection since our focus is on bi-fidelity computer simulations. There is a possibility that a proper tuning of the regression term reduces the necessity of kernel selection, which is why this study compares the implementation of bi-fidelity GP with tunable regression term and that with the computed smallest regression term to ensure stable computation (Ranjan et al. 2011).

In particular, the current manuscript (1) compares the impact of various kernel functions, (2) investigates the potential of CKL for bi-fidelity GP, (3) investigates kernel selection as a cheaper alternative to CKL, (4) thoroughly investigates the composition of composite kernels for better understanding on the mechanism of CKL, (5) discusses the computational cost associated with all the models, and (6) investigates the effect of tuning the regression term on the choice of the kernel function.

2 Gaussian process surrogate model

A surrogate model is an approximation model $\hat{y}(\mathbf{x})$, where $\mathbf{x} = \{x_1, x_2, \dots, x_m\}$ and $m \geq 1$, that tries to mimic the actual input-output relationship $y(\mathbf{x})$. Our primary interest is the GP model which handles the black-box function as a realization of a multivariate Gaussian process $Y(\mathbf{x})$:

$$Y(\mathbf{x}) = \mu(\mathbf{x}) + \delta(\mathbf{x}), \tag{1}$$

where $\mu(\mathbf{x})$ is a deterministic regression function, constructed by observed data, and $\delta(\mathbf{x})$ is a Gaussian process, constructed through the residuals, with zero mean and covariance function cov($\mathbf{x}$,$\mathbf{x}'$). Hence, building $\hat{y}(\mathbf{x})$ requires evaluating $y(\mathbf{x})$ at an experimental design set $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(n)}\}$, where n is the sample size, to obtain corresponding responses, i.e., $\boldsymbol{y} = \{y^{(1)}, y^{(2)}, \dots, y^{(n)}\}^T$, where $y^{(i)} = y(\mathbf{x}^{(i)})$ for $i = 1, 2, \dots, n$.

Moreover, it is necessary to specify the correlation between the responses of different points in the input space. The correlation between the responses at two arbitrary points, say, $\mathbf{x}$ and $\mathbf{x}'$, is modeled by the kernel function $k(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}) = \text{corr}(\mathbf{x},\mathbf{x}')$, where $\boldsymbol{\theta} = \{\theta_1, \theta_2, \dots, \theta_m\}$ is the vector of lengthscale. The covariance function is defined as $\text{cov}(\mathbf{x},\mathbf{x}') = \sigma^2 k(\mathbf{x},\mathbf{x}';\boldsymbol{\theta})$, where σ^2 is the GP variance.

The most widely used kernel is probably the Gaussian kernel, which one-dimensional definition reads as

$$k(x, x';\theta) = \exp\left(-0.5\left(\frac{x - x'}{\theta}\right)^2\right). \tag{2}$$

In the CKL formulation, besides Gaussian, other choices include the Matern-3/2, Matern-5/2 and cubic kernels which are, respectively, written as

$$k(x, x';\theta, \nu = 3/2) = \left(1 + \frac{\sqrt{3}|x - x'|}{\theta}\right)\exp\left(-\frac{\sqrt{3}|x - x'|}{\theta}\right), \tag{3}$$

$$k(x, x';\theta, \nu = 5/2) = \left(1 + \frac{\sqrt{5}|x - x'|}{\theta} + \frac{5(x - x')^2}{3\theta^2}\right) \times \exp\left(-\frac{\sqrt{5}|x - x'|}{\theta}\right), \tag{4}$$

and

$$k(x, x';\theta) = \begin{cases} 1 - 6\left|\frac{x-x'}{\theta}\right|^2 + 6\left|\frac{x-x'}{\theta}\right|^3 & \text{for} \quad 0 \leq \left|\frac{x-x'}{\theta}\right| \leq 0.5 \\ 2\left(1 - \left|\frac{x-x'}{\theta}\right|\right)^3 & \text{for} \quad 0.5 \leq \left|\frac{x-x'}{\theta}\right| \leq 1 \\ 0 & \text{for} \quad \left|\frac{x-x'}{\theta}\right| \geq 1 \end{cases}. \tag{5}$$

A multidimensional kernel is constructed simply as the product of one-dimensional kernels reads as

$$k(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}) = \prod_{i=1}^{m} k(x_i, x_i';\theta_i). \tag{6}$$

Let us define the $n \times n$ correlation matrix $\mathbf{R}$ with its i, j-th component $R_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)};\boldsymbol{\theta})$, where $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are the subset of $\mathcal{X}$. We can also define the correlation vector between an arbitrary point $\mathbf{x}$ and $\mathcal{X}$, that is, $\mathbf{r}(\mathbf{x}) = \{k(\mathbf{x},\mathbf{x}^{(1)};\boldsymbol{\theta}), k(\mathbf{x},\mathbf{x}^{(2)};\boldsymbol{\theta}), \ldots, k(\mathbf{x},\mathbf{x}^{(n)};\boldsymbol{\theta})\}$.

This paper uses a constant μ as the trend function (i.e., the ordinary Kriging). The trend function is set to a constant because our aim is to investigate the effect of the kernel, while setting the trend function to a simple but useful form. In this regard, the most commonly used trend function in engineering applications is an estimated constant. It is worth noting that using automatic trend selection, such as in polynomial chaos Kriging (Kersaudy et al. 2015), will lead to difficulty in investigating the impact of the kernel itself. The prediction structure for GP with constant mean reads as

$$\hat{y}(\mathbf{x}) = \mu + \mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}(\mathbf{y} - \mu\mathbf{1}), \tag{7}$$

where $\mathbf{1}$ is a vector of ones with length n. Due to its probabilistic nature, GP also outputs the mean-squared error of the prediction which reads as

$$\hat{s}^2(\mathbf{x}) = \sigma^2(1 - (\mathbf{r}(\mathbf{x})^T \mathbf{R}^{-1}\mathbf{r}(\mathbf{x})) + (1 - \mathbf{1}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}))^2(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}). \tag{8}$$

One important step in GP is to calibrate the set of hyperparameters, which include $\boldsymbol{\theta}, \sigma^2$, and μ. Let us denote the set of hyperparameters as $\boldsymbol{\gamma} = \{\boldsymbol{\theta}, \sigma^2, \mu\}$. The calibration is done by performing the following optimization:

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathcal{L}(\boldsymbol{\gamma}), \tag{9}$$

where $\mathcal{L}(\boldsymbol{\gamma})$ is the likelihood function defined as

$$\mathcal{L}(\boldsymbol{\gamma}) = \frac{1}{\sqrt{(2\pi\sigma^2)^{n/2}|\mathbf{R}(\boldsymbol{\theta})|}} \exp\left(-\frac{1}{2}\frac{(\mathbf{y} - \mu\mathbf{1})^T\mathbf{R}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mu\mathbf{1})}{\sigma^2}\right). \tag{10}$$

According to the maximum likelihood estimation, $\hat{\mu}$ can be analytically computed by

$$\hat{\mu} = (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{y}. \tag{11}$$

If the model interpolates the data, the Kriging variance can be analytically computed as follows:

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mu\mathbf{1})^T\mathbf{R}(\boldsymbol{\theta})^{-1}(\mathbf{y} - \mu\mathbf{1}). \tag{12}$$

Notice that there is no analytical formulation of σ^2 for a regression GP, so it needs to be tuned together with the lengthscales. The regression term λ is added to the correlation matrix such that it becomes $\mathbf{R} + \lambda\mathbf{I}$, where $\mathbf{I}$ is an $n \times n$ identity matrix. In our experiments, the regression term is tuned or set to the smallest value that keeps the calculation stable (Ranjan et al. 2011). For tunable λ case, the regression term should be set as one hyperparameter (Bostanabad et al. 2018), i.e., we have $\boldsymbol{\gamma} = \{\boldsymbol{\theta}, \sigma^2, \mu, \lambda\}$. The lower and upper bounds of λ are set to 10^{-12} and 10^{-1}, respectively (the optimization is performed in a log scale). In this paper, the hyperparameters are optimized using a genetic algorithm with a population size of 100 and 300 generations, restarted five times to increase the chance of discovering the global optimum. Each genetic algorithm run is also followed by a local optimization using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm to exploit the solution found by GA. The combination of genetic algorithm and L-BFGS-B for hyperparameter optimization has been applied with success in some previous publications (Palar and Shimoyama 2018, 2019).

3 Bi-fidelity Gaussian process regression

One advantage of GP is that it can be conveniently modified to take MF data into account. The basic concept is to fuse information from real and/or computational experiments with various levels of fidelity. In general, the high-fidelity experiment is more expensive (in time and/or money) but more accurate than the low-fidelity experiment. The main objective is to accurately estimate the high-fidelity responses, with the low-fidelity experiments primarily help in capturing the general trend of the high-fidelity function. It is worth noting that there are several variants of MF GP. However, in this paper, we only use the recursive version of the GP with the linear autoregressive model.

3.1 Recursive formulation

As the name suggests, the recursive version of MF GP uses a recursive framework to fuse data from multiple levels of fidelity. The explanation of the recursive autoregressive model that follows assumes that there are l fidelity levels, where $l > 1$. However, we only use $l = 2$ in our MF model (i.e., bi-fidelity model).

For l levels of fidelity, the recursive framework builds l independent GP models for $t = 2, \ldots, l$, where $t = 1$ is the least accurate fidelity. The main concept of recursive GP is based on the classical autoregressive model by Kennedy and O-Hagan, which reads as follows:

$$\begin{cases} y_t(\mathbf{x}) = \rho_{t-1}(\mathbf{x})y_{t-1}(\mathbf{x}) + \delta_t(\mathbf{x}), \\ y(\mathbf{x}) \perp \delta_t(\mathbf{x}) \end{cases} \tag{13}$$

where $\perp$ indicates an independence relationship, $\delta_t(\mathbf{s})$ is a GP that is independent of Gaussian processes in other levels, i.e., $y_{t-1}(\mathbf{x}), \ldots, y_1(\mathbf{x})$, and the term $\rho_{t-1}(\mathbf{x})$ denotes the

degree of correlation and the scale factor between two successive models (i.e., $y_t(\mathbf{x})$ and $y_{t-1}(\mathbf{x})$).

Let us denote $\mathbf{y}^{(t-1)} = (\mathbf{y}_1, \ldots, \mathbf{y}_{t-1})$ as the responses at the experimental design $(\boldsymbol{\mathcal{X}}_i)_{i=1,\ldots,t-1}$, for $t = 2, \ldots, l$ we then have

$$\hat{y}_t(\mathbf{x}) = \rho_{t-1}(\mathbf{x})\hat{y}_{t-1}(\mathbf{x}) + \mu_{KR,t} + \mathbf{r}_t(\mathbf{x})^T\mathbf{R}_t^{-1} (\mathbf{y}_t - \rho_{t-1}(\boldsymbol{\mathcal{X}}_t) \odot \mathbf{y}_{t-1}(\boldsymbol{\mathcal{X}}_t) - \mu_{KR,t}\mathbf{1}) \tag{14}$$

and

$$\hat{s}_t^2(\mathbf{x}) = \rho_{t-1}^2(\mathbf{x})\hat{s}_{t-1}^2(\mathbf{x}) + \sigma_t^2\big(1 - \mathbf{r}_t(\mathbf{x})^T\mathbf{R}_t^{-1}\mathbf{r}_t(\mathbf{x}) + \big(1 - \mathbf{1}^T\mathbf{R}_t^{-1}\mathbf{r}_t(\mathbf{x})\big)^2\big(\mathbf{1}^T\mathbf{R}_t^{-1}\mathbf{1}\big)^{-1}\big), \tag{15}$$

with the condition that the low-fidelity experimental design should be the subset of the high-fidelity set, i.e., $\boldsymbol{\mathcal{X}}_t \subset \boldsymbol{\mathcal{X}}_{t-1}$. It can be seen that the mean and the variance of the GP at level t (i.e., $y_t(\mathbf{x})$) are functions of the corresponding mean and variance at the corresponding lower level (i.e., $y_{t-1}(\mathbf{x})$), which makes the framework recursive. Compared to the original Kennedy and O'Hagan formulation (Kennedy and O'Hagan 2000), due to its recursive nature, the recursive approach provides the GP models of all fidelity levels. In this paper, the scale factor is defined as a constant throughout the input space, that is, $\rho_{t-1}(\boldsymbol{x}) = \rho_{t-1}$. In the subsequent discussions, we primarily use the term bi-fidelity since our focus is on modeling with two levels of fidelity.

3.2 Composite kernel for recursive co-Kriging

The primary objective of this paper is to investigate the potential of CKL for bi-fidelity modeling via GP in engineering design analysis. Consider K non-identical kernels, i.e., $k_i(\mathbf{x},\mathbf{x}')$ where $i = 1, 2, \ldots, K$, we can combine these kernels into a single kernel through the following weighted-sum formulation:

$$k_{ckl}(\mathbf{x},\mathbf{x}';\boldsymbol{\theta},\boldsymbol{w}) = \sum_{i=1}^{K} w_i k_i(\mathbf{x},\mathbf{x}';\boldsymbol{\theta}), \tag{16}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_K)$ is the weight vector for CKL such that $\sum_{i=1}^{K} w_i = 1$. The equality constraint ensures that the resulting kernel is a valid correlation function, i.e., $k_{ckl}(\mathbf{x},\mathbf{x}) = 1$. One can create various kernels by tuning the weights in Eq. (16). Notice that Eq. (16) assumes that all individual kernels share the same values of lengthscales $\boldsymbol{\theta}$. Using variable lengthscales that differ for each kernel is possible, but it increases the risk of overfitting due to a large number of tunable hyperparameters (Satria Palar et al. 2020).

Combining kernels will alter the correlation matrix and affect the value of the likelihood function. The combined correlation matrix reads as

$$\mathbf{R}_{ckl}(\boldsymbol{\theta},\boldsymbol{w}) = w_1\mathbf{R}_1 + w_2\mathbf{R}_2 + \cdots + w_K\mathbf{R}_K. \tag{17}$$

Since $\boldsymbol{w}$ is treated as a part of hyperparameters, the likelihood function now also takes $\boldsymbol{w}$ as its argument. However, satisfying $\sum_{i=1}^{K} w_i = 1$ is not a trivial task. Following (Satria Palar et al. 2020), we use a set of dummy variables $\boldsymbol{z} = \{z_1, z_2, \ldots, z_k\}$, where $w_j = z_j / \sum_{i=1}^{K} z_i$ for $j = 1, 2, \ldots, K$ and the range for each z_j is set to [0, 1]. This approach ensures that the equality constraint is satisfied and results in an unconstrained hyperparameter optimization with $n + 3 + K$ parameters to tune. The objective is now to optimize $\mathcal{L}(\boldsymbol{\gamma}_{\text{CKL}})$, where $\boldsymbol{\gamma}_{\text{CKL}} = \{\boldsymbol{\theta}, \sigma^2, \mu, \lambda, \boldsymbol{z}\}$. Notice that λ is excluded from the hyperparameter set for the stable λ case.

Our CKL implementation uses four constituents kernels: Gaussian, Matern-3/2, Matern-5/2, and cubic (thus, $K = 4$). This choice of the kernel is based on the types of kernel frequently used in engineering design and analysis. Notice that the authors also use the exponential kernel in their single-fidelity implementation of CKL. However, we do not use an exponential kernel in our current implementation of CKL in MF GP since it has been shown that the exponential kernel is not essential in single-fidelity CKL (Satria Palar et al. 2020). In contrast, the other four kernels frequently contribute to the new kernel yielded by the hyperparameter optimization in CKL. Therefore, removing inessential kernels is advantageous because it decreases the training time of CKL.

The implementation of CKL within a bi-fidelity GP framework is relatively straightforward. Because the recursive model builds two different GP models, each model is optimized with respect to its independent set of hyperparameters. A separate CKL procedure can then be applied for the two models, each with a different level of fidelity. CKL can then find different optimal weighting of kernels for different fidelity levels. Thus, the CKL method can potentially further improve the accuracy and robustness of a bi-fidelity GP model.

3.3 Investigated issues

The first research question is how important is the choice of the kernel function in the context of bi-fidelity GP for engineering problems. This aspect is often overlooked, even though the choice of kernel function might impact the accuracy of bi-fidelity GP. From our experience with single-fidelity GP models, it is natural to assume that kernel selection is crucial in bi-fidelity GP. Our research aims to shed light on the issue through experiments on four engineering test problems. We use a single identical kernel for both fidelity levels to simplify the investigation while retaining meaningful comparison. For example, a comparison of Gaussian and cubic kernel means that the bi-fidelity GP model is built by using Gaussian on both the low- and high-fidelity levels, while the second model uses cubic on both levels.

The second issue we investigated, considering the difficulty in the manual selection of the kernel, is the automatic selection of the kernel by a specific criterion. The idea is as simple as building multiple bi-fidelity GP models, each with different kernel functions, and then picking the most proper model. The most proper model is assessed through either minimum leave-one-out-cross validation (LOOCV) error or minimum negative log-likelihood value (essentially, the highest likelihood). The LOOCV error is assessed via a quick cross-validation procedure as detailed in Le Gratiet and Garnier (2012). Answering this research question is crucial because it might potentially ease the process of kernel selection. We call this procedure 'model selection' in the subsequent discussions.

The third research question is about the use of CKL in bi-fidelity GP models. Although the implementation of CKL is relatively straightforward, some unresolved research questions need to be tackled regarding the utilization of CKL for bi-fidelity GP. For example, is CKL effective in improving the predictive power of bi-fidelity GP? This is a fundamental research question since the main objective of CKL is to improve its accuracy, but it also incurs an extra computational cost. Hence, the need to verify that the application of CKL leads to actual benefit. Furthermore, we also want to enquire about the effectiveness of CKL compared to model selection, which is more straightforward than the former.

The last question, which is also related to the three previous questions, is to what extent tuning the regression term affects the choice of kernel. Answering this question is particularly important in the context of bi-fidelity modeling of deterministic computer simulation-based problems, which is sometimes assumed to be noiseless. We hypothesized that the choice of the kernel is still important, but its importance would be diminished by the effect of regularization via regression term.

4 Numerical experiment

This paper investigates the open issues mentioned above through numerical experiments on a set of four non-algebraic engineering problems, consisting of (1) a two-variable aeroelastic test problem, (2) a three-variable vibration rig problem, (3) an eight-variable subsonic wing problem, and (4) a ten-variable heat conduction problem. Notice that we only focus on non-analytical problems to make the comparison more meaningful for engineering problems. The accuracy is measured via the normalized root-mean-squared error (denoted as ε), expressed as

$$\varepsilon = \frac{1}{IQR}\sqrt{\frac{1}{n_{val}}\sum_{i=1}^{n_{val}}(f(\mathbf{x}^{(i)}) - \hat{f}(\mathbf{x}^{(i)}))^2}, \tag{18}$$

where IQR is the interquartile range estimated from all the available samples and n_{val} is the size of the validation samples. The NRMSE is calculated on a separate test set of high-fidelity samples. Due to the limited amount of samples, we reshuffled the training and validation set several times for the NRMSE to be averaged. We experimented with various combinations of low- and high-fidelity sample sizes (denoted as n_h and n_l, respectively) and observed the effect on the global accuracy.

As mentioned above, there are four kernels of interest: Gaussian, Matern-3/2, Matern-5/2, and cubic. We are also interested in model selection methods based on the likelihood and LOOCV error values. In addition to accuracy, we also measured the training time of all methods. The performance is measured for the stable and tuned λ case to investigate the impact of tuning the regression term to accuracy.

We denote the single kernel GP as simply the abbreviation of the kernel, i.e., "Gss," "M-3/2," "M-5/2," and "Cub" for the Gaussian, Matern-3/2, Matern-5/2, and cubic kernel, respectively. For example, "M-3/2" denotes the GP that uses the Matern-3/2 kernel for both the high- and low-fidelity levels. The CKL approach is simply denoted as "CKL." Also, we denote "low-fidelity" and "high-fidelity" as LF and HF, respectively. The model selection methods that use LOOCV and maximum likelihood criteria are referred to as "MS-CV" and "MS-ML," respectively. We use the Mann–Whitney U test to check whether one method is statistically significantly better than the other. We also compute the performance score for each method, that is, the number of methods significantly outperformed by the method being investigated. Because there are seven methods to compare (four individual kernels, CKL, and two model selection methods), the maximum achievable performance score is six. Notice that the performance score is computed separately for each combination of low- and high-fidelity sample sizes and for the stable and tuned λ case.

In this paper, we do not focus on the comparison between single- and bi-fidelity GP. However, for the sake of completeness, we show the result from the comparison with the single-fidelity GP with CKL on Appendix 1. All experiments were performed in MATLAB-2019B™ using a personal computer with the following specifications: Intel ® Core™ i5-6200U CPU at 2.40 GHz and with 4 GB of RAM.

4.1 Two-variable Isogai case

The first test problem is the Isogai case which consists of the 2D airfoil NACA 64A010 with two degrees of freedom: pitch and plunge motion (Isogai 1979). The input variables are the Mach number (M) and the flutter speed index (V_{f}), which are evaluated within [0.6, 0.9] and [0.4, 2.0], respectively. The primary objective is to approximate the damping coefficient as a function of M and V_{f}. We evaluated 300 low- and high-fidelity samples using the Euler solver from SU2 (Economon et al. 2016) with 16,937 and 4383 mesh elements for the HF and LF simulations, respectively (see Fig. 1 for the high-fidelity mesh). The time ratio between the low- and high-fidelity solver is about one-fourth. The experiment is repeated 40 times.

The response surface in the Isogai case is evidently nonlinear (see Fig. 2), with the computed sample Pearson correlation coefficient between the low- and high-fidelity simulations equals 0.91. Despite the similarity in the global trend, the disparity between the LF and HF responses of the Isogai case can be clearly seen. For this problem, we experimented with three combinations of low- and high-fidelity sample sizes: (1) $n_h = 20, n_l = 60$, (2) $n_h = 20, n_l = 120$, and (3) $n_h = 40, n_l = 120$.

The NRMSE and the performance score results for the Isogai case are shown in Tables 1 and 2, respectively. First, we can see that tuning the regression term significantly improves the stability of the predictive power. The most notable example is at $n_h = 20/n_l = 60$ case, in which the errors from the bi-fidelity GP with Gaussian, Matern-5/2, and cubic significantly decrease when the regression term is tuned. However, when comparing all individual kernels, the Gaussian kernel still yielded the worst performance, implying that this problem's response surface is not entirely smooth. Another likely cause is the inherently stable nature of Gaussian, even when we already tune the regularization term that improves stability.

The poorer performance of the identical Gaussian kernel is the first evidence showing that kernel choice affects the accuracy of bi-fidelity GP. When the regression term is tuned, the performance of individual kernels is mostly the same, especially when the sample size is small (see the performance scores). Matern-3/2 kernel yields the lowest NRMSE among all identical kernels under the stable λ scenario. However, under the tuned λ scenario, Matern-3/2 only yields significantly better performance for $n_h = 40/n_l = 120$. For this problem, tuning the regression term seems to decrease the importance of the choice of the kernel in terms of accuracy. Upon analyzing the estimated regression term from hyperparameter optimization for the tuned λ case (see Table 3), we observe the tendency of the GPR to create a regression instead of a near-interpolating model, as

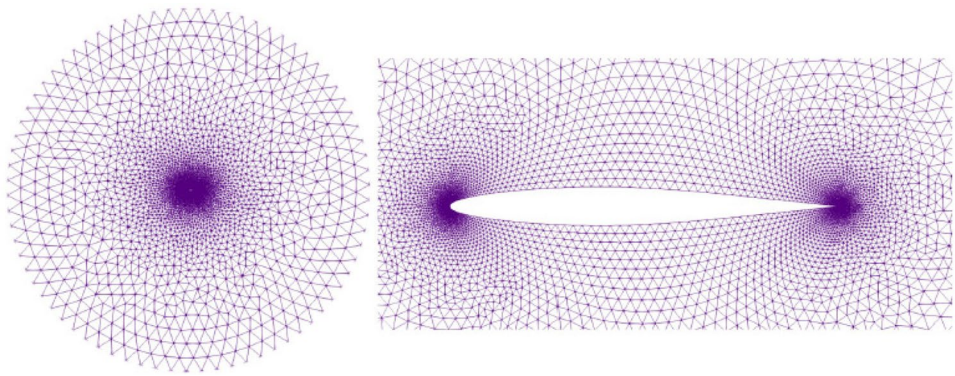
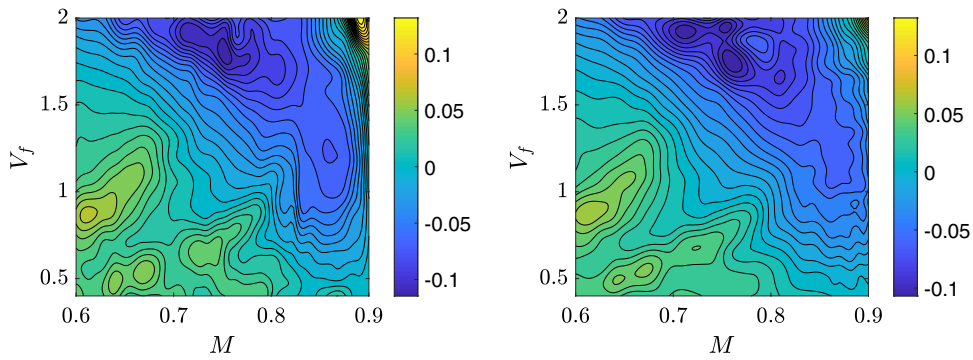


Fig. 1 O-grid mesh for the NACA 64A010 airfoil used in the Isogai test case



(a) High-fidelity (b) Low-fidelity

Fig. 2 Response surfaces of the damping coefficient for the two-variable Isogai case estimated using GPR with 300 samples

Table 1 Averaged NRMSE ($\bar{\varepsilon}$) results for the two-variable Isogai case

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 20, n_l = 60$							
$\bar{\varepsilon}$ (stable λ)	0.2939	0.2546	0.2626	0.2826	**0.2534**	0.2747	0.2838
$\bar{\varepsilon}$ (tuned λ)	0.2638	0.2572	**0.2483**	0.2587	0.2524	0.2535	0.2577
$n_h = 20, n_l = 120$							
$\bar{\varepsilon}$ (stable λ)	0.2490	0.2157	0.2227	0.2311	**0.2142**	0.2270	0.2264
$\bar{\varepsilon}$ (tuned λ)	0.2254	0.2147	0.2131	0.2255	**0.2067**	0.2140	0.2129
$n_h = 40, n_l = 120$							
$\bar{\varepsilon}$ (stable λ)	0.2205	0.1882	0.1957	0.2056	0.1839	0.1949	0.1960
$\bar{\varepsilon}$ (tuned λ)	0.2001	0.1871	0.1906	0.1982	**0.1827**	0.1871	0.1870

Bold values show the lowest mean NRMSE

Table 2 Performance scores (PS) for the two-variable Isogai case

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 20, n_l = 60$							
PS (stable λ)	0	3	1	0	3	0	0
PS (tuned λ)	0	0	0	0	0	0	0
$n_h = 20, n_l = 120$							
PS (stable λ)	0	1	1	1	1	1	1
PS (tuned λ)	0	0	0	0	2	0	0
$n_h = 40, n_l = 120$							
PS (stable λ)	0	2	1	0	2	1	1
PS (tuned λ)	0	1	0	0	2	1	1

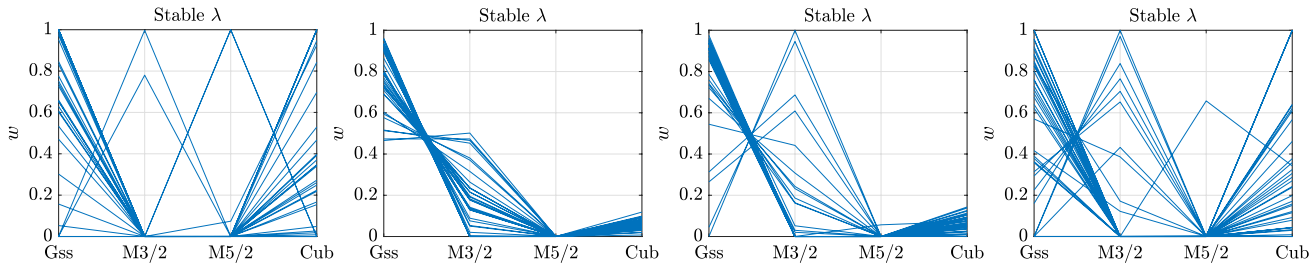
Table 3 Averaged values of the regression term (shown and averaged in the log-scale) estimated from hyperparameter optimization of the HF and LF models for the two-variable Isogai case

Model	Mean of $\log_{10}(\lambda)$				
	Gauss.	M-3/2	M-5/2	Cub.	CKL
HF, $n_h = 20$	− 7.5559	− 9.2493	− 7.7710	− 7.8627	− 10.5225
HF, $n_h = 40$	− 5.8760	− 6.7412	− 6.3373	− 5.8731	− 8.6000
LF, $n_l = 60$	− 4.7193	− 7.8716	− 5.5103	− 5.7152	− 7.2908
LF, $n_l = 120$	− 4.7718	− 7.0024	− 5.5662	− 4.9403	− 6.7291

evidenced by the relatively high values of the regression term. Such a result makes sense since the data are generated from unsteady aerodynamic simulations, in which the corresponding quantities of interest tend to exhibit higher numerical noise than steady simulations.

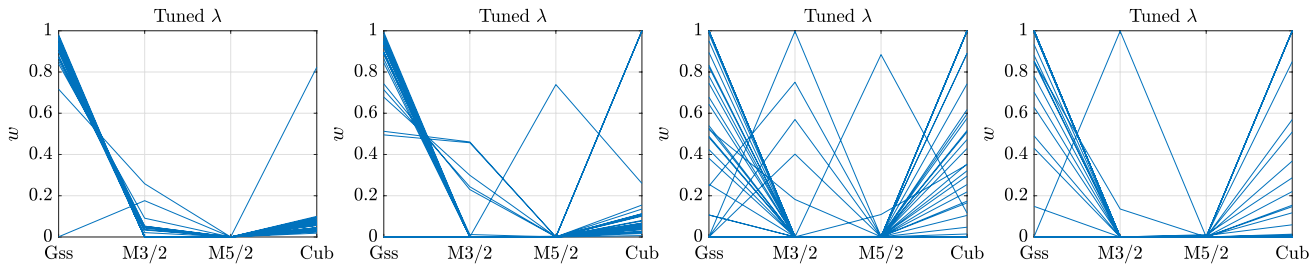
Although CKL is better than the other methods for the stable $\lambda, n_h = 20/n_l = 60$, case, there is no evidence to support the same statement when the regression term is tuned (no single method is significantly better than the other methods since the performance scores are all zero). Interestingly, CKL yields a significantly beneficial effect on higher sample sizes (i.e., $n_h = 20/n_l = 120$ and $n_h = 40/n_l = 120$). According to the single-fidelity result, the CKL approach requires a sufficient sample size to have enough information to construct suitable composite kernels (Satria Palar et al. 2020). Otherwise, CKL tends to mimic the performance of the best-performing individual kernels, which is also a desirable trait. Even though model selection methods are not statistically significantly better than other methods for $n_h = 20/n_l = 60$ and $n_h = 20/n_l = 120$, tuned λ case, they are still useful in the sense that they avoid the choice of poor-performing kernel (in this case, Gaussian and cubic). In addition, we observe no significant difference between the errors from the LOOCV- and likelihood-based model selection.

From Table 4, which shows the frequency of kernel selected according to the LOOCV and likelihood criterion together with that of the lowest actual NRSME, we can see that the best kernel for the stable λ case is the Matern-3/2 as it frequently yielded the lowest actual NRSME. Despite that, the kernel that yields the lowest NRMSE changes from one independent run to the other, although the differences are slight in some cases. It is worth noting that tuning the λ changes the dynamic of the result, in which Matern-5/2 now also plays important role. Regardless, it can be seen that Gaussian is rarely the best kernel that yields the lowest actual NRMSE. The LOOCV error and likelihood values



(a) LF, $n_l = 60$ (b) LF, $n_l = 120$ (c) HF, $n_h = 20$ (d) HF, $n_h = 40$

Fig. 3 Parallel coordinate plots of weights of the composite kernels in CKL from 40 independent runs for the two-variable Isogai case, stable λ case



(a) LF, $n_l = 60$ (b) LF, $n_l = 120$ (c) HF, $n_h = 20$ (d) HF, $n_h = 40$

Fig. 4 Parallel coordinate plots of weights of the composite kernels in CKL from 40 independent runs for the two-variable Isogai case, tuned λ case

Table 4 Frequency of kernel selected according to the LOOCV and likelihood criterion for the two-variable Isogai problem

n_h, n_l	Criterion	Frequency (stable λ/tuned λ)			
		Gaussian	Matern-3/2	Matern-5/2	Cubic
20, 60	LOOCV	7/7	11/14	3/9	9/10
20, 60	Likelihood	12/7	11/21	2/5	15/7
20, 60	Actual NRMSE	3/4	29/9	4/12	4/15
20, 120	LOOCV	4/11	8/8	1/10	27/11
20, 120	Likelihood	5/9	7/10	5/10	23/11
20, 120	Actual NRMSE	3 /6	25/15	6/15	6/4
40, 120	LOOCV	4/2	22/20	0/8	14/10
40, 120	Likelihood	2/4	28/20	1/12	9/4
40, 120	Actual NRMSE	3/6	28/15	8/13	1/6

Table 5 Averaged computational training time of various modeling techniques for the two-variable Isogai case

n_h, n_l	Averaged training time (stable λ/tuned λ)		
	Identical	Model selection	CKL
20, 60	3 s/5 s	10 s/16 s	74 s/116 s
20, 120	6 s/9 s	19 s/35 s	152 s/242 s
40, 120	10 s/12 s	41 s/46 s	242 s/250 s

are a good predictor for the actual NRMSE only for $n_h = 40$ /$n_l = 120$.

Analysis of the contribution of composite kernels shows that Gaussian is the main contributor for both the low- and high-fidelity GP (see Figs. 3 and 4 for the stable and tuned λ case, respectively, which shows the weights in parallel coordinate plots). The parallel coordinate plots become more ordered as the sample size increases, as in the case for $n_l = 120$. Despite some differences, the trend is roughly the same for both stable and tuned λ cases. The composite kernels of the low-fidelity GP mainly consist of the Gaussian, Matern-3/2, and cubic kernels. It is interesting to see that combining these three kernels can produce better kernel functions than the best-performing individual kernel, i.e., the Matern-3/2.

However, the better performance of CKL comes with a high computational cost. Table 5 clearly shows that the CKL is notably more expensive than the identical approaches (about 35–50 times more expensive). As expected, model selection is about four times more expensive than the identical approach. Nevertheless, the long training time of CKL

is justified as it performs better than the model selection. Furthermore, setting λ as one extra hyperparameter increases the training time due to the more complex hyperparameter optimization.

4.2 Eight-variable subsonic wing problem

The next case is the quantification of the aerodynamic performance of an untapered, unswept, and rectangle subsonic wing problem. The airfoil under analysis is NACA2412. The wing has a wingspan of $b = 6$ m and a chord length of $c = 1$ m. The input variables are the twist angles at eight different sections in the semi-span location, i.e., 0%, 20%, 40%, 60%, 80%, 90%, 95%, and 100% from the root chord, where each angle is varied between -10 and 10 degrees of twist. The output of interest is the drag coefficient, calculated using a coupled panel method and vortex wake solver. On the other hand, the low-fidelity solver uses a simple flat wake panel method to compute the coefficients. For this problem, we use the FLOW5 code to simulate both the low- and high-fidelity cases (Cère-Aéro 2022). The high-fidelity simulation is about six times more expensive than the low-fidelity simulation. The experiment is performed with the following combinations: (1) $n_h = 20, n_l = 60$, (2) $n_h = 20, n_l = 120$ and (3) $n_h = 40, n_l = 120$, from 1000 samples that are reshuffled 20 times (Fig. 5).

The NRMSE results and the corresponding performance scores are shown in Tables 6 and 7. In contrast to the Isogai case, tuning the regression term does not significantly increase the model accuracy, although the difference is still noticeable for larger sample sizes. Upon analyzing the values of λ selected by hyperparameter optimization shown in Table 8 (for the tuned λ case), we observe that the HF response surface tends to be smoother than the LF response surface (the estimated regression terms are much smaller for the former). This phenomenon is counter-intuitive since the LF samples are collected from the panel method, which was simple and thought to produce small numerical noises. The result highlights again the importance of tuning the regression term since there is no guarantee that the numerical noise is small for any case. Furthermore, it is also interesting to see that CKL tends to select small values of regression terms for both LF and HF data, indicating the stable nature of CKL.

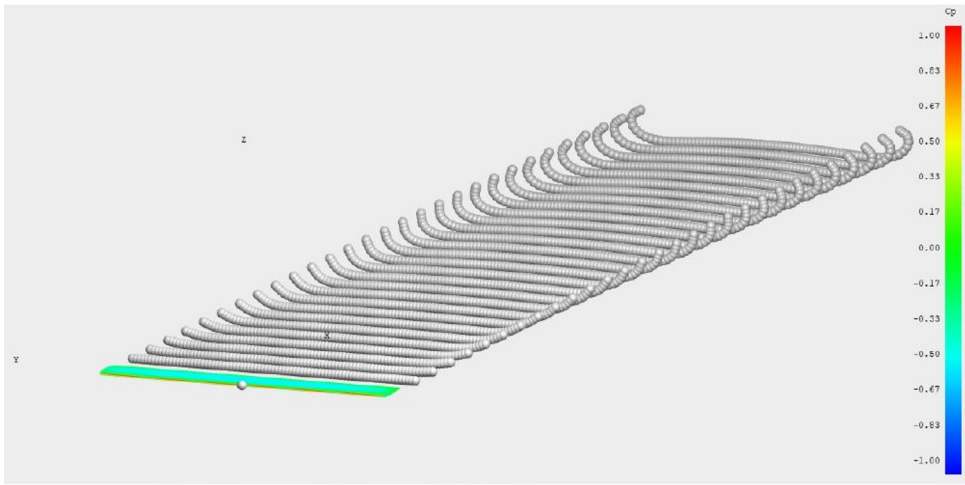


Fig. 5 Depiction of the wing used in the eight-variable subsonic wing problem with the result from the vortex particle wake method

Table 6 Averaged NRMSE ($\bar{\varepsilon}$) results for the eight-variable subsonic wing case

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 20, n_l = 60$							
$\bar{\varepsilon}$ (stable λ)	0.1198	0.1404	0.1307	**0.1187**	0.1207	0.1215	0.1233
$\bar{\varepsilon}$ (tuned λ)	0.1196	0.1424	0.1341	**0.1189**	0.1200	0.1223	0.1228
$n_h = 20, n_l = 120$							
$\bar{\varepsilon}$ (stable λ)	0.0869	0.0916	0.0872	0.0849	**0.0829**	0.0866	0.0869
$\bar{\varepsilon}$ (tuned λ)	0.0836	0.0909	0.0867	0.0831	**0.0797**	0.0848	0.0861
$n_h = 40, n_l = 120$							
$\bar{\varepsilon}$ (stable λ)	0.0831	0.0865	0.0833	0.0801	**0.0782**	0.0823	0.0831
$\bar{\varepsilon}$ (tuned λ)	0.0802	0.0857	0.0820	0.0791	**0.0755**	0.0809	0.0811

Bold values show the lowest mean NRMSE

Table 7 Performance scores (PS) for the eight-variable subsonic wing problem

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 20, n_l = 60$							
PS (stable λ)	2	0	1	2	1	1	1
PS (tuned λ)	2	0	0	2	2	1	1
$n_h = 20, n_l = 120$							
PS (stable λ)	1	0	1	1	3	1	1
PS (tuned λ)	1	0	1	2	5	1	1
$n_h = 40, n_l = 120$							
PS (stable λ)	1	0	0	3	5	1	1
PS (tuned λ)	1	0	1	1	6	1	1

Table 8 Averaged values of the regression term (shown and averaged in the log-scale) estimated from hyperparameter optimization of the HF and LF models for the eight-variable subsonic wing problem

Model	Mean of $\log_{10}(\lambda)$				
	Gauss.	M-3/2	M-5/2	Cub.	CKL
HF, $n_h = 20$	− 10.7425	− 11.4271	− 11.9931	− 11.9681	− 11.9942
HF, $n_h = 40$	− 11.3522	− 11.7472	− 11.0417	− 8.7065	− 11.5127
LF, $n_l = 60$	− 6.3966	− 7.5822	− 6.5493	− 8.8257	− 10.0440
LF, $n_l = 120$	− 4.3358	− 10.3714	− 6.4333	− 7.9927	− 10.4616

For this problem, the Gaussian and cubic functions yield the lowest mean NRMSEs for the identical kernel approach. By analyzing Table 8 again, it can be seen that the numerical noise of the HF data is relatively minimal compared to that of the Isogai case, which makes Gaussian more suitable for this problem. Conversely, Matern kernels are usually recommended for non-smooth cases, which is not always the case. The better performance of cubic compared to Matern kernels is quite surprising since the former is rarely picked as a kernel of choice in many applications.

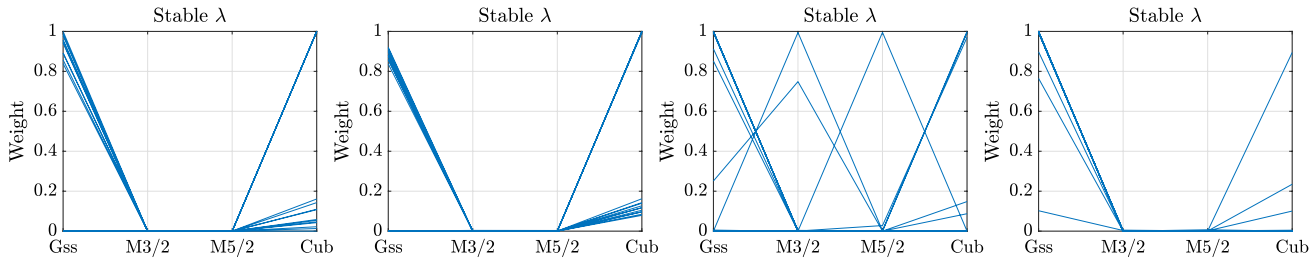
The CKL yields the highest accuracy, as shown by the lowest NRMSE and highest performance scores. That is, the CKL performs similarly with the best-performing kernel on $n_h = 20, n_l = 60$ (i.e., cubic), while outperforming all individual kernels on the $n_h = 20, n_l = 120$ and $n_h = 40, n_l = 120$ case. The comparison of CKL with model selections reveals that the former yields better accuracy, although one might argue that the difference is slight. However, even a small difference might lead to, for example, a faster optimization process. One might argue that the performance difference compared to the individual kernel approach is slight. From such a viewpoint, CKL and model selection can at least be seen as a mechanism to avoid relatively poor-performing kernels. In a sense, CKL successfully built more suitable kernels, making it more efficient than model selection. There is an evident and frequent mismatch between the kernel with the lowest actual NRMSE and that selected by model selection (see Table 9). The upside is that both likelihood and LOOCV criteria tend to avoid Matern-3/2, which is the worst-performing kernel for the subsonic wing problem.

CKL is notably more expensive than the other approaches. Table 10, which shows the averaged training time for the subsonic wing problem, indicates that the training time of CKL ranges from about seven to 20 min. On the other hand, the model selection is only about four times more expensive than the identical approach.

The constituents of the composite kernels for CKL are, essentially, only the Gaussian and cubic (see Figs. 6 and 7). Such a trend reflects that of the individual kernels, in which both Gaussian and cubic outperform the Matern kernels. It then makes sense that the CKL infers this trend from the data to make better kernels. The inclusion of tunable regression term slightly changes the dynamic of kernel compositions. To be exact, the Matern-5/2 kernel also plays a role together with Gaussian and cubic, with the Matern-3/2 kernel still excluded from the composition. Combination of kernel functions lead to the combined characteristics of the constituents (see Appendix 2, which discusses the sample paths generated from composite kernels). The likely explanation is that combining the three kernels yield a fine mixture between the smoothness of Gaussian and the more rough characteristics of the other kernels, which eventually matches the behavior of this problem. The total effect is such that the accuracy is better than identical kernel and model selection approaches.

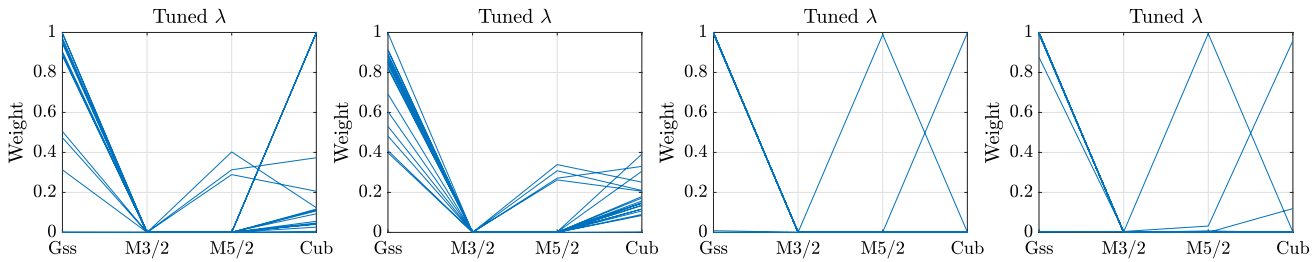
4.3 Three-variable vibration rig problem

The next problem is a three-variable vibration rig problem, which is unique because the high-fidelity data were evaluated from a physical experiment. The low fidelity is sampled from a numerical model to complement the high-fidelity



(a) LF, $n_l = 60$ (b) LF, $n_l = 120$ (c) HF, $n_h = 20$ (d) HF, $n_h = 40$

Fig. 6 Parallel coordinate plots of weights of the composite kernels in CKL from 20 independent runs for the eight-variable subsonic wing problem, stable λ case



(a) LF, $n_l = 60$ (b) LF, $n_l = 120$ (c) HF, $n_h = 20$ (d) HF, $n_h = 40$

Fig. 7 Parallel coordinate plots of weights of the composite kernels in CKL from 20 independent runs for the eight-variable subsonic wing problem, tuned λ case

Table 9 Frequency of kernel selected according to the LOOCV and likelihood criterion for the eight-variable subsonic wing problem

n_h, n_l	Criterion	Frequency (stable λ/tuned λ)			
		Gaussian	Matern-3/2	Matern-5/2	Cubic
20, 60	LOOCV	10/4	1/1	2/4	7/11
20, 60	Likelihood	10/7	1/2	1/3	8/8
20, 60	Actual NRMSE	11/11	0/0	0/0	9/9
20, 120	LOOCV	14/12	2/2	1/3	3/3
20, 120	Likelihood	16/13	0/2	2/3	2/2
20, 120	Actual NRMSE	2/5	0/1	4/1	14/13
40, 120	LOOCV	6/9	0/2	6/2	8/7
40, 120	Likelihood	9/8	0/1	2/4	9/7
40, 120	Actual NRMSE	5/7	0/2	1/3	14/8

Table 10 Averaged computational training time of various modeling techniques for the eight-variable subsonic wing problem

n_h, n_l	Averaged training time (stable λ/tuned λ)		
	Identical	Model selection	CKL
20, 60	55 s/67 s	201 s/292 s	461 s/477 s
20, 120	103 s /112 s	417 s/438 s	909 s/1036 s
40, 120	128 s/154 s	524 s/558 s	1112 s/1241 s

data. For this problem, the output of interest is the peak value of the radial vibrations as measured from the accelerometers on the bearing of a test rig.

The test rig simulates a simple industrial application composed of a motor and a user. The two parts are rigidly connected through a shaft, supported by roller bearings in a self-alignment configuration. The testing configurations can vary depending on the rpm, the global unbalance, and the load acting on the system. The motor rpm is controlled by an inverter that keeps the rotational velocity constant at the selected operational speeds (4 conditions); the unbalances and the loads are modified by varying the amount of dedicated masses, respectively, on the rotor (3 conditions) and the shaft (3 conditions). The measurements explored the full range of operational conditions, with a total of 58 sets containing the raw and processed data of two accelerometers positioned on the bearing in the radial direction. The vibration data were collected with a Dewesoft Sirius acquisition system with a sampling frequency of 5kHz, in an operational range not containing the system's resonances. As a first approximation, the system can be represented by

Table 11 Averaged NRMSE ($\bar{\varepsilon}$) results for the three-variable vibration rig problem

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 24, n_l = 112$							
$\bar{\varepsilon}$ (stable λ)	0.2856	0.2474	0.2598	0.2893	**0.2340**	0.2600	0.2575
$\bar{\varepsilon}$ (tuned λ)	**0.2160**	0.2261	0.2237	0.2225	0.2200	0.2223	0.2231
$n_h = 30, n_l = 168$							
$\bar{\varepsilon}$ (stable λ)	0.3001	0.2618	0.2764	0.3152	**0.2440**	0.2818	0.2766
$\bar{\varepsilon}$ (tuned λ)	**0.2142**	0.2256	0.2216	0.2314	0.2238	0.2193	0.2234

Bold values show the lowest mean NRMSE

Table 12 Performance scores (PS) for the three-variable vibration rig problem

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 24, n_l = 112$							
PS (stable λ)	0	2	2	0	5	2	2
PS (tuned λ)	0	0	0	0	0	0	0
$n_h = 20, n_l = 120$							
PS (stable λ)	0	2	2	0	5	1	2
PS (tuned λ)	2	0	0	0	0	1	0

an SDOF (Single Degree Of Freedom) system, in which a rotating unbalance excites a mass supported by stiffness and damping elements. The numerical model of such a system has been implemented in Simulink and represents the low-fidelity numerical model. Interested readers are referred to Palar et al. (2022) for more details regarding this problem.

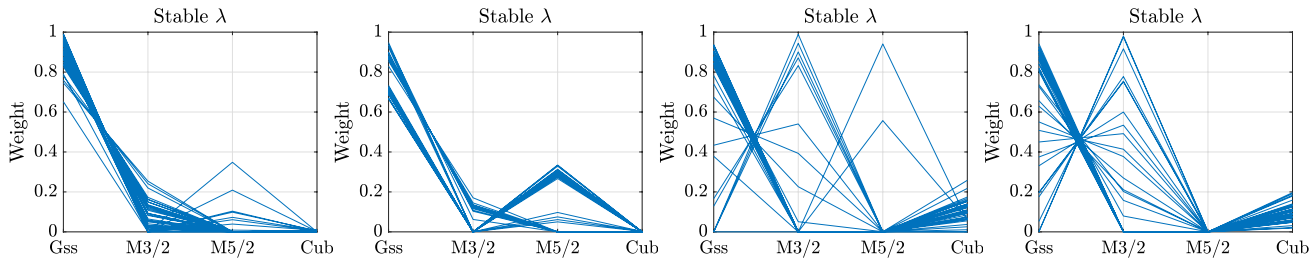
The first challenge of this problem is the low correlation between the low- and high-fidelity data. To be exact, the Pearson correlation coefficient between the low- and high-fidelity sample is 0.68. The second challenge is experimental noise, which corrupts the high-fidelity data, making approximating this problem more difficult. For this problem, we experimented with the two following combinations of low- and high-fidelity sample size: $n_h = 24, n_l = 112$ and $n_h = 30, n_l = 168$. Using cost ratio is not relevant here since we use physical and computer experiment for collecting the high- and low-fidelity data, respectively. The experiment is repeated 40 times, each with a different set of randomly reshuffled sampling points.

Although we are aware that the high-fidelity data are corrupted by experimental noise, we still performed the experiment with stable λ for the sake of completeness. The results are shown in Tables 11 and 12. First, it can be seen that the obtained accuracy is notably lower than the other problems due to the complexity of the high-fidelity data, despite the relatively abundant amount of low-fidelity data. However, the bi-fidelity framework still yields lower errors than the pure HF model (see Appendix 1). Such a trend provides evidence regarding the usefulness of the bi-fidelity framework for these data despite the low Pearson correlation (although it should be noted here that this statement only applies to the current sample size).

Table 13 Averaged values of the regression term (shown and averaged in the log scale) estimated from hyperparameter optimization of the HF and LF models for the three-variable vibration rig problem

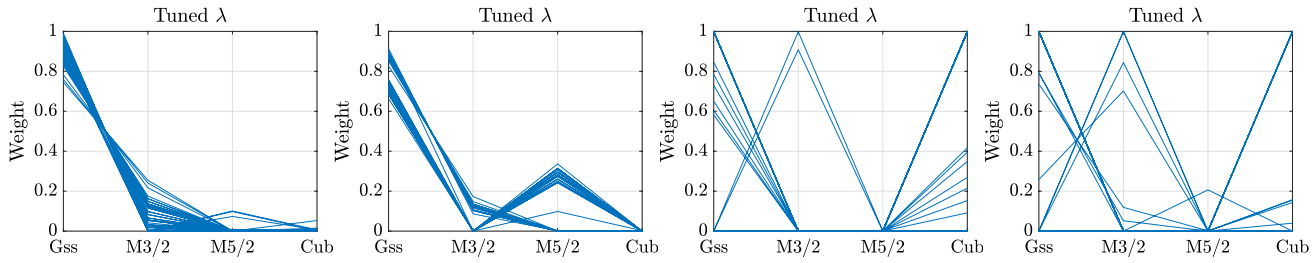
Model	Mean of $\log_{10}(\lambda)$				
	Gauss.	M-3/2	M-5/2	Cub.	CKL
HF, $n_h = 24$	− 1.8109	− 1.7538	− 1.6107	− 1.3348	− 2.8972
HF, $n_h = 30$	− 1.2605	− 1.3070	− 1.5188	− 1.2586	− 1.6081
LF, $n_l = 112$	− 6.3608	− 11.1464	− 9.3688	− 8.7662	− 11.7828
LF, $n_l = 168$	− 7.7731	− 11.6224	− 10.8683	− 10.6291	− 11.9989

Even though CKL is notably better than other methods for the stable λ, such a better performance is not useful in a practical sense since the experimental data are corrupted with noise. As expected, tuning the λ significantly decreases the error level. In fact, there is no statistically significant difference between the performance of all methods when the regression term is tuned for the $n_h = 24/n_l = 112$ case. However, the Gaussian kernel is notably better than other methods for the $n_h = 30/n_l = 168$, tuned λ, case. The superior accuracy of Gaussian is somehow unforeseen since the use of Gaussian itself is not recommended for noisy data (which is the case for experimental data). Furthermore, in contrast to the results of the other problems, the CKL yields no extra observable benefits in terms of accuracy. The presence of large noise in the experimental data, as evidenced by the high value of



(a) LF, $n_l = 112$ (b) LF, $n_l = 168$ (c) HF, $n_h = 24$ (d) HF, $n_h = 30$

Fig. 8 Parallel coordinate plots of weights of the composite kernels in CKL from 40 independent runs for the three-variable vibration rig problem, stable λ case



(a) LF, $n_l = 112$ (b) LF, $n_l = 168$ (c) HF, $n_h = 24$ (d) HF, $n_h = 30$

Fig. 9 Parallel coordinate plots of weights of the composite kernels in CKL from 40 independent runs for the three-variable vibration rig problem, tuned λ case

Table 14 Frequency of kernel selected according to the LOOCV and likelihood criterion for the three-variable vibration rig problem

n_h, n_l	Criterion	Frequency (stable λ/tuned λ)			
		Gaussian	Matern-3/2	Matern-5/2	Cubic
24, 112	LOOCV	5/16	24/9	6/7	5/8
24, 112	Likelihood	4/17	31/2	2/4	3/17
24, 112	Actual NRMSE	5/17	29/3	4/9	2/11
30, 168	LOOCV	9/25	24/4	2/9	5/2
30, 168	Likelihood	6/26	31/3	1/3	2/8
30, 168	Actual NRMSE	5/24	29/5	6/6	0/5

Table 15 Averaged computational training time of various modeling techniques for the three-variable vibration rig problem

n_h, n_l	Averaged training time (stable λ/tuned λ).		
	Identical	Model selection	CKL
24, 112	24 s/32 s	95 s /137 s	238 s/315 s
30, 168	29 s /41 s	120 s/161 s	525 s/736 s

λ shown in Table 13, seems to decrease the importance of the proper choice of kernel. Finally, both model selection methods cannot outperform the best-performing individual kernels, implying that both LOOCV and likelihood are not excellent predictors of the actual error for this problem.

Table 14 shows the frequency of kernels for the vibration rig problem. Interestingly, the Matern-3/2 kernel frequently yields the lowest actual NRMSE for the stable λ case. Conversely, Gaussian is the most proper kernel for the tuned λ case. The fact that model selection with both LOOCV and likelihood criterion frequently selects the Gaussian kernel for the tuned λ case is appreciable. Tuning the regression term interestingly also changes the dynamic of kernel selection. We also observe that the model selection did not perform similarly to Gaussian.

The mixture of the low-fidelity kernel in CKL tends to pick Gaussian and Matern kernels as its constituents (see Figs. 8 and 9 for the stable and tuned λ case, respectively). On the other hand, although the plots are still quite cluttered due to the relatively low high-fidelity sample size, the CKL tends to mix primarily Gaussian, Matern-3/2, and cubic kernels. Despite this effort, CKL is still no better than the best-performing kernel. There are two possible causes for the ineffectiveness of CKL and model selection in this problem. Firstly, the noisy nature of the experimental data (i.e., the high-fidelity data) makes this problem harder

to approximate. Secondly, the relatively low correlation between the low- and high-fidelity data might also contribute to the ineffectiveness of CKL. As shown in Table 15, the very high computational cost of CKL compared to the single kernel approach does not translate to better accuracy for this problem.

4.4 Ten-variable heat conduction problem

Finally, we consider a heat conduction problem with random diffusivity coefficient in a square-shaped domain $D = (-0.5, 0.5) \times (-0.5, 0.5)\,\mathrm{m}^2$, as in Konakli and Sudret (2016) and Zuhal et al. (2021). The problem sketch is shown in Fig. 10. The governing PDE is the steady-state heat equation $-\nabla(\kappa(z)\nabla T(z)) = QI_A(z)$, with the Dirichlet boundary condition $T = 0\,^\circ\mathrm{C}$ at the top of the domain and Neumann boundary condition on the other sides. The domain A produces a heat source $Q = 2 \times 10^3 \mathrm{W/m}^3$ within D, where $A = (0.2, 0.3) \times (0.2, 0.3)\mathrm{m}^2$ and the quantity of interest is the average temperature at domain $B = (-0.3, -0.2) \times (-0.3, -0.2)\mathrm{m}^2$.

The spatial randomness of the conduction coefficient is characterized by a lognormal random field with expansion optimal linear estimation (EOLE) discretization:

$$\kappa(z) = \exp[a_\kappa + b_\kappa g(z)], \tag{19}$$

where a_κ and b_κ denote the mean and standard deviation of κ which are $\mu_\kappa = 1\mathrm{W}/(\mathrm{m}\,^\circ\mathrm{C})$ and $\sigma_\kappa = 0.3\mathrm{W}/(\mathrm{m}\,^\circ\mathrm{C})$, respectively, and $g(z)$ is the Gaussian random field. The size of the EOLE grid is set to 0.1m and $\ell = 0.2\mathrm{m}$. The expansion is truncated to the first ten bases, in which the respective coefficients are taken as the random input variables with standard normal distribution.

The low-fidelity simulation uses a 30×30 grid, while the high-fidelity samples use 100×100 grid to discretize the domain. We use our in-house finite-difference solver to solve the heat conduction equation, with the cost of the LF and HF simulation being 0.4 and 12 s, respectively (i.e., cost ratio of 30). The Pearson correlation between the two responses is strong, yielding a value of about 0.99, but the LF value underestimates that of HF. For this problem, we experimented with four combinations of LF and HF samples, namely, (1) $n_h = 10/n_l = 50$, (2) $n_h = 10/n_l = 100$, (3) $n_h = 20/n_l = 100$, and (4) $n_h = 40/n_l = 120$. The sample sizes are adjusted to investigate the effect of LF and HF experimental design size on the accuracy. Due to the expensive training time, we only repeated the experiment 20 times, each with random samples reshuffled from the available 10,000 samples.

The NRMSE results and the performance scores are shown in Tables 16 and 17. In contrast to the previous three problems, tuning the regression term decreased the accuracy of the model; the high mean averaged errors are due to the poor-performing outliers. Such a low accuracy primarily stems from the high dimensionality of the problem, which makes the models with tunable λ find it difficult to differentiate between the data and the noise. Hence, an interpolating model is more suitable for approximating high-dimensional problems.

Individual kernel wise, the result indicates that Matern-5/2 and Gaussian yield the most accurate performance, although one might argue that the difference with the other kernels is slight. The effect of kernel starts to take effect when the high-fidelity sample size is increased to $n_h = 20$ and higher. It is interesting to see that, in contrast to all previous problems, the pair of the best-performing kernel now comprises Gaussian and Matern-5/2. The Matern-5/2 kernel yields the best performance for the $n_h = 20/n_l = 100$ and $n_h = 40/n_l = 120$ case.

We observe that the CKL, LOOCV-based kernel selection, and likelihood-based kernel selection successfully improve the general robustness of bi-fidelity GP for the stable λ case. Furthermore, there is also no significant difference in accuracy between the three methods, although CKL yields the lowest overall NRMSE for high sample sizes. Given the computational cost of CKL, as shown in Table 20, model selection based on either LOOCV or likelihood is

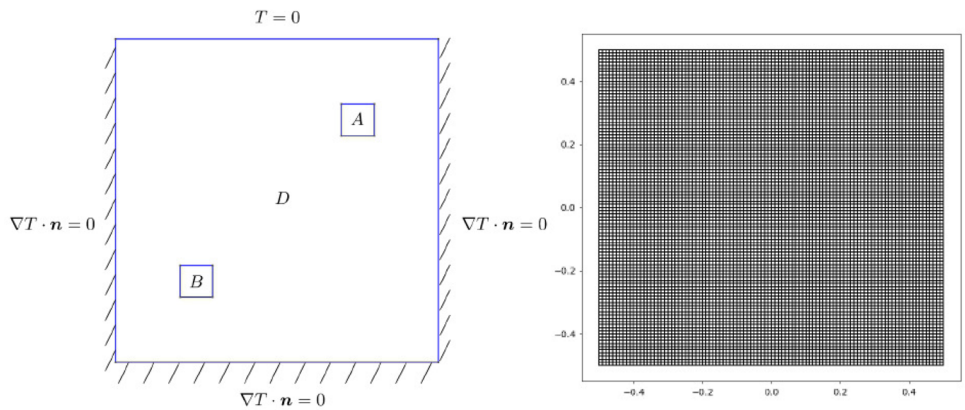


Fig. 10 Computational domain for the 10-variable heat conduction problem (left) and the finite-difference mesh for the HF simulation (100×100 grid)

Table 16 Averaged NRMSE ($\bar{\varepsilon}$) results for the ten-variable heat conduction problem

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 10, n_l = 50$							
$\bar{\varepsilon}$ (stable λ)	0.0968	0.1102	0.0992	0.1069	**0.0962**	0.0974	0.0966
$\bar{\varepsilon}$ (tuned λ)	0.3477	0.2503	0.2767	0.6622	**0.2412**	0.4168	0.4186
$n_h = 10, n_l = 100$							
$\bar{\varepsilon}$ (stable λ)	0.0842	0.0857	0.0826	0.0823	0.0835	0.0818	**0.0815**
$\bar{\varepsilon}$ (tuned λ)	**0.0828**	0.1883	0.0865	0.3281	0.1203	0.1880	0.1908
$n_h = 20, n_l = 100$							
$\bar{\varepsilon}$ (stable λ)	0.0599	0.0611	0.0513	0.0621	**0.0403**	0.0432	0.0413
$\bar{\varepsilon}$ (tuned λ)	0.1361	0.1386	0.1343	0.5705	**0.1026**	0.1187	0.1181
$n_h = 40, n_l = 120$							
$\bar{\varepsilon}$ (stable λ)	0.0254	0.0343	0.0205	0.0278	**0.0188**	0.0204	0.0197
$\bar{\varepsilon}$ (tuned λ)	**0.0445**	0.1235	0.0487	0.5996	0.0786	0.0615	0.0616

Bold values show the lowest mean NRMSE

Table 17 Performance scores (PS) for the ten-variable heat conduction problem

	Gauss.	M-3/2	M-5/2	Cub.	CKL	MS-CV	MS-ML
$n_h = 10, n_l = 50$							
PS (stable λ)	1	0	0	0	1	0	1
PS (tuned λ)	1	1	1	0	1	0	0
$n_h = 10, n_l = 100$							
PS (stable λ)	0	0	0	0	0	0	0
PS (tuned λ)	1	0	1	0	1	0	0
$n_h = 20, n_l = 100$							
PS (stable λ)	0	0	1	0	3	2	3
PS (tuned λ)	1	1	1	0	1	2	2
$n_h = 40, n_l = 120$							
PS (stable λ)	2	0	2	1	6	2	2
PS (tuned λ)	2	2	1	0	2	3	3

Table 18 Averaged values of the regression term (shown and averaged in the log scale) estimated from hyperparameter optimization of the HF and LF models for the ten-variable heat conduction problem

Model	Mean of $\log_{10}(\lambda)$				
	Gauss.	M-3/2	M-5/2	Cub.	CKL
HF, $n_h = 10$	− 6.6040	− 5.2308	− 5.4847	− 5.5912	− 5.7331
HF, $n_h = 20$	− 6.1651	− 6.8841	− 6.7213	− 4.8115	− 6.8031
HF, $n_h = 40$	− 6.4420	− 7.6324	− 5.3938	− 4.9965	− 6.8913
LF, $n_l = 50$	− 7.9041	− 9.8201	− 8.8966	− 7.8141	− 9.0158
LF, $n_l = 100$	− 7.4488	− 11.9261	− 7.3064	− 9.1443	− 9.0119
LF, $n_l = 120$	− 5.2768	− 11.6214	− 6.9564	− 7.4757	− 7.3334

then preferable to CKL for this problem. Analysis of the estimated regression terms (see Table 18) shows that the values of λ are quite high for the high-fidelity model. However, it is not likely that the high-fidelity data are corrupted with noise, but rather the low accuracy is due to the combination of small sample size and high input dimensionality. It can be seen that the estimated λ is lower for the low-fidelity data, thanks to the large sample size.

The parallel coordinate plots for the weights of the composite kernel are shown in Figs. 11 and 12 for the stable and tuned λ case, respectively. The combination of kernels seems to favor the Gaussian kernel as the main constituent, as shown primarily in Fig. 11. However, notice that the mixture is never perfectly Gaussian. The weights are also disordered for low sample sizes since there is only a few available information for CKL. Furthermore, the plots are more disordered for the tuned λ case. Although CKL yields models with high accuracy, remember that a similar effect can be obtained by proper kernel selection, in which LOOCV and likelihood-based selection proved useful for such a task. Table 19 shows that the best individual kernel with the lowest actual NRMSE might differ for each independent run and sample size, with primarily Gaussian and Matern-5/2 alternately becoming the best for both the stable and tuned λ case. It is also interesting that both LOOCV

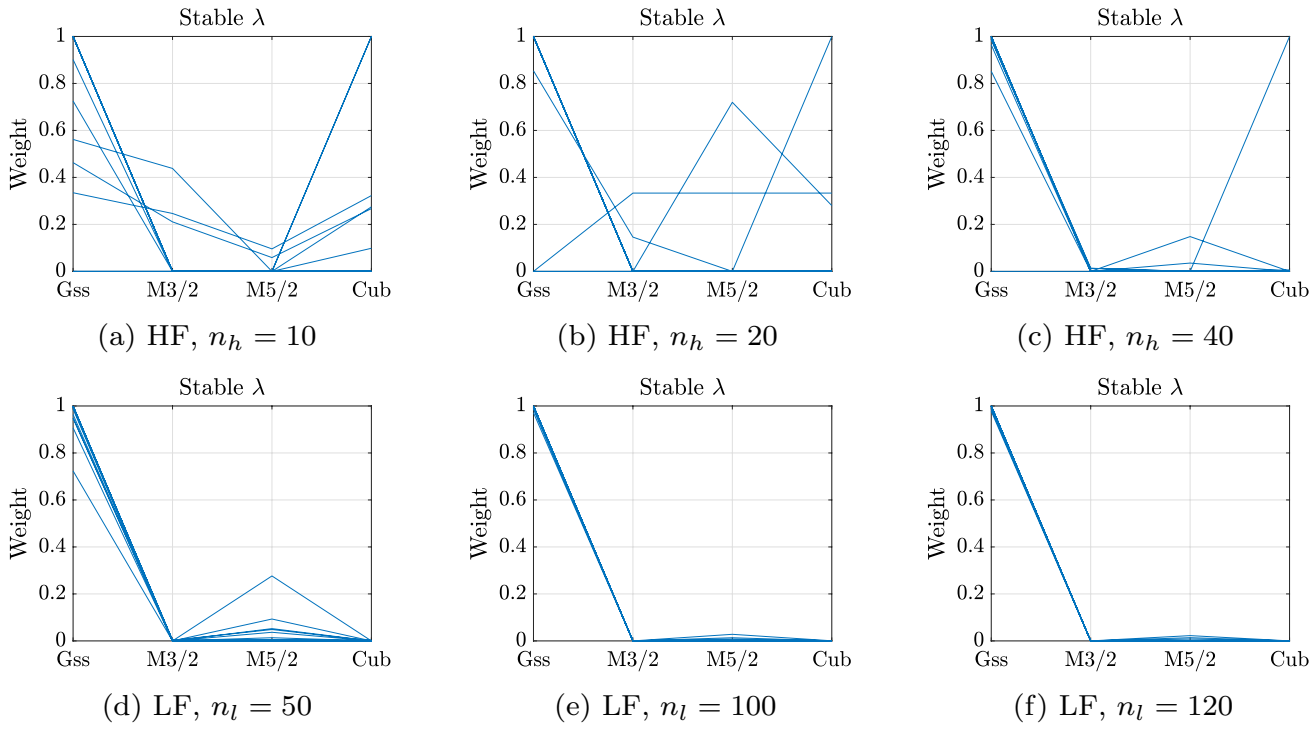


Fig. 11 Parallel coordinate plots of weights of the composite kernels in CKL from 20 independent runs for the ten-variable heat conduction problem, stable λ case

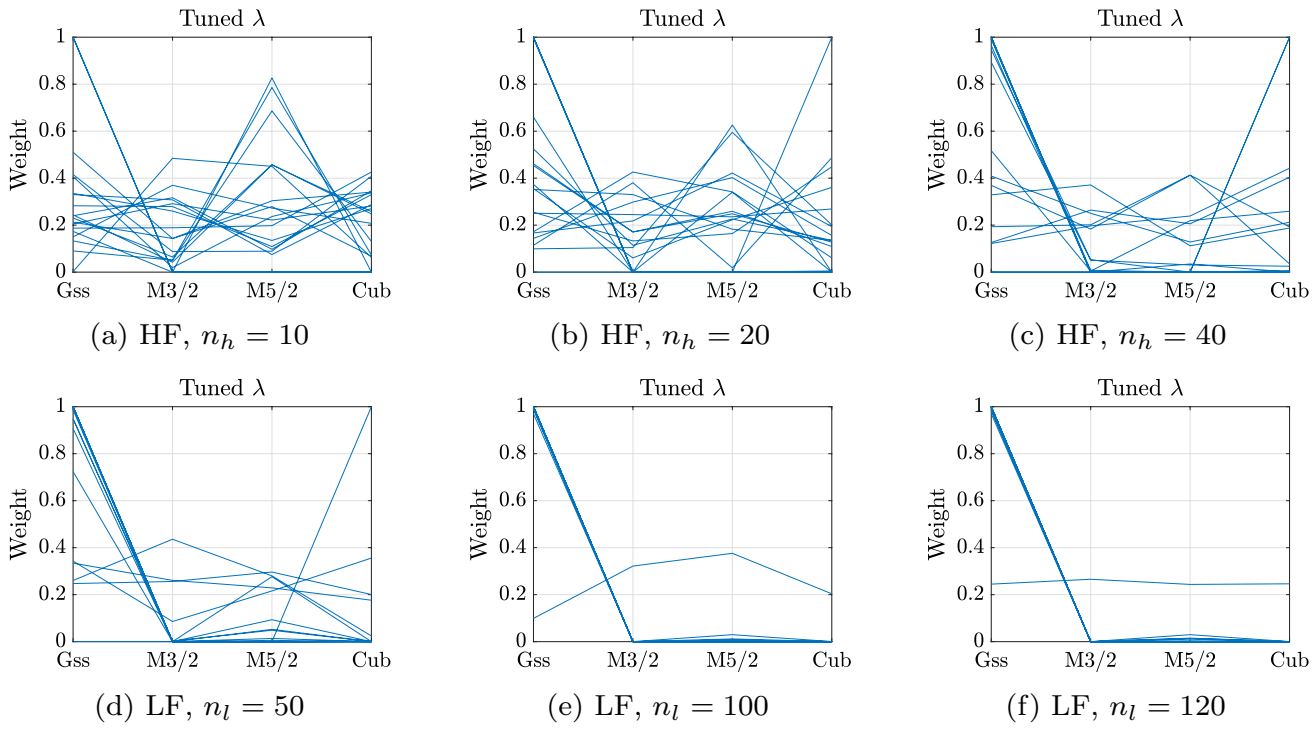


Fig. 12 Parallel coordinate plots of weights of the composite kernels in CKL from 20 independent runs for the ten-variable heat conduction problem, tuned λ case

Table 19 Frequency of selected kernel according to the LOOCV and likelihood criterion for the 10-variable heat conduction problem from 20 independent runs

n_h, n_l	Criterion	Frequency (stable λ/tuned λ)			
		Gaussian	Matern-3/2	Matern-5/2	Cubic
10, 50	LOOCV	7/14	4/1	7/3	2/2
10, 50	Likelihood	7/12	3/3	8/3	2/2
10, 50	Actual NRMSE	8/10	1/3	9/5	2/2
10, 100	LOOCV	8/8	5/7	4/3	3/2
10, 100	Likelihood	7/5	2/7	8/6	3/2
10, 100	Actual NRMSE	8/10	2/3	6/5	4/2
20, 100	LOOCV	7/4	2/9	9/7	2/0
20, 100	Likelihood	7/4	1/8	10/8	2/0
20, 100	Actual NRMSE	7/4	3/7	8/9	2/0
40, 120	LOOCV	13/8	0/1	6/8	1/3
40, 120	Likelihood	15/9	0/1	5/7	0/3
40, 120	Actual NRMSE	11/10	0/2	9/7	0/1

error and likelihood values are relatively good predictors for the actual NRMSE on this problem. Judging from the relatively low values of NRMSE, we infer that approximating this problem is not as difficult as the eight-variable subsonic wing problem (even so, the choice of kernel makes a difference).

Table 20 shows the averaged computational training time for the ten-variable heat conduction problem. The most noticeable observation is that CKL is highly computationally expensive for a high sample size. Such a high computational cost is the main drawback of CKL for high-dimensional problems, which is problematic if CKL is to be employed with an adaptive sampling methodology. Although a cheaper alternative, the model selection is still computationally expensive. The Matern-5/2 kernel is an attractive choice for this problem since it is generally robust.

4.5 Comparison to the previous single-fidelity study

Compared to the previous similar study for single-fidelity GPR (Satria Palar et al. 2020), both the single-fidelity study and the present study agree that the best choice of the kernel is problem dependent. Furthermore, we observe that the gain in accuracy obtained from CKL or model selection applied on bi-fidelity GPR is less significant than that of the single-fidelity case. The reason is that low-fidelity data already help improve the accuracy of the bi-fidelity GPR model, with the proper choice of kernel further aiding in filtering poor kernels from the model. On the other hand, the accuracy gain

Table 20 Averaged computational training time of various modeling techniques for the ten-variable heat conduction problem

n_h, n_l	Averaged training time (stable λ/tuned λ).		
	Identical	Model selection	CKL
10, 50	123 s/145 s	469 s/609 s	683 s/715 s
10, 100	201 s/262 s	830 s/1034.7 s	1743 s/1998 s
20, 100	264 s/303 s	1087 s/1217 s	1741 s/2001 s
40, 120	346 s/411 s	1430 s/1694 s	2591 s/2817 s

is higher for the single-fidelity case since there is no other auxiliary information besides the single-fidelity data; hence, the accuracy of the GPR is improved primarily through the CKL mechanism. Nevertheless, in some cases, CKL can improve the predictive power of bi-fidelity GPR. It will be interesting to investigate how such improved accuracy can translate to more time-efficient optimization or other applications involving adaptive sampling (e.g., structural reliability analysis); we believe this study should be performed for both single- and bi-fidelity cases.

5 Conclusion

This paper investigates the impact of kernel functions and the potential of automatic kernel selection on the global accuracy of bi-fidelity GP for approximating engineering problems. The kernels of interest include Gaussian, Matern-3/2, Matern-5/2, and cubic function. Furthermore, this paper also discusses the possibility of composite kernels to increase the predictive power of the bi-fidelity GP. In the context of engineering applications, we consider the implementation of a weighted-sum combination of kernels into the bi-fidelity GPR formulated as a linear autoregressive model. The study is performed on four engineering problems with various dimensionalities and complexities, focusing on accuracy and training time, which are decisive factors in choosing a surrogate model.

The results from our experiments suggest that the specific choice of kernel function is important, even when the GPR is already assisted by a large amount of low-fidelity data. First, given a sufficient amount of samples, tuning the regression term is essential to improve the accuracy of the bi-fidelity GPR model. However, one should be careful since tuning the regression term might produce low-accuracy models if the sample size is sparse, especially for high-dimensional problems. We observe that no kernel performs best in all problems. Instead, automatically selecting the kernel based on either likelihood and LOOCV error criterion is recommended to avoid poor-performing kernel. Alternatively, using CKL in both low- and high-fidelity potentially can further reduce the approximation error. However, the difference

Table 21 Averaged NRMSE ($\bar{\varepsilon}$) of single-fidelity (SF) and bi-fidelity (BF) CKL for all test problems, tuned λ case

Isogai case ($m = 2$, CR $= 4$)			
n_h, n_l	n_{equiv}	$\bar{\varepsilon}$ (SF)	$\bar{\varepsilon}$ (BF)
20, 60	35	0.3637	0.2524
20, 120	50	0.3279	0.2067
40, 120	70	0.3085	0.1827
Subsonic wing problem ($m = 8$, CR $= 6$)			
n_h, n_l	n_{equiv}	SF	BF
20, 60	30	0.2131	0.1200
20, 120	40	0.2178	0.0797
40, 120	60	0.1955	0.0755
Vibration rig problem ($m = 3$)			
n_h, n_l	n_{equiv}	$\bar{\varepsilon}$ (SF)	$\bar{\varepsilon}$ (BF)
24, 112	–	0.2733	0.2175
30, 168	–	0.2601	0.2139
Heat conduction problem ($m = 2$, CR $= 30$)			
n_h, n_l	n_{equiv}	$\bar{\varepsilon}$ (SF)	$\bar{\varepsilon}$ (BF)
10, 50	11.66	0.8099	0.2412
10, 100	13.33	0.6720	0.1203
20, 100	23.33	0.2367	0.1026
40, 120	44	0.0708	0.0786

in terms of error between CKL and kernel selection is sometimes slight. So, kernel selection is an excellent alternative to full CKL since it avoids kernel misspecification, and its training time is proportional to the number of tested kernels. Further, we observe no significant difference between using LOOCV error and likelihood as the selection criterion. Conversely, the training time of CKL is significantly more expensive than the identical kernel with a selection criterion. CKL is preferable to kernel selection for static sampling (i.e., no adaptive sampling) or, under the adaptive sampling situation, when the simulation cost is significantly expensive in order of hours or days. Lastly, the presence of large noise (as in the case of the three-variable vibration rig problem) leads to the ineffectiveness of CKL and significantly decreases the importance of kernel selection.

Potential subjects for future work include the study and implementation of multiple kernel strategies on other variants of MF GP or kernel-based models. Regarding CKL, future composite kernel-based methods should be able to handle problems with severe non-stationarity or complex relationship between the low- and high-fidelity responses. Further investigation should also deal with multi-fidelity engineering problems of complex correlation to see the combined impact of the type of MF GP model and the choice of kernel function. It is also worth noting that the present study only focuses on bi-fidelity data sets. Therefore, future works should also consider more than two fidelity levels to further investigate the impact of kernel functions. Finally, it would also be interesting to investigate a computationally efficient composite kernel technique for dealing with large multi-fidelity data sets.

Appendix 1: Comparison to single-fidelity Gaussian process

For the sake of completeness, Table 21 shows the comparison of the NRMSE between the single- and bi-fidelity CKL for all problems and combinations of n_h and n_l. This appendix shows that the bi-fidelity model is better than the single-fidelity model for equivalent cost. Consider a combination of low- and high-fidelity data sets in which the cost ratio is defined as

$$CR = \frac{t_{\text{high}}}{t_{\text{low}}}, \tag{20}$$

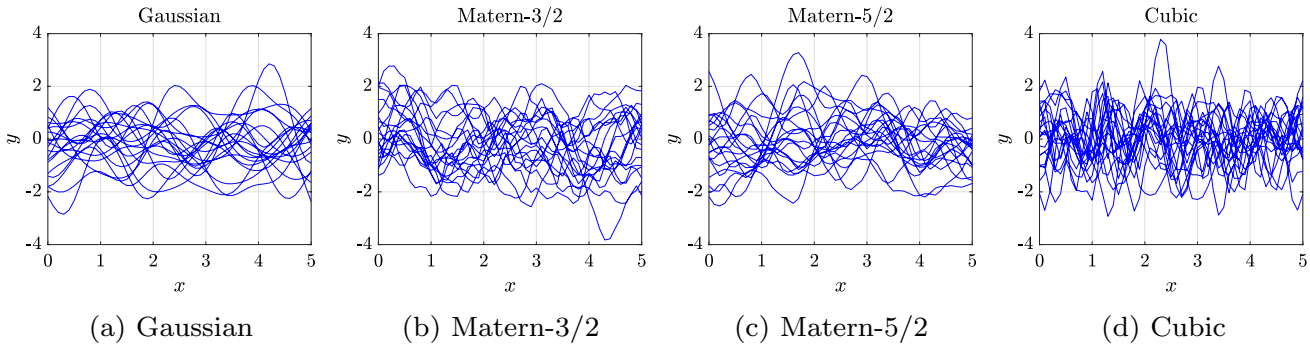
where t_{high} and t_{low} are the wall-clock time of the high- and low-fidelity simulations, respectively. The equivalent sample size of the single-fidelity model, given n_h and n_l, is defined as

$$n_{\text{equiv}} = n_h + \frac{1}{CR} n_l. \tag{21}$$

In this regard, a single-fidelity GP with n_{equiv} samples is then compared with the corresponding bi-fidelity GP with n_h high-fidelity and n_l low-fidelity samples. Notice that there is no n_{equiv} for the vibration rig problem since the high-fidelity data are evaluated experimentally. Hence, we use the single-fidelity GP using n_h samples for the vibration rig problem. The averaged NRMSE results for all problems, tuned λ case, are shown in Table 21. It can be seen that the bi-fidelity GP with CKL always outperforms its single-fidelity GP counterparts, with the only exception being on the heat conduction case with $n_h = 40/n_l = 120$.

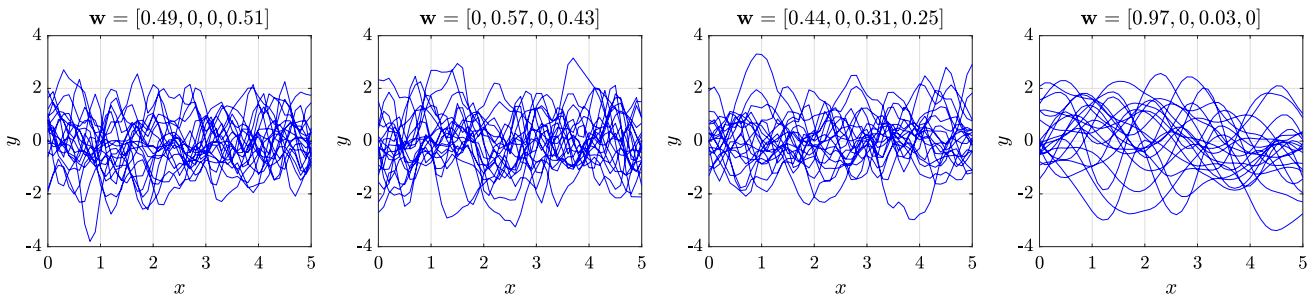
Appendix 2: Examples of sample paths from Gaussian processes

To illustrate the behavior of the composite kernels, Figs. 13 and 14 show the sample paths generated from the four individual kernels and CKL with various weights, respectively. The sample paths were generated using the lengthscale of 0.1 for all kernels. As shown in Fig. 14, the sample paths from composite kernels reflect the behavior of the constituents. Let us denote the vector of weight as follows: $\boldsymbol{w} = [\text{Gaussian, Matern-3/2, Matern-5/2, Cubic}]$. The weights shown in The CKL with $\boldsymbol{w} = [0.49, 0, 0, 0.51]$ (extracted from the two-variable Isogai problem) is a combination of only the Gaussian and cubic kernel, with the sample paths reflect the smooth nature of Gaussian and the rapid change of the cubic. On the other hand, the combination of Matern-3/2 and cubic kernel ($\boldsymbol{w} = [0, 0.57, 0, 0.43]$, extracted from the subsonic wing problem) yields a rough characteristic that primarily comes from the former. The combination of Gaussian, Matern-5/2, and cubic ($\boldsymbol{w} = [0.44, 0, 0.31, 0.25]$, extracted from the eight-variable subsonic wing problem) primarily exhibits the smoothness of Gaussian, with slight roughness from Matern-5/2 and cubic. Finally, the combination of dominant Gaussian kernel and cubic ($\boldsymbol{w} = [0.97, 0, 0.03, 0]$, extracted from



(a) Gaussian (b) Matern-3/2 (c) Matern-5/2 (d) Cubic

Fig. 13 Examples of 20 sample paths generated from various kernel functions using $\theta = 0.1$



(a) $\boldsymbol{w} = [0.49, 0, 0, 0.51]$ (b) $\boldsymbol{w} = [0, 0.57, 0, 0.43]$ (c) $\boldsymbol{w} = [0.44, 0, 0.31, 0.25]$ (d) $\boldsymbol{w} = [0.97, 0, 0.03, 0]$

Fig. 14 Examples of 20 sample paths generated from composite kernels with various weights using $\theta = 0.1$. The composition of the weight vector is as follows: $\boldsymbol{w} = [\text{Gaussian,Matern-3/2,Matern-5/2,Cubic}]$

the ten-variable heat conduction problem) almost looks similar to Gaussian; however, the slight change of this characteristic can lead to a better prediction as observed in the ten-variable heat conduction problem.

Acknowledgements The authors acknowledge financial support from Penelitian Dasar Unggulan Perguruan Tinggi research scheme administered by Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi, Republic of Indonesia.

Declarations

Competing interests On behalf of all authors, the corresponding author states that there is no conflict of interest.

Replications of results The code and the data needed to replicate the results can be downloaded from the following link: https://github.com/optimuspram/MF-GPR-code.

References

Bachoc F (2013) Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. Comput Stat Data Anal 66:55–69

Bertram A, Othmer C, Zimmermann R (2018) Towards real-time vehicle aerodynamic design via multi-fidelity data-driven reduced order modeling. In: 2018 AIAA/ASCE/AHS/ASC structures. structural dynamics, and materials conference 0916

Bonfiglio, L., Perdikaris P, Brizzolara S, Karniadakis G (2017) A multi-fidelity framework for investigating the performance of super-cavitating hydrofoils under uncertain flow conditions. In: 19th AIAA non-deterministic approaches conference, p 1328

Bostanabad R, Kearney T, Tao S, Apley DW, Chen W (2018) Leveraging the nugget parameter for efficient Gaussian process modeling. Int J Numer Methods Eng 114(5):501–516

Brevault L, Balesdent M, Hebbal A (2020) Overview of Gaussian process based multi-fidelity techniques with variable relationship between fidelities, application to aerospace systems. Aerosp Sci Technol 107:106339

Bryson DE, Rumpfkeil MP (2017) All-at-once approach to multifidelity polynomial chaos expansion surrogate modeling. Aerosp Sci Technol 70:121–136

Cère-Aéro (2022) Flow5 v7.15—documentation. https://flow5.tech/

Cutajar K, Pullin M, Damianou A, Lawrence N, González J (2019) Deep Gaussian processes for multi-fidelity modeling. arXiv preprint. arXiv:1903.07320

de Baar J, Roberts S, Dwight R, Mallol B (2015) Uncertainty quantification for a sailing yacht hull, using multi-fidelity Kriging. Comput Fluids 123:185–201

Economon TD, Palacios F, Copeland SR, Lukaczyk TW, Alonso JJ (2016) Su2: an open-source suite for multiphysics simulation and design. AIAA J 54(3):828–846

Fernández-Godino MG, Park C, Kim NH, Haftka RT (2016) Review of multi-fidelity models. arXiv preprint. arXiv:1609.07196

Han Z-H, Görtz S (2012) Hierarchical Kriging model for variable-fidelity surrogate modeling. AIAA J 50(9):1885–1896

Isogai K (1979) On the transonic-dip mechanism of flutter of a swept-back wing. AIAA J 17(7):793–795

Jin S-S (2020) Compositional kernel learning using tree-based genetic programming for Gaussian process regression. Struct Multidisc Optim 62:1313–1351

Jofre L, Geraci G, Fairbanks H, Doostan A, Iaccarino G (2018) Multi-fidelity uncertainty quantification of irradiated particle-laden turbulence. arXiv preprint. arXiv:1801.06062

Kennedy MC, O'Hagan A (2000) Predicting the output from a complex computer code when fast approximations are available. Biometrika 87(1):1–13

Kersaudy P, Sudret B, Varsier N, Picon O, Wiart J (2015) A new surrogate modeling technique combining Kriging and polynomial chaos expansions—application to uncertainty analysis in computational dosimetry. J Comput Phys 286:103–117

Konakli K, Sudret B (2016) Reliability analysis of high-dimensional models using low-rank tensor approximations. Probab Eng Mech 46:18–36

Kronberger G, Kommenda M (2013) Evolution of covariance functions for Gaussian process regression using genetic programming. In: Moreno-Díaz R, Pichler F, Quesada-Arencibia A (eds) Computer aided systems theory—EUROCAST 2013. Springer, Berlin, pp 308–315

Le Gratiet L, Garnier J (2012) Recursive co-Kriging model for design of computer experiments with multiple levels of fidelity. Int J Uncertain Quant 4(5):365–386

Liu B, Koziel S, Zhang Q (2016) A multi-fidelity surrogate-model-assisted evolutionary algorithm for computationally expensive optimization problems. J Comput Sci 12:28–37

Liu X, Zhao W, Wan D (2022) Multi-fidelity co-Kriging surrogate model for ship hull form optimization. Ocean Eng 243:110239

Maolin S, Liye L, Sun W, Xueguan S (2020) A multi-fidelity surrogate model based on support vector regression. Struct Multidisc Optim 61(6):2363–2375

Meng X, Wang Z, Fan D, Triantafyllou MS, Karniadakis GE (2021) A fast multi-fidelity method with uncertainty quantification for complex data correlations: application to vortex-induced vibrations of marine risers. Comput Methods Appl Mech Eng 386:114212

Ng LWT, Eldred M (2012) Multifidelity uncertainty quantification using non-intrusive polynomial chaos and stochastic collocation. In: 53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference, 20th AIAA/ASME/AHS adaptive structures conference, 14th AIAA, p 1852

Palar PS, Shimoyama K (2018) On efficient global optimization via universal Kriging surrogate models. Struct Multidisc Optim 57(6):2377–2397

Palar PS, Shimoyama K (2019) Efficient global optimization with ensemble and selection of kernel functions for engineering design. Struct Multidisc Optim 59(1):93–116

Palar PS, Tsuchiya T, Parks GT (2016) Multi-fidelity non-intrusive polynomial chaos based on regression. Comput Methods Appl Mech Eng 305:579–606

Palar PS, Zuhal LR, Shimoyama K, Tsuchiya T (2018) Global sensitivity analysis via multi-fidelity polynomial chaos expansion. Reliab Eng Syst Saf 170:175–190

Palar PS, Parussini L, Bregant L, Shimoyama K, Izzaturrahman MF, Baehaqi FA, Zuhal L (2022) Composite kernel functions for surrogate modeling using recursive multi-fidelity Kriging. In: AIAA SCITECH 2022 forum, p 0506

Pang G, Perdikaris P, Cai W, Karniadakis GE (2017) Discovering variable fractional orders of advection–dispersion equations from field data using multi-fidelity bayesian optimization. J Comput Phys 348:694–714

Park C, Haftka RT, Kim NH (2017) Remarks on multi-fidelity surrogates. Struct Multidisc Optim 55(3):1029–1050

Perdikaris P, Raissi M, Damianou A, Lawrence ND, Karniadakis GE (2017) Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. Proc R Soc A Math Phys Eng Sci 473(2198):20160751

Ranjan P, Haynes R, Karsten R (2011) A computationally stable approach to Gaussian process interpolation of deterministic computer simulation data. Technometrics 53(4):366–378

Satria Palar P, Rizki Zuhal L, Shimoyama K (2020) Gaussian process surrogate model with composite kernel learning for engineering design. AIAA J 58(4):1864–1880

Serani A, Pellegrini R, Wackers J, Jeanson C-E, Queutey P, Visonneau M, Diez M (2019) Adaptive multi-fidelity sampling for CFD-based optimisation via radial basis function metamodels. Int J Comput Fluid Dyn 33(6–7):237–255

Song X, Lv L, Sun W, Zhang J (2019) A radial basis function-based multi-fidelity surrogate model: exploring correlation between high-fidelity and low-fidelity models. Struct Multidisc Optim 60(3):965–981

Sundar V, Shields MD (2019) Reliability analysis using adaptive kriging surrogates with multimodel inference. ASCE-ASME J Risk Uncertain Eng Syst Part A Civ Eng 5(2):04019004

Tao J, Sun G (2019) Application of deep learning based multi-fidelity surrogate model to robust aerodynamic design optimization. Aerosp Sci Technol 92:722–737

Toal DJ (2015) Some considerations regarding the use of multi-fidelity Kriging in the construction of surrogate models. Struct Multidisc Optim 51(6):1223–1245

Yoo K, Bacarreza O, Aliabadi MF (2020) A novel multi-fidelity modelling-based framework for reliability-based design optimisation of composite structures. Eng Comput 38:595–608

Zhang X, Xie F, Ji T, Zhu Z, Zheng Y (2021) Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. Comput Methods Appl Mech Eng 373:113485

Zuhal LR., Faza GA, Palar PS, Liem RP (2021) On dimensionality reduction via partial least squares for kriging-based reliability analysis with active learning. Reliab Eng Syst Saf 215:107848