Structural and Multidisciplinary Optimization (2018) 58:399–414
https://doi.org/10.1007/s00158-018-2031-2

RESEARCH PAPER

Low-fidelity scale factor improves Bayesian multi-fidelity prediction by reducing bumpiness of discrepancy function

Chanyoung Park[1] · Raphael T. Haftka[1] · Nam H. Kim[1]

Received: 5 January 2018 / Revised: 5 June 2018 / Accepted: 12 June 2018 / Published online: 23 June 2018

Abstract

This study explores why the use of the low-fidelity scale factor can substantially improve the accuracy of the Bayesian multi-fidelity surrogate (MFS). It is shown analytically that the Bayesian MFS framework utilizes the scale factor to reduce the waviness and variation of the discrepancy function by maximizing the Gaussian process-based likelihood function. Less wavy functions are more accurately fitted, and variation reduction mitigates the effect of fitting error. Bumpiness is another way used to combine waviness and variation. Two examples, Borehole3 and Hartmann6, illustrated that indeed the Bayesian MFS reduced bumpiness using the scale factor. The finding may be useful for MFS using surrogates lacking uncertainty structure, so that likelihood is not an option, but bumpiness may be.

Keywords Bayesian · multi-fidelity · surrogate · scale factor · bumpiness · Gaussian process

1 Introduction

Design optimization and uncertainty quantification often require numerous expensive simulations to find an optimum design and to propagate uncertainties. Instead, a surrogate fit to dozens of simulations is often employed as a cheap alternative. However, for computationally expensive high-fidelity simulations, even evaluating sufficient samples for building a surrogate is often unaffordable. To address this challenge, multi-fidelity surrogates (MFS) combine inexpensive low-fidelity models with a small number of high-fidelity simulations. For example, MFS can be employed to predict the response of an expensive finite element model with a few runs of the model and many runs from a less accurate model with a coarse mesh. MFS have been applied extensively in the literature (Fernández-Godino et al. 2016; Gano et al. 2006).

The regression-based MFS framework has been used extensively in design optimization, for example, by combining two- and three-dimensional finite element models (Mason et al. 1998) or coarse and fine finite element models (Balabanov et al. 1998). More recently, Bayesian MFS frameworks have become popular. Qian and Wu (2008) proposed the use of Markov chain Monte Carlo and sample average approximation algorithm for hyperparameter estimation of the Bayesian MFS framework. Co-Kriging followed with better computational efficiency (Forrester, 2007; Le Gratiet 2013). A Gaussian process (GP) based Bayesian MFS framework was introduced by Kennedy and O'Hagan (2000). The use of GP model provides flexibility and their prediction is not limited to a specific form of the trend function. Thus, the Bayesian MFS can also be useful when there is no prior information, which is also called non-informative prior (Rasmussen, 2006).

The Bayesian MFS framework can be expressed as

$$\hat{y}_H(\mathbf{x}) = \rho\hat{y}_L(\mathbf{x}) + \hat{\delta}(\mathbf{x}) \tag{1}$$

where $\hat{y}_H(\mathbf{x})$ is high-fidelity function prediction at $\mathbf{x}$, $\hat{y}_L(\mathbf{x})$ is low-fidelity function prediction, $\hat{\delta}(\mathbf{x})$ is discrepancy function prediction, and ρ is a low-fidelity scale factor. This scale factor has rarely been used in the past. However, the combined use of a scale factor and a discrepancy function has been common for recently developed GP-based MFS frameworks (Fernández-Godino et al. 2016; Zhou et al. 2018).

The MFS frameworks can handle noisy data using (1) with a noise model, where random noise follows a normal distribution defined with zero mean and a noise standard deviation, which needs to be estimated. However, this paper focuses on

Responsible Editor: Helder C. Rodrigues

✉ Nam H. Kim
nkim@ufl.edu

[1] Department of Mechanical and Aerospace Engineering, University of Florida, PO Box 116250, Gainesville, FL 32611-6250, USA

MFS prediction with a few high-fidelity samples without random noise. This is because filtering noise based on a few high-fidelity samples is often not reliable (Matsumura et al. 2015).

It was found that the Bayesian framework and co-Kriging often gave significantly more accurate predictions than other MFS frameworks with a discrepancy function only (Park et al. 2017). An interesting observation was the influence of the scale factor. The Bayesian framework gave much more accurate discrepancy predictions with the scalar and so did the MFS predictions, but it gave mediocre predictions without the scalar. The objective of this paper is to discover the reason why the use of the scalar made the Bayesian MFS significantly more accurate. Understanding the reason behind the Bayesian MFS is likely to help extend the success to other non-Bayesian MFS frameworks that may have advantages for some applications.

This paper is organized as follows. Section 2 discusses the importance of using the scalar for making MFS prediction. One-dimensional examples show how the scalar improves the accuracy of the discrepancy function to improve MFS prediction. It is discussed that large waviness and variation of the discrepancy function tend to increase errors. Bumpiness is introduced to combine them. Section 3 explains how the Bayesian framework characterizes a discrepancy function with variation and waviness. It finds the scalar by combining them through the likelihood function based on a Gaussian process model. Section 4 uses multi-dimensional examples to illustrate the correlation between bumpiness and error. These include the Borehole3 physical function and Hartmann6 algebraic function. Concluding remarks are presented in Section 5.

2 The importance of having the scale factor to reduce the bumpiness of a discrepancy

In MFS frameworks, it is usually assumed that the low-fidelity function is well approximated due to a relatively large number of samples. With the model expressed in (1), the low-fidelity prediction is based on a sufficient number of low-fidelity samples. Once the low-fidelity model is determined, the differences between the high-fidelity samples and the low-fidelity predictions are used to fit the discrepancy function. Since the discrepancy samples are based on the high-fidelity samples, the discrepancy function prediction depends on a small number of high-fidelity samples. Therefore, if the discrepancy function is wavy or has a high amplitude of oscillation, a small number of high-fidelity samples may lead to large errors.

Fortunately, the discrepancy function depends on the scalar ρ because it is defined as the difference between the high-fidelity prediction and the scaled low-fidelity prediction. Therefore, it is possible to manage the discrepancy function by changing the scalar. Based on our study about MFS frameworks, the Bayesian MFS framework was particularly effective with ρ. We found that it was because the Bayesian MFS determines the scalar to make the discrepancy simple. That is, it reduces the waviness and variation of the discrepancy, which tends to improve the prediction accuracy.

In order to show the above-mentioned characteristic, Fig. 1 shows two analytical examples where the Bayesian MFS framework was applied with and without ρ. In Fig. 1(a) and (b), the red and blue curves are the true high- and low-fidelity functions, and the crosses and the hollow circles are the high- and low-fidelity samples, respectively. Figure 1(c) and (d) show the true discrepancy function (dashed curve) without ρ (or with $\rho = 1$) and the corresponding predictions (red solid curve) using the Bayesian framework. The figures also show the two-sigma prediction uncertainty. Due to a large variation of the true discrepancy function, the four samples were not enough to predict the discrepancy accurately. Also, the 2σ-confidence intervals (blue areas) failed to cover the true discrepancy function.

On the other hand, when the scale factor is present, the Bayesian framework found $\rho = 2.5$ for the first example, whose discrepancy is shown in Fig. 1(e). Note that the discrepancy samples in Fig. 1(c) and (e) are different because they are discrepancies between the high-fidelity samples to that of scaled low-fidelity samples. By choosing $\rho = 2.5$, the variation in the original discrepancy in Fig. 1(c) is drastically reduced as shown in Fig. 1(e), and thus, four samples were enough to accurately predict it. Due to the reduction of variation, the root-mean-squared-error (RMSE) in the discrepancy is reduced by more than a factor of 8. Although the true discrepancy function is still wavy, the bumpiness is significantly reduced by reducing the variation in this case.

In the case of the second example, Fig. 1(f) shows that the Bayesian method found $\rho = 2$ that turns the wavy discrepancy function in Fig. 1(e) into a linear function, and the prediction becomes almost perfect. In this case, the Bayesian MFS framework reduces the waviness of the discrepancy while the variation is still large.

Note that the results also illustrate the flexibility of the GP-based Bayesian MFS framework. The trend functions of the GP models of the Bayesian MFS framework were set to a constant function. Figure 1(f) shows that the GP-based discrepancy prediction gave a prediction like a linear function based on the data while its trend function is a constant function. In addition to the two analytical examples, a cantilever beam example is also included in Appendix B.

The concept of variation and waviness can be combined into the concept of bumpiness. The notion of bumpiness, which is also referred to as roughness, was introduced for measuring function roughness (Duchon 1977; Gutmann 2001; Cressie 2015). Salem and Tomaso (2018) uses bumpiness to select surrogate and surrogate weighting for an

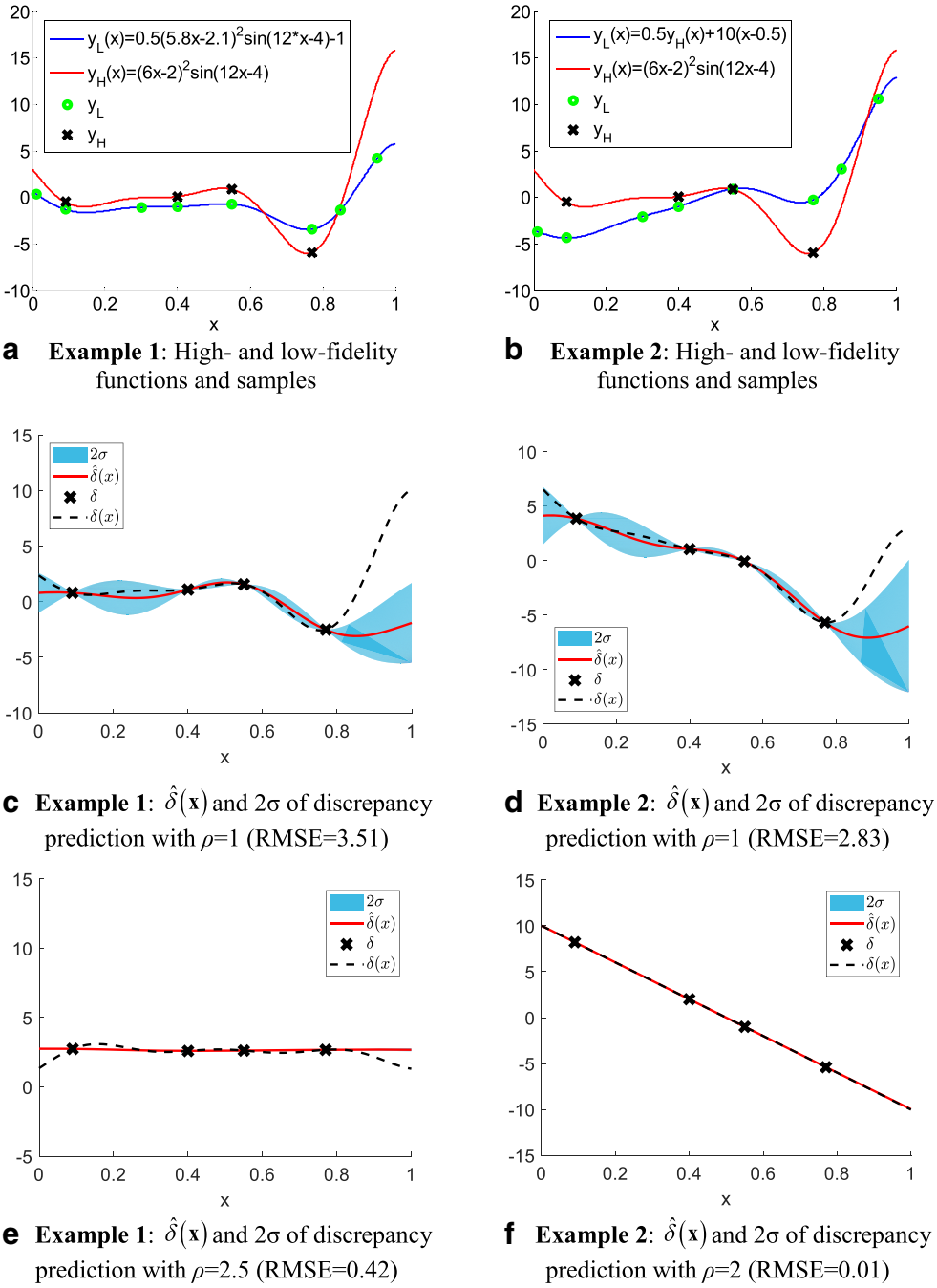


Fig. 1 Options to improve the accuracy by including ρ for MFS prediction ($\delta(x)$: True discrepancy function; $\hat{\delta}(x)$: Discrepancy prediction; $\boldsymbol{\delta}$: Discrepancy function data)

a **Example 1**: High- and low-fidelity functions and samples

b **Example 2**: High- and low-fidelity functions and samples

c **Example 1**: $\hat{\delta}(\mathbf{x})$ and 2σ of discrepancy prediction with ρ=1 (RMSE=3.51)

d **Example 2**: $\hat{\delta}(\mathbf{x})$ and 2σ of discrepancy prediction with ρ=1 (RMSE=2.83)

e **Example 1**: $\hat{\delta}(\mathbf{x})$ and 2σ of discrepancy prediction with ρ=2.5 (RMSE=0.42)

f **Example 2**: $\hat{\delta}(\mathbf{x})$ and 2σ of discrepancy prediction with ρ=2 (RMSE=0.01)

ensemble. Bumpiness is an integral of the square of second derivative of the associated function, as

$$B(f(x)) = \int \left| f''(x) \right|^2 \mathrm{d}x \tag{2}$$

In the following section, we describe how the Bayesian MFS framework combines the effect of variation and waviness through the likelihood function. In the Bayesian method, finding a scalar value that maximizes the likelihood function is related to reducing variation and waviness, which also leads to the reduction of bumpiness in (2). However, maximizing the likelihood function does not mean exactly minimizing the bumpiness.

3 Bayesian MFS framework: Finding the scale factor that reduces bumpiness

The two examples in the previous section used the Bayesian framework to determine ρ, as shown in Fig. 1(e) and (f). The Bayesian framework finds ρ using the method of maximum

likelihood estimation (MLE) that estimates ρ at the mode of the likelihood function. While the ρ obtained by MLE is not exactly the same as the ρ minimizing the bumpiness, for the examples we analyzed they were close, as will be seen in the next section. This section discusses why the Bayesian formulation tends to reduce the bumpiness in the discrepancy function using variation reduction and waviness reduction.

The likelihood function, which is in the form of the multivariate normal distribution, can be reformulated to find ρ as the minimizer, as

$$\underset{\rho}{\text{argmin}} \quad \hat{\sigma}_{\Delta}^{2}(\rho)|\mathbf{R}_{\Delta}(\boldsymbol{\omega}_{\Delta})|^{1/n_H} \tag{3}$$

where $\hat{\sigma}_{\Delta}(\rho)$ and $|\mathbf{R}_{\Delta}(\boldsymbol{\omega}_{\Delta})|$ are, respectively, the process standard deviation and the determinant of the correlation matrix obtained based on discrepancy data $\mathbf{y}_H - \rho\mathbf{y}_L^c$. $\hat{\sigma}_{\Delta}(\rho)$ represents the variation, while $|\mathbf{R}_{\Delta}(\boldsymbol{\omega}_{\Delta})|$ represents the waviness of the discrepancy data. $|\mathbf{R}_{\Delta}(\boldsymbol{\omega}_{\Delta})|$ can be interpreted as a waviness measure, which is a function of waviness vector $\boldsymbol{\omega}_{\Delta}$. The detailed derivation of (3) from the likelihood function is given in Appendix A.

$\hat{\sigma}_{\Delta}$ and $\boldsymbol{\omega}_{\Delta}$ are estimated based on the discrepancy data for given ρ using auto-covariance, which was introduced to quantify the probabilistic similarity of two values in space (Ripley 1981). The auto-covariance is also applicable to random function generation based on the variation and waviness parameters of the auto-covariance. The inverse use of the auto-covariance allows estimating the variation and waviness of the true function based on the data (Rasmussen 2004).

The Bayesian framework uses the Gaussian correlation function to model the auto-covariance of the uncertainties in discrepancy function predictions at different locations. Let $\Delta(\mathbf{x})$ and $\Delta(\mathbf{x}')$ be the discrepancy predictions at two data locations $\mathbf{x}$ and $\mathbf{x}'$. The covariance between them is expressed using their distance as

$$\begin{aligned}&\text{cov}\left(\Delta(\mathbf{x}), \Delta\left(\mathbf{x}'\right)\right)\\ &= \sigma_{\Delta}^{2}\exp\left(-\left(\mathbf{x}-\mathbf{x}'\right)^{\mathrm{T}}\text{diag}(\boldsymbol{\omega}_{\Delta})\left(\mathbf{x}-\mathbf{x}'\right)\right)\end{aligned} \tag{4}$$

where $\text{diag}(\boldsymbol{\omega}_{\Delta})$ is a diagonal matrix with the waviness vector $\boldsymbol{\omega}_{\Delta}$, which has the same dimension with $\mathbf{x}$.

Figure 2 shows two sets of samples from (a) a wavy function with small variation and (b) a less wavy function with large variation. The process standard deviation and waviness were estimated based on the data sets using (4). It is clear that a wavy function with small variation has a small σ_{Δ} and a large $\boldsymbol{\omega}_{\Delta}$. On the other hand, a less wavy function with large variation has a large σ_{Δ} and a small $\boldsymbol{\omega}_{\Delta}$.

The effect of the process standard deviation on (3) is obvious because the objective function decreases as with the process standard deviation. The influence of $|\mathbf{R}_{\Delta}|$ on (3) is a function of the waviness parameter and discrepancy data locations. Since data location remains the same for finding ρ, there is no influence of the data location.

The correlation matrix $\mathbf{R}_{\Delta}$ is a symmetric square matrix. The size of the matrix is the number of discrepancy data. The diagonal elements of $\mathbf{R}_{\Delta}$ are one and the off-diagonal elements are obtained by the exponential part of the auto-correlation in (4). The off-diagonal elements measure the correlations between discrepancy values at two different data locations. The correlation matrix is expressed as

$$\mathbf{R}_{\Delta} = \begin{bmatrix} 1 & \cdots & \exp\left(-\left(\mathbf{x}_{\delta}-\mathbf{x}_{\delta}'\right)^{T}\text{diag}(\boldsymbol{\omega}_{\Delta})\left(\mathbf{x}_{\delta}-\mathbf{x}_{\delta}'\right)\right) \\ & \ddots & \vdots \\ symm & & 1 \end{bmatrix}_{(n_H \times n_H)} \tag{5}$$

The properties of the determinant of a correlation matrix are well known (Reddon et al. 1985; Lophaven et al. 2002; Johnson and Wichern 2007). The determinant has a minimum value of zero when $\boldsymbol{\omega}_{\Delta} = \mathbf{0}$, which makes all the off-diagonal elements one. On the other hand, the determinant has a maximum value of one when $\boldsymbol{\omega}_{\Delta} \to \infty$, which makes all the off-diagonal elements zero. As shown by Johnson and Wichern (2007), the determinant decreases monotonically with decreasing $\boldsymbol{\omega}_{\Delta}$. Since waviness is proportional to $\boldsymbol{\omega}_{\Delta}$, reducing the determinant is equivalent to reducing the waviness.

In summary, the minimization of (3) is achieved by reducing the product of the terms representing variation and waviness. When there is no way to reduce the variation and waviness using ρ simultaneously, (3) trade-off between them to minimize the objective function. Equation (3) drives bumpiness reduction through the reduction of variation and/or waviness. However, minimizing (3) is not theoretically equivalent to minimizing the bumpiness of the discrepancy function defined in (2).

The reader is referred to Section A.2 of Appendix for the detailed formulas to measure variation and waviness using the auto-covariance model of the Bayesian framework.

4 Multi-dimensional examples

In this section, the influence of ρ on increasing MFS prediction accuracy by reducing bumpiness will be presented through multi-dimensional examples: (a) physical borehole function and (b) numerical Hartmann 6 function.

Firstly, in this section, three different MFS frameworks are compared, along with two single-fidelity Kriging surrogates. Table 1 shows the framework descriptions with the corresponding abbreviations. "H" and "L" denote Kriging surrogates using only high- and low-fidelity samples, respectively. "B" is the Bayesian framework without ρ. "BR" is the Bayesian framework with ρ that is found by minimizing the

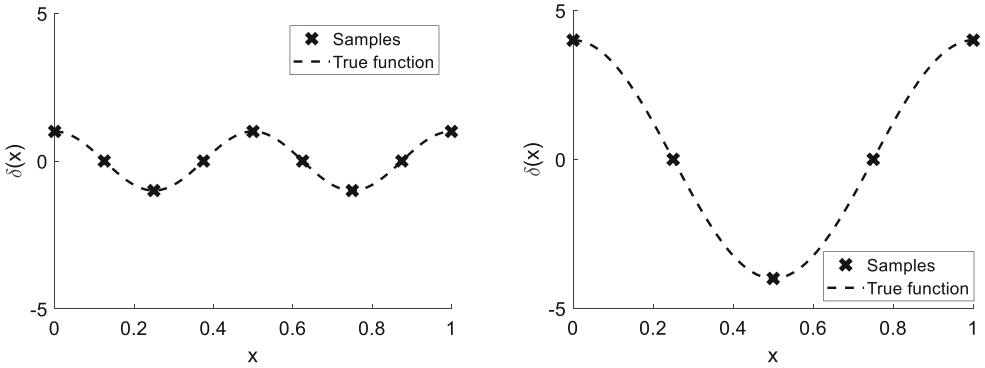


Fig. 2 The variation and the waviness of a discrepancy function

a A wavy function with small variation: σ_Δ=1.05 and ω_Δ=21.61 for δ(x)=cos(4πx)

a A less wavy function with large variation: σ_Δ=3.40 and ω_Δ=7.07 for δ(x)=4cos(2πx)

bumpiness, while ρ in BR2 is found by minimizing error. The comparison between BR and B shows the effect of including ρ on the prediction. The comparison between BR and BR2 shows the effect of different criteria for finding ρ: reducing bumpiness versus minimizing error.

Secondly, the influences of ρ on the bumpiness of the discrepancy function and the accuracy of MFS prediction were measured in the form of graphs by gradually changing ρ. However, since evaluating second-order derivatives of a multi-dimensional function in (2) is a computational challenge (Cressie, N., 2015; Duchon, J., 1977), one-dimensional bumpiness measures are used along $N_{line}=1000$ randomly generated lines. Each line was generated by connecting two randomly generated points and extending the line to the boundary of the sampling domain. The average of the one-dimensional bumpiness measures is used as a representative bumpiness measure for a given value of ρ, as

$$B(\rho)=\frac{1}{N_{line}}\sum_{i=1}^{N_{line}}\int\left|\delta''(s_i,\rho)\right|^2 ds_i \tag{6}$$

where $s_i$is the parameter along the i^{th} line, and $\delta''(s_i,\rho)$is the second-order derivative of the discrepancy function along the line for a given ρ.

The graph of bumpiness with respect to ρ explains the influence of ρ on the bumpiness, but it does not explain whether the change in bumpiness is caused by the variation and/or the waviness. To measure their individual contributions, graphs of variation and waviness are also obtained in terms of ρ. The discrepancy variation is measured using the variance of the discrepancy along all lines as

$$V(\rho)=\frac{1}{N_{line}}\sum_{i=1}^{N_{line}}\sigma_i^2 \tag{7}$$

$$\mu_i=\int\delta(s_i,\rho)/L_i ds_i \quad \text{and} \quad \sigma_i^2=\int(\delta(s_i,\rho)-\mu_i)^2/L_i ds_i \tag{8}$$

where $L_i$is the length of the i^{th} line.

To quantify the effect of waviness, a normalized bumpiness is used. For example, the variation of $\delta_1(x)=2\sin(100\times)$ is four times of the variation of $\delta_2(x)=\sin(100\times)$ while their waviness is the same. If these two functions are normalized, then they have the same variation, so that only waviness can be measured. The waviness measure is defined as

$$W(\rho)=\frac{1}{N_{line}}\sum_{i=1}^{N_{line}}\int\left|\bar{\delta}''(s_i,\rho)\right|^2 ds_i \tag{9}$$

where $\bar{\delta}''(s_i,\rho)=\delta''(s_i,\rho)/\sigma_i$ is the normalized second-order derivative of the discrepancy function using the standard deviation.

The accuracy graph of the Bayesian framework is measured in terms of RMSE with respect to ρ. Since the MFS prediction depends on samples; i.e., design of experiments (DOE), 100 DOEs were randomly generated using the nearest neighbor sampling method (Forrester et al. 2007; Jin et al. 2005). Since the Bayesian framework is applied for each DOE to calculate RMSE, the above process yields 100 RMSEs. In the following examples, the median, 25 and 75 percentiles of RMSEs were obtained as a function of ρ.

Table 1 Frameworks and the corresponding labels

Label	Framework
BR	Bayesian discrepancy framework including ρ
BR2	Bayesian discrepancy framework and finding ρ for maximizing agreement
B	Bayesian framework excluding ρ (or BR with $\rho=1$)
H	Kriging surrogate based on high-fidelity samples
L	Kriging surrogate based on low-fidelity samples

4.1 Borehole function example: Reducing discrepancy variation

The empirical Borehole function calculates the water flow rate from an aquifer through a borehole. The function was obtained

based on assumptions of steady-state flow from an upper aquifer to the borehole and from the borehole to the lower aquifer, no groundwater gradient, and laminar, isothermal flow through the borehole (Morris et al. 1993).

In this example, the borehole function is considered as the high-fidelity function and an approximate function is used as a low-fidelity function. The high-fidelity function is defined as

$$f_H(R_w, L, K_w) = \frac{2\pi T_u(H_u - H_l)}{\ln(R/R_w)\left(1 + \frac{2LT_u}{\ln(R/R_w)R_w^2 K_w} + \frac{T_u}{T_l}\right)} \quad (10)$$

The flow rate $f_H(R_w, L, K_w)$ is a function of three input variables, R_w, L, and K_w, which are the borehole radius, borehole length and hydraulic conductivity of borehole, respectively. The ranges of the input variables and other environmental parameters are presented in Table 2. The values of the parameters were determined as nominal values of the parameters based on Morris et al. (1993).

A low-fidelity function of the borehole function is obtained from the literature (Xiong et al. 2013) as

$$f_L(R_w, L, K_w) = \frac{5T_u(H_u - H_l)}{\ln(R/R_w)\left(1.5 + \frac{2LT_u}{\ln(R/R_w)R_w^2 K_w} + \frac{T_u}{T_l}\right)} \quad (11)$$

Note that bounds of [0.5, 1.5] were used for ρ, and constant trend functions were used for the Bayesian framework.

Since MFS are built with low- and high-fidelity samples, there are many different combinations of low- and high-fidelity samples possible for the same total computational budget. MFS performances for different combinations were measured, and then, the one that shows the highest accuracy was selected and analyzed further.

All the frameworks in Table 1 were applied for different sample combinations for the same total budget. Table 3 shows

Table 2 Input variables and environmental parameters

Input	Bounds	Description
R_w	[0.05, 0.15] m	Borehole radius
L	[1120,1680] m	Borehole length
K_w	[1500,15,000] m/yr	Hydraulic conductivity of borehole
Parameters	**Value**	**Description**
R	25,050 m	Radius of aquifer influence
T_u	89,335 m^2/yr	Transmissivity of upper aquifer
H_u	1050 m	Potentiometric head of upper aquifer
T_l	89.55 m^2/yr	Transmissivity of lower aquifer
H_l	760 m	Potentiometric head of lower aquifer

Table 3 Cases of sample size combinations for a total computational budget of evaluating 10 high-fidelity samples (10H) and ratio of 30 between the cost of high-fidelity and low-fidelity simulation (Borehole3 function)

Total budget	Sample cost ratio	High- and low-fidelity samples
10H	30	8/60, 7/90, 6/120, 5/150, 4/180, 3/210, 2/240

sample size ratios for the total budget of 10H, which means the computational budget for evaluating ten high-fidelity samples. The sample cost ratio of 30 means that the cost of evaluating 30 low-fidelity samples is equivalent to that of evaluating a single high-fidelity sample. With the total budge of 10H, we can use either 10 high-fidelity samples, 300 low-fidelity samples, or any combinations as shown in Table 3. These combinations are expressed with the numbers of high- (n_H) and low- (n_L) fidelity samples, such as 7/90.

For each sample size ratio, 100 DOEs were randomly generated using the nearest neighbor sampling method (Forrester et al. 2007), and the statistics of their RMSEs are used for evaluating the performance of each MFS framework. Note that RMSE of each DOE was calculated based on 100,000 test points in the sampling domain. For each sample size ratio and MFS framework, the median, 25 and 75 percentiles of RMSEs were obtained.

Figure 3 shows the median RMSEs of all five frameworks for different sample size ratios. Since the low-fidelity Kriging surrogate used 300 samples, the prediction error is small against the true low-fidelity function, but RMSE is high against the true high-fidelity function. On the other hand, the error in the high-fidelity Kriging surrogate comes from the prediction error because 10 high-fidelity samples are too

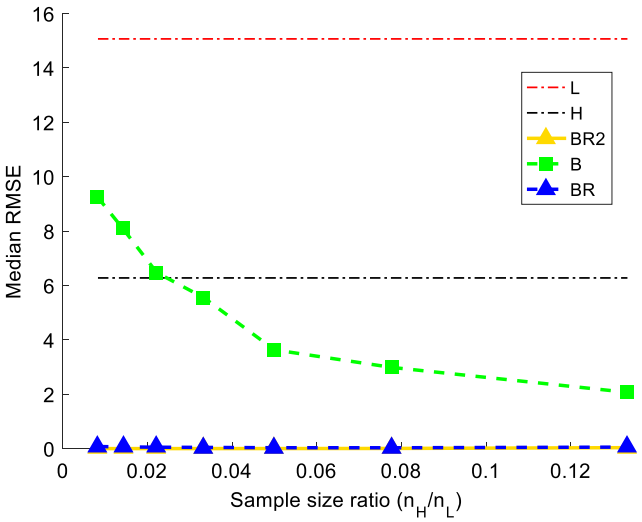


Fig. 3 The median (of 100 DOEs) accuracy for different sample size ratios (Note that the BR2 curve is overlapped with the BR curve) (Borehole example)

few. In general, the RMSEs of the MFS frameworks were significantly lower than that of single-fidelities.

The difference between these MFS frameworks is from the contributions of ρ and the criteria for finding ρ. BR and BR2 significantly outperformed B. That indicates that the inclusion of ρ is a key factor for the accuracy in this example. In addition, the difference in finding ρ (BR and BR2) did not lead to a significant difference, which means that the different criteria for finding ρ did not change the results. In the case of the sample size ratio of 7/90 (BR and BR2 were most accurate at this ratio), the estimated ρ from 100 DOEs has the mean of 1.25 and standard deviation of 9.3×10^{-7}; that is, the effect of different DOEs is negligible. In this example, the directions of finding ρ for reducing bumpiness (BR) and maximizing agreement (BR2) are consistent, which is not always true. The reason will be discussed in a later section.

Figure 4(a), (b) and (c) show the graphs of bumpiness, variation, and waviness of the true discrepancy function with respect to ρ. In Fig. 4(d), the MFS accuracy was calculated based on 100 DOEs of 7/90 at which the prediction accuracy of BR is closest to the minimum median RMSE of BR shown in Fig. 3. From Fig. 4(a), $\rho = 1.25$ minimizes the bumpiness, which is consistent with the mean of the estimated ρ from BR; that is, BR found ρ that is identical to minimizing the bumpiness. The bumpiness behavior is related to the variation and waviness behavior. Figure 4(b) and (c) show that ρ affects the variation but not the waviness of the discrepancy function, so all the changes in the bumpiness are due to variation. Figure 4(d) shows that the MFS accuracy as well as its uncertainty is the best when the bumpiness is minimum. In summary, BR found $\rho = 1.25$ by minimizing the bumpiness, or equivalently, by minimizing variation. Such behavior is related to the characteristics that both the high- and low-fidelity functions are convex.

Figure 5 illustrates a prediction comparison between BR and B along the line connecting $\mathbf{x}_1 = \{0.2, 1120, 1500\}$ and $\mathbf{x}_2 = \{0, 1680, 15{,}000\}$ for a chosen DOE. Note that the characteristics along other lines are similar to this one. Since the low-fidelity prediction was very accurate, the prediction accuracy was determined by the error in the discrepancy function. As the result shows, seven high-fidelity samples could not capture the curvature of the discrepancy in Fig. 5(b) without ρ (or $\rho = 1$). ρ reduced the variation of the discrepancy function and so does the bumpiness as shown in Fig. 5(a); and it increased the MFS prediction accuracy significantly. Note that the magnitude of the discrepancy function in Fig. 5(a) and (b) are different by a factor of about 100. Since the variation of the discrepancy function was reduced so much, the errors in fitting the discrepancy function have also been reduced by two orders of magnitude. The comparison between Fig. 5(c) and (e) shows that the high-fidelity response has a similar trend with the low-fidelity response. Therefore, magnifying

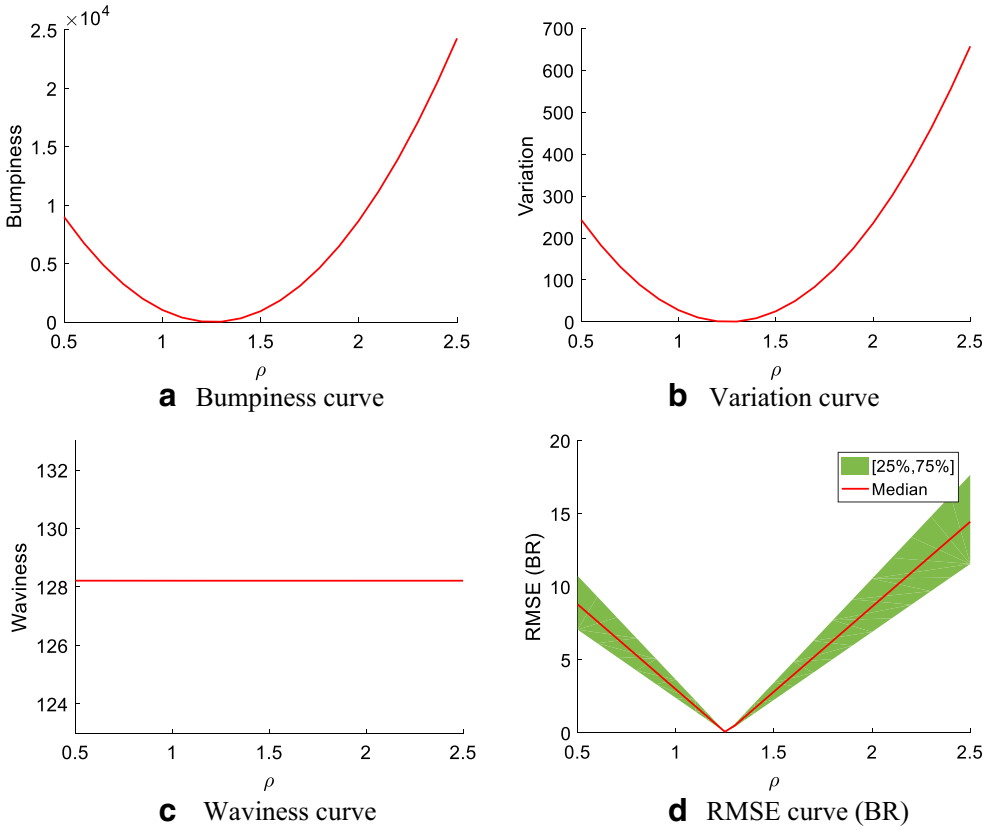
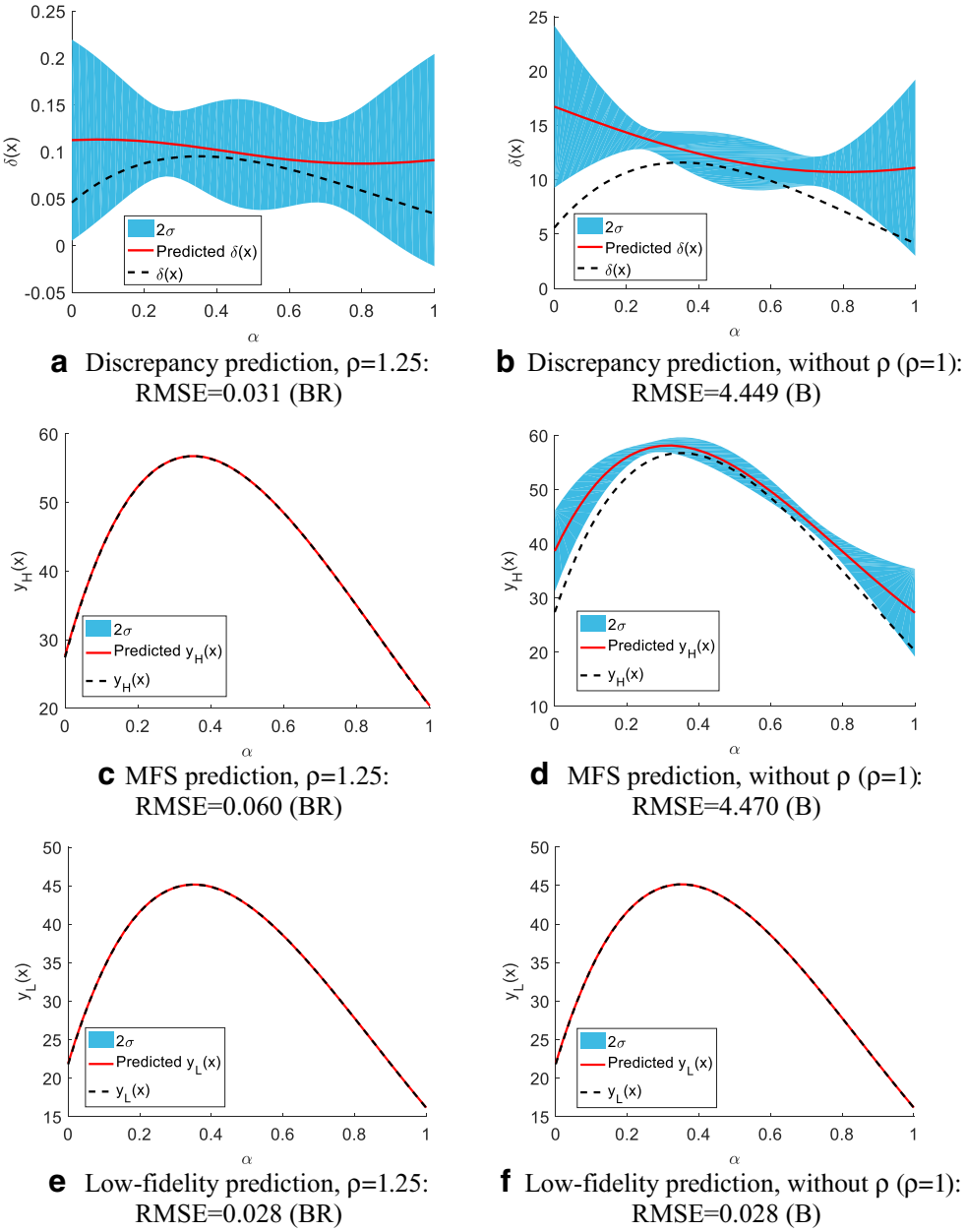


Fig. 4 The bumpiness, variation and waviness graphs and the RMSE from BR in terms of ρ for the borehole example



a Discrepancy prediction, ρ=1.25: RMSE=0.031 (BR)

b Discrepancy prediction, without ρ (ρ=1): RMSE=4.449 (B)

c MFS prediction, ρ=1.25: RMSE=0.060 (BR)

d MFS prediction, without ρ (ρ=1): RMSE=4.470 (B)

e Low-fidelity prediction, ρ=1.25: RMSE=0.028 (BR)

f Low-fidelity prediction, without ρ (ρ=1): RMSE=0.028 (B)

Fig. 5 Comparisons between the predictions based on the BR and the B on the line between $\mathbf{x}_1 = \{0.2, 1120, 1500\}$ and $\mathbf{x}_2 = \{0, 1680, 15{,}000\}$ and RMSEs along the line (Borehole example)

the low-fidelity function reduces the variation of the discrepancy function.

It is recalled that the performances of the BR and the BR2 were almost identical. This is because for this example, the reduction in variation of the discrepancy function is achieved by scaling down its magnitude. However, this is not always the case, as the Hartman 6 example will show.

4.2 Hartmann 6 function example

The Hartmann 6 function example also shows that Bayesian frameworks with ρ increase the prediction accuracy of MFS. However, in contrast to the borehole example, the BR2 did not give as good prediction as BR, which confirms that reducing bumpiness was more effective than minimizing error. For a high-fidelity function, the six-dimensional Hartmann 6 function is defined as

$$f_H(\mathbf{x}) = -\frac{1}{1.94}\left(2.58 + \sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{6} A_{ij}\left(x_j - P_{ij}\right)^2\right)\right) \tag{12}$$

where the domain of input variables is defined as $\{0.1, \ldots 0.1\} \le \mathbf{x} \le \{1, \ldots, 1\}$. For model parameters, $\boldsymbol{\alpha} = \{1\ 1.2\ 3\ 3.2\}^T$, and the following two constant matrices are used:

$$\mathbf{A} = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix} \quad \text{and}$$

$$\mathbf{P} = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix}.$$

The approximated Hartmann 6 function was invented to be used as a low-fidelity function as

$$f_L(\mathbf{x}) = -\frac{1}{1.94}\left(2.58 + \sum_{i=1}^{4} \alpha_i' f_{\exp}\left(-\sum_{j=1}^{6} A_{ij}\left(x_j - P_{ij}\right)^2\right)\right) \tag{13}$$

where $\boldsymbol{\alpha}' = \{0.5 \quad 0.5 \quad 2.0 \quad 4.0\}^T$ and $f_{\exp}(x)$ is the approximated exponential function as

$$f_{\exp}(x) = \left(\exp\left(\frac{-4}{9}\right) + \exp\left(\frac{-4}{9}\right)\frac{(x+4)}{9}\right)^9 \tag{14}$$

Note that the total function variation of the Hartmann 6 function is 0.33 and the RMSE of the low-fidelity function is 0.11.

In this example, the total computational budget is the cost of evaluating 56 high-fidelity samples (56H), and the sample cost ratio between high- and low-fidelity functions is 30. Table 4 shows the considered sample size ratios. The notations and the repetitions of DOE are the same as the previous example.

Figure 6 shows the median of RMSEs of all the frameworks for different sample size ratios. In this example, BR outperformed both B and BR2, which shows that not only the inclusion of ρ but also reducing the bumpiness is important for prediction accuracy. Finding ρ by reducing bumpiness yielded much more accurate prediction than by minimizing error.

Figure 7(a) and (b) show the histograms of ρ estimated from BR and BR2 for the sample size ratio of 42/420, at which the BR is the most accurate. The histograms clearly show they estimated significantly different ρ. The mode of the histogram was 1.49 for BR, while it was 1.03 for BR2. Since there is no difference in making low-fidelity predictions between the BR and the BR2, the difference between the two frameworks is caused by the difference in the ways of finding ρ.

Table 4 Cases of sample size combinations for a total computational budget of evaluating 56 high-fidelity samples (56H) and sample cost ratio of 30 (Hartmann6 example)

Total budget	Sample size ratio n_H/n_L
56H	48/240, 46/300, 44/360, 42/420, 40/480, 38/540, 28/840, 18/1140

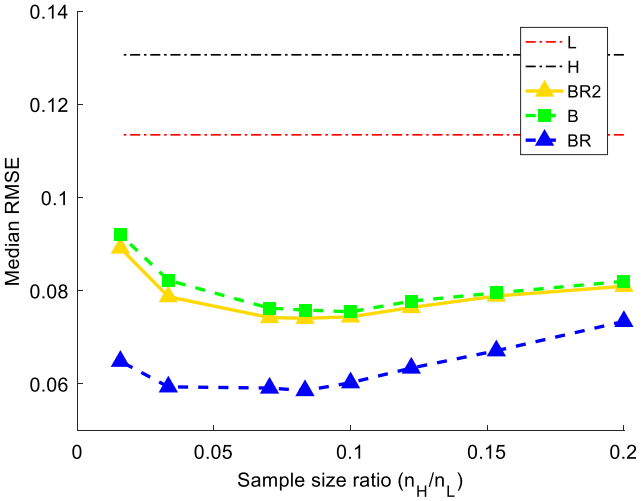


Fig. 6 The median (of 100 DOEs) RMSEs versus sample size ratio (Hartman6 example)

The graphs of MFS discrepancy bumpiness, variation and waviness with respect to ρ are shown in Fig. 8 as well as the graph of prediction accuracy for the sample size ratio of 42/420. Figure 8(a) shows the bumpiness graph of the true discrepancy function, where the minimum bumpiness occurred at $\rho = 1.41$. The mode of ρ from BR (1.49) is close to ρ at the minimum bumpiness, which indicates that BR found ρ, which reduces the bumpiness as discussed in Section 3. The contributions of the variation and the waviness are shown as a graph in Fig. 8(b) and (c). It shows that the bumpiness is strongly correlated with the variation, while the waviness shows an opposite behavior, but its contribution is overwhelmed by that of the variation. Figure 8(d) shows the RMSE of BR for varying ρ. Note that the corresponding RMSE graph of BR2 is identical to that of BR. That means, BR and BR2 gave identical predictions for the same ρ. The difference between BR and BR2 shown in Fig. 6 was because they used different ρ estimates. RMSE is closely correlated with the bumpiness, where the maximum accuracy occurred at $\rho = 1.55$. ρ at the minimum variation (1.55 from Fig. 8(b)) is consistent with ρ at the minimum RMSE (1.55 from Fig. 8(d)).

The result indicates that bumpiness reduction was effective to reduce prediction error. However, minimizing bumpiness is not equivalent to maximizing accuracy. An explanation of the observation is that the bumpiness of the true discrepancy function is compared with the accuracy of the predictions based on samples. Since there are infinite true functions for a given set of samples, the bumpiness of the true function cannot be perfectly correlated with the error.

In order to visualize the effect of different variation reductions between BR with BR2, a DOE was chosen for the sample size ratio of 42/420. Figure 9 shows predictions along a line in the sampling domain that has the maximum difference between BR and BR2. The low-fidelity prediction is reasonably accurate with RMSE of 0.048 for both frameworks (Fig. 9(e) and (f)). Therefore, the error in MFS mostly comes

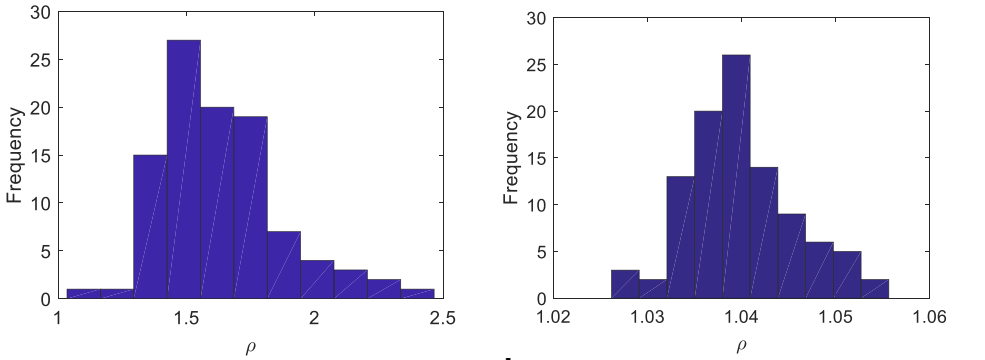


Fig. 7 Histogram of ρ estimates (Hartman6 example)

a Histogram of ρ for BR (mode of 1.49) **b** Histogram of ρ for BR2 (mode of 1.03)

from the error in the discrepancy function. This is because 42 high-fidelity samples are not sufficient to capture the bumpy behavior of the discrepancy in the six-dimensions. BR determined $\rho = 1.48$ by minimizing the bumpiness, while BR2 found $\rho = 1.04$ by minimizing the error.

As in the borehole function, Figs. 9(a) and (b) show that the fit along the line is poor for both BR and BR2. However, the improvement by BR is not accomplished by reducing the magnitude of the discrepancy but by reducing its variation. The RMSEs along the line are much higher than the median RMSEs of the whole sampling domain shown in Fig. 6. This is because the line is a tiny part of the domain. However, in terms of the RMSE reduction, they are consistent: 23% reduction for the line and 20% reduction for the whole domain.

5 Concluding remarks

This paper discussed that the Bayesian discrepancy framework uses the low-fidelity scaling scalar to reduce variation and waviness of the discrepancy function through the likelihood based on the Gaussian process model. The variation and waviness reductions lead to reduction of bumpiness that combines the two without a Gaussian process model. The importance of including the low-fidelity scaling factor is that it allows to reduce bumpiness of the discrepancy function that tends to reduce error as the examples show. For the examples studied, the success of the Bayesian framework was largely based on the use of the scale factor. Without the scalar, the Bayesian method gave mediocre predictions. The three-

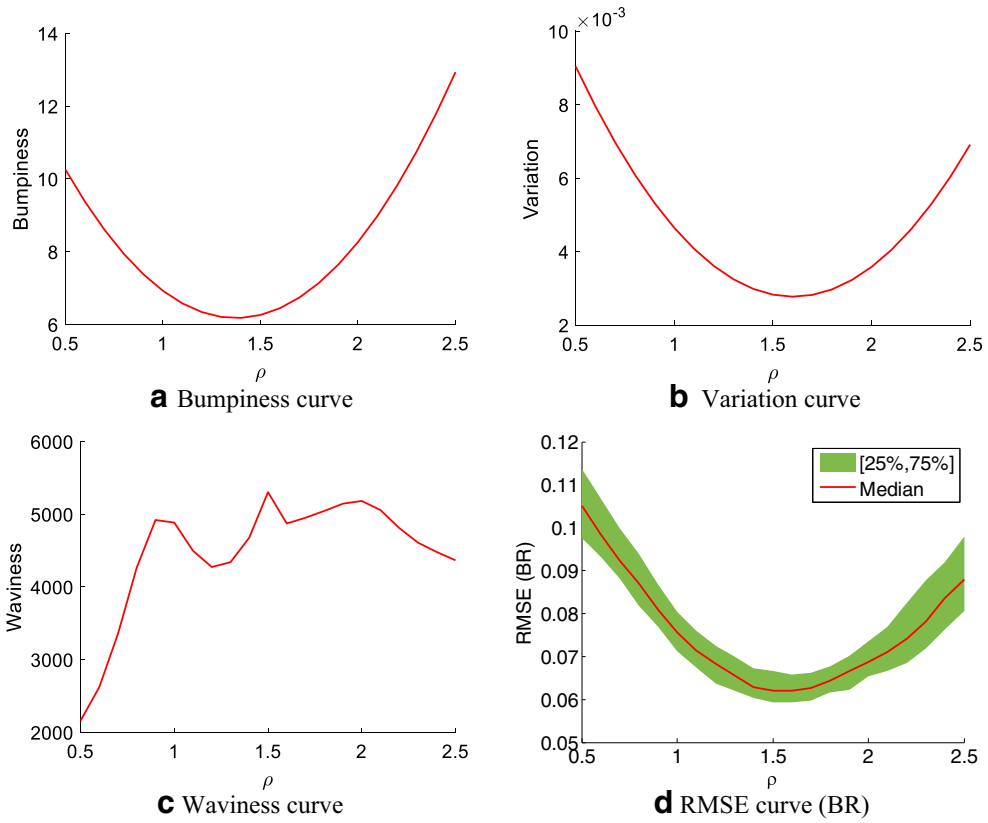


Fig. 8 The bumpiness, variation and waviness graphs and the RMSE from BR in terms of ρ for the Hartman6 example

a Bumpiness curve **b** Variation curve **c** Waviness curve **d** RMSE curve (BR)

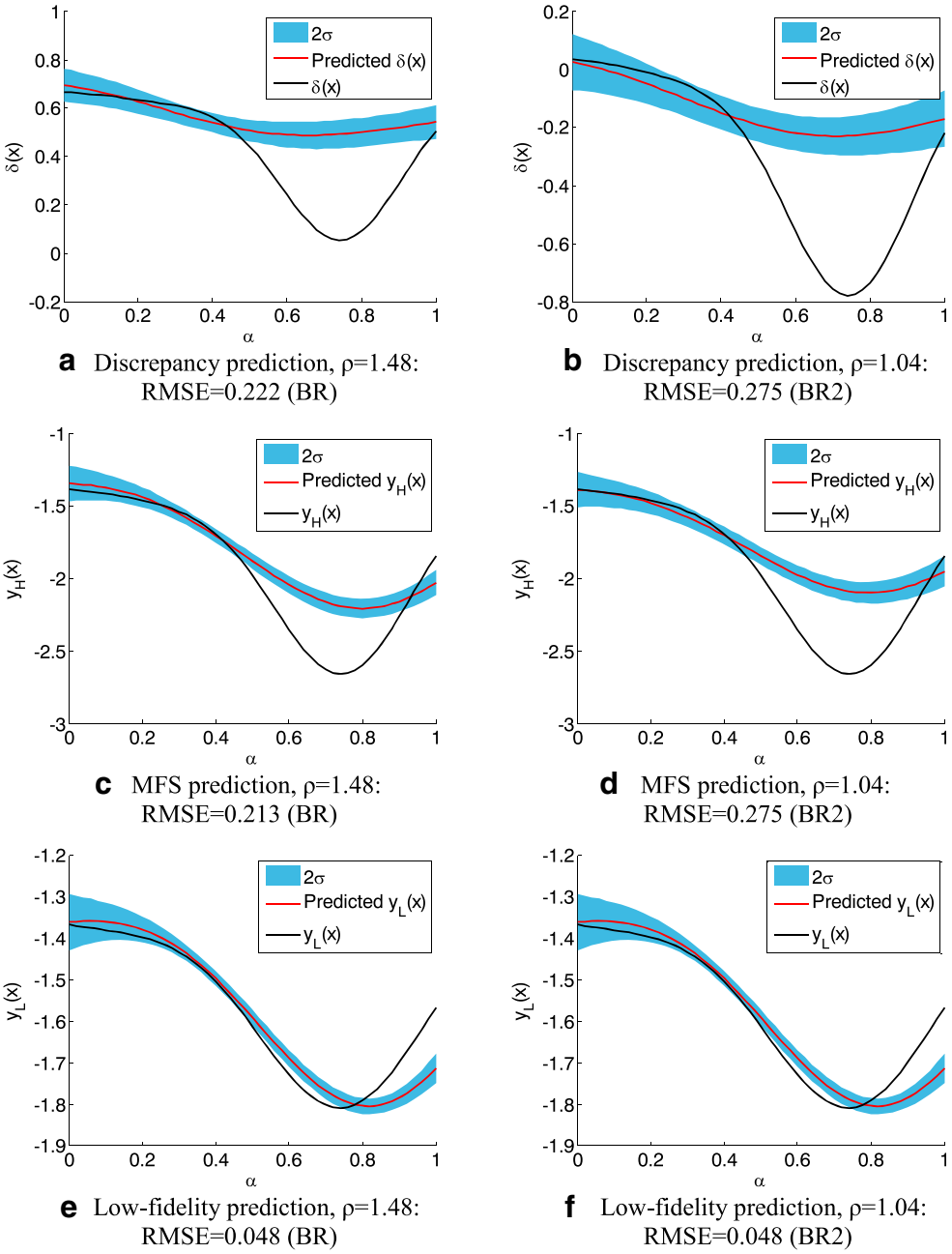


Fig. 9 Comparisons between the BR and the BR2 on hyper-line between {0.35,0.32,0.63,0.14,0.88,0.10} and {0.30,0.39,0.31,0.36,0.10,0.80} and RMSEs along the line (Hartman6 example)

dimensional borehole and the six-dimensional Hartmann6 examples demonstrated that the accuracy of the MFS predictions was strongly correlated with the bumpiness of the discrepancy function. For the Borehole3 example, the minimum RMSE was achieved with the scalar minimizing bumpiness. For the Hartmann6 example, the scalar minimizing RMSE was not identical to the scalar minimizing the bumpiness but they were very close and the behaviors of RMSE and bumpiness were strongly correlated. However, a perfect correlation cannot be expected between the bumpiness of the true function and the prediction error due to infinite possible true functions passing through a finite number of samples.

The Bayesian framework characterizes a discrepancy function with two factors variation and waviness through the Gaussian process model. And the maximum likelihood method combines variation and waviness reduction. Bumpiness is another way to combine them without using a Gaussian process model. For the examples using the Bayesian framework, variation reduction dominated bumpiness reduction for the Hartmann6 and Borehole3 examples. That can be interpreted as the low-fidelity function captured the trend of the high-fidelity function, but it did not capture the high-frequency behavior. Whereas waviness reduction requires the scaled low-fidelity model to capture the high-frequency behavior of

the high-fidelity model without necessarily capturing the low-frequency behavior. We suspect that such a case may be rare, so reducing variation would be more common. The lessons learned from the Bayesian framework can be utilized for other MFS predictions.

Acknowledgements This work is supported by the U.S. Department of Energy, National Nuclear Security Administration, Advanced Simulation and Computing Program, as a Cooperative Agreement under the Predictive Science Academic Alliance Program, under Contract No. DE-NA0002378.

Appendix A: Bayesian MFS framework

The discrepancy function based Bayesian framework (equivalent to co-Kriging) predicts a high-fidelity response based on a scaled low fidelity prediction and the corresponding discrepancy prediction, $\rho\hat{y}_L(\mathbf{x})$ and $\hat{\delta}(\mathbf{x})$, respectively, as

$$\hat{y}_H(\mathbf{x}) = \rho\hat{y}_L(\mathbf{x}) + \hat{\delta}(\mathbf{x}) \tag{15}$$

where ρ is a scalar for the low-fidelity response. Both predictions are based on a low fidelity Gaussian process (GP) model $Y_L(\mathbf{x})$ and a discrepancy function GP model $\Delta(\mathbf{x})$, respectively. MFS prediction is made with the high-fidelity GP model, which is the sum of the two GP models as

$$Y_H(\mathbf{x}) = \rho Y_L(\mathbf{x}) + \Delta(\mathbf{x}) \tag{16}$$

MFS prediction is made with two steps: 1) estimating hyperparameters of the GP models and ρ using the Bayesian inference and 2) updating the high-fidelity GP models defined with the estimates using data. The MFS predictor is the mean of the updated high-fidelity GP model $Y_H(\mathbf{x}) \mid \mathbf{y}_H, \mathbf{y}_L$, which is expressed as

$$\hat{y}_H(\mathbf{x}) = E(Y_H(\mathbf{x})|\mathbf{y}_H, \mathbf{y}_L) \tag{17}$$

where $\mathbf{y}_H$ and $\mathbf{y}_L$ are, respectively, the vectors of high- and low-fidelity sample sets. The corresponding prediction uncertainty estimate is

$$\hat{\sigma}^2_H(\mathbf{x}) = Var(Y_H(\mathbf{x})|\mathbf{y}_H, \mathbf{y}_L) \tag{18}$$

Gaussian process models

The GP models assume that the prediction uncertainties (epistemic uncertainty) follow normal distributions. The low-fidelity and discrepancy GP models are respectively defined at $\mathbf{x}$ as

$$\begin{aligned} Y_L(\mathbf{x}) &\sim N\left(\boldsymbol{\xi}_L(\mathbf{x})\boldsymbol{\beta}_L, \sigma_L^2\right) \\ \Delta(\mathbf{x}) &\sim N\left(\boldsymbol{\xi}_\Delta(\mathbf{x})\boldsymbol{\beta}_\Delta, \sigma_\Delta^2\right) \end{aligned} \tag{19}$$

where the subscript L and Δ denote low-fidelity function and discrepancy function, respectively. Since the GP models are functions of $\mathbf{x}$, the GP models are defined with mean functions, where a polynomial regression model is often employed. $\boldsymbol{\xi}_L(\mathbf{x})$ and $\boldsymbol{\xi}_\Delta(\mathbf{x})$ are basis vectors of the mean functions at $\mathbf{x}$ and $\boldsymbol{\beta}_L$ and $\boldsymbol{\beta}_\Delta$ are coefficient vectors. Note that the unknown coefficient vectors are to be estimated based on data. Equation (24) and (31) are the estimators of the coefficient vectors. The variances remain constant but the GP models define the relation between responses at two different points with covariance functions based on the distance between the points, which are defined as

$$\begin{aligned} \mathrm{cov}\left(Y_L(\mathbf{x}), Y_L\left(\mathbf{x}'\right)\right) &= \sigma_L^2 corr\left(Y_L(\mathbf{x}), Y_L\left(\mathbf{x}'\right)\right) \\ \mathrm{cov}\left(\Delta(\mathbf{x}), \Delta\left(\mathbf{x}'\right)\right) &= \sigma_\Delta^2 corr\left(\Delta(\mathbf{x}), \Delta\left(\mathbf{x}'\right)\right) \end{aligned} \tag{20}$$

where σ_L^2 and σ_Δ^2 are process variances of the GP models of the low-fidelity and discrepancy function, respectively. $corr(Y_L(\mathbf{x}), Y_L(\mathbf{x}'))$ and $corr(\Delta(\mathbf{x}), \Delta(\mathbf{x}'))$ are correlation functions. A Gaussian kernel is used for the correlation functions as

$$\begin{aligned} corr\left(Y_L(\mathbf{x}), Y_L\left(\mathbf{x}'\right)\right) &= \exp\left(-\left(\mathbf{x}-\mathbf{x}'\right)^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega}_L)\left(\mathbf{x}-\mathbf{x}'\right)\right) \\ corr\left(\Delta(\mathbf{x}), \Delta\left(\mathbf{x}'\right)\right) &= \exp\left(-\left(\mathbf{x}-\mathbf{x}'\right)^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega}_\Delta)\left(\mathbf{x}-\mathbf{x}'\right)\right) \end{aligned} \tag{21}$$

where $\mathrm{diag}(\boldsymbol{\omega}_L)$ and $\mathrm{diag}(\boldsymbol{\omega}_\Delta)$ are diagonal matrices. Their (i,i) component is the i^{th} component of $\boldsymbol{\omega}_L$ and $\boldsymbol{\omega}_\Delta$, respectively.

Finally, the high fidelity GP model is obtained as the sum of the GP models as

$$Y_H(\mathbf{x}) \sim N\left(\rho\boldsymbol{\xi}_L^T(\mathbf{x})\boldsymbol{\beta}_L + \boldsymbol{\xi}_\Delta^T(\mathbf{x})\boldsymbol{\beta}_\Delta, \rho^2\sigma_L^2 + \sigma_\Delta^2\right) \tag{22}$$

The covariance function of the GP model is

$$\begin{aligned} \mathrm{cov}\left(Y_H(\mathbf{x}), Y_H\left(\mathbf{x}'\right)\right) &= \rho^2\sigma_L^2 corr\left(Y_L(\mathbf{x}), Y_L\left(\mathbf{x}'\right)\right) \\ &\quad + \sigma_\Delta^2 corr\left(\Delta(\mathbf{x}), \Delta\left(\mathbf{x}'\right)\right) \end{aligned} \tag{23}$$

For making MFS prediction, $\{\boldsymbol{\beta}_L, \boldsymbol{\omega}_L, \sigma_L, \boldsymbol{\beta}_\Delta, \boldsymbol{\omega}_\Delta, \sigma_\Delta, \rho\}$ of the high fidelity GP model need to be estimated.

Estimating hyper parameters of the GP models and ρ

A common condition for MFS sampling is that high fidelity sampling locations are selected from low fidelity sampling locations. With the condition, the discrepancy can be directly obtained at the common locations, i.e. at the high-fidelity sample locations. For the Bayesian framework, the condition allows computational advantage of estimating the parameters.

The condition yields independent estimations of $\{\boldsymbol{\beta}_L, \boldsymbol{\omega}_L, \sigma_L\}and\{\boldsymbol{\beta}_\Delta, \boldsymbol{\omega}_\Delta, \sigma_\Delta, \rho\}$. This is because the discrepancy function prediction does not depend on the low fidelity prediction that makes$\{\boldsymbol{\beta}_\Delta, \boldsymbol{\omega}_\Delta, \sigma_\Delta, \rho\}$estimation independent to $\{\boldsymbol{\beta}_L, \boldsymbol{\omega}_L, \sigma_L\}$estimation.

The Bayesian inference uses the GP model inversely to obtain the likelihood function with respect to$\{\boldsymbol{\beta}_L, \boldsymbol{\omega}_L, \sigma_L\}$. Fortunately, $\{\boldsymbol{\beta}_L, \sigma_L\}$ can be analytically expressed as a function of $\boldsymbol{\omega}_L$. Thus, the likelihood function is reformulated as a function of $\boldsymbol{\omega}_L$only by substituting $\{\boldsymbol{\beta}_L, \sigma_L\}$:

$$p(\boldsymbol{\omega}_L|\mathbf{y}_L) = \frac{1}{\sqrt{(2\pi)^{n_L}\left|\hat{\sigma}_L^2(\boldsymbol{\omega}_L)\mathbf{R}_L(\boldsymbol{\omega}_L)\right|}} \exp\left(-\frac{1}{2\hat{\sigma}_L^2(\boldsymbol{\omega}_L)}\left(\mathbf{y}_L-\mathbf{X}_L\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L)\right)^T\mathbf{R}_L^{-1}(\boldsymbol{\omega}_L)\left(\mathbf{y}_L-\mathbf{X}_L\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L)\right)\right) \tag{24}$$

where $\left\{\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L), \hat{\sigma}_L^2(\boldsymbol{\omega}_L)\right\}$ are the analytical estimates of $\{\boldsymbol{\beta}_L, \sigma_L\}$for given $\boldsymbol{\omega}_L$. These estimates can be expressed as a function of$\boldsymbol{\omega}_L$:

$$\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L) = \left(\mathbf{X}_L^T\mathbf{R}_L^{-1}(\boldsymbol{\omega}_L)\mathbf{X}_L\right)^{-1}\mathbf{X}_L^T\mathbf{R}_L^{-1}(\boldsymbol{\omega}_L)\mathbf{y}_L \tag{25}$$

$$\hat{\sigma}_L^2(\boldsymbol{\omega}_L) = \frac{1}{n_L}\left(\mathbf{y}_L-\mathbf{X}_L\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L)\right)^T\mathbf{R}_L^{-1}(\boldsymbol{\omega}_L)\left(\mathbf{y}_L-\mathbf{X}_L\hat{\boldsymbol{\beta}}_L(\boldsymbol{\omega}_L)\right) \tag{26}$$

Correlation matrix$\mathbf{R}_L(\boldsymbol{\omega}_L)$is also a function of $\boldsymbol{\omega}_L$, which is defined as

$$\mathbf{R}_L(\boldsymbol{\omega}_\Delta) = \begin{bmatrix} 1 & \mathrm{L} & \exp\left(-(\mathbf{x}_{L,n_L}-\mathbf{x}_{L,1})^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega}_L)(\mathbf{x}_{L,n_L}-\mathbf{x}_{L,1})\right) \\ \mathrm{M} & \mathrm{O} & \mathrm{M} \\ symm & \mathrm{L} & 1 \end{bmatrix}_{(n_L\times n_L)} \tag{27}$$

where $\mathbf{x}_L = \left\{\mathbf{x}_{L,1}, \ldots, \mathbf{x}_{L,\mathrm{n}_L}\right\}$ is the vector of n_L low-fidelity sample locations. The moment matrix is defined by applying the sample locations to the basis vectors as

$$\mathbf{X}_L = \begin{Bmatrix} \xi_L^T(\mathbf{x}_{L,1}) \\ \vdots \\ \xi_L^T(\mathbf{x}_{L,n_L}) \end{Bmatrix}_{(n_L\times p_1)} \tag{28}$$

where p_1 is the number of basis of the mean function.

Using the maximum likelihood estimation, the mode of the likelihood function is used to determine $\boldsymbol{\omega}_L$. Thus, the likelihood function can be reformulated as long as the mode remains the same. By substituting (26) into the likelihood function and taking $2/\mathrm{n_L}^{\mathrm{th}}$ root, the likelihood function in (23) and (24) can be simplified as

$$p(\boldsymbol{\omega}_L|\mathbf{y}_L)\propto-\hat{\sigma}_L^2(\boldsymbol{\omega}_L)|\mathbf{R}_L(\boldsymbol{\omega}_L)|^{-1/n_L} \tag{29}$$

Finally, $\boldsymbol{\omega}_L$estimate is obtained by solving an optimization problem defined as

$$\underset{\boldsymbol{\omega}_\Delta}{\operatorname{argmin}} \quad \hat{\sigma}_L^2(\boldsymbol{\omega}_L)|\mathbf{R}_L(\boldsymbol{\omega}_L)|^{-1/n_L} \tag{30}$$

Note that Bayesian statistics uses the maximum a posteriori (MAP) estimation, which takes the mode of the posterior distribution as a parameter estimate. However, in the case of non-informative prior, which is the case we used, the MAP estimation gives the same estimate with the method of maximum likelihood estimation. Because of this reason, we simply describe that hyper parameters were estimated with the method of maximum likelihood in this paper.

In the same sense, $\{\boldsymbol{\beta}_\Delta, \boldsymbol{\omega}_\Delta, \sigma_\Delta, \rho\}$are estimated using the discrepancy GP model. Since$\{\boldsymbol{\beta}_\Delta, \sigma_\Delta\}$can be analytically obtained for given$\{\boldsymbol{\omega}_\Delta, \rho\}$, the likelihood function is expressed as

$$p(\boldsymbol{\omega}_\Delta, \rho|\mathbf{y}_H, \mathbf{y}_L^c) = \frac{1}{\sqrt{(2\pi)^{n_H}\left|\hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho)\mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta)\right|}} \exp\left[-\frac{1}{2\hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho)}\left\{\begin{pmatrix}\mathbf{y}_H-\rho\mathbf{y}_L^c-\mathbf{X}_\Delta\hat{\boldsymbol{\beta}}_\Delta(\boldsymbol{\omega}_\Delta, \rho)\end{pmatrix}^T\mathbf{R}_\Delta^{-1}(\boldsymbol{\omega}_\Delta)\begin{pmatrix}\mathbf{y}_H-\rho\mathbf{y}_L^c-\mathbf{X}_\Delta\hat{\boldsymbol{\beta}}_\Delta(\boldsymbol{\omega}_\Delta, \rho)\end{pmatrix}\right\}\right] \tag{31}$$

where $\mathbf{y}_L^c$ is a subset of the low fidelity data at the common data locations. The analytical estimates of $\{\boldsymbol{\beta}_\Delta, \sigma_\Delta\}$ are

$$\hat{\boldsymbol{\beta}}_\Delta(\boldsymbol{\omega}_\Delta, \rho) = \left(\mathbf{X}_\Delta^T\mathbf{R}_\Delta^{-1}(\boldsymbol{\omega}_\Delta)\mathbf{X}_\Delta\right)^{-1}\mathbf{X}_\Delta^T\mathbf{R}_\Delta^{-1}(\boldsymbol{\omega}_\Delta)\left(\mathbf{y}_H-\rho\mathbf{y}_L^c\right) \tag{32}$$

$$\hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho) = \frac{1}{n_y}\left(\mathbf{y}_H-\rho\mathbf{y}_L^c-\mathbf{X}_\Delta\hat{\boldsymbol{\beta}}_\Delta(\boldsymbol{\omega}_\Delta, \rho)\right)^T\mathbf{R}_\Delta^{-1}(\boldsymbol{\omega}_\Delta)\left(\mathbf{y}_H-\rho\mathbf{y}_L^c-\mathbf{X}_\Delta\hat{\boldsymbol{\beta}}_\Delta(\boldsymbol{\omega}_\Delta, \rho)\right) \tag{33}$$

$\mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta)$ is defined as

$$\mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta) = \begin{bmatrix} 1 & \mathrm{L} & \exp\left(-(\mathbf{x}_{H,n_H}-\mathbf{x}_{H,1})^{\mathrm{T}}\mathrm{diag}(\boldsymbol{\omega}_\Delta)(\mathbf{x}_{H,n_H}-\mathbf{x}_{H,1})\right) \\ & \mathrm{O} & \mathrm{M} \\ symm & & 1 \end{bmatrix}_{(n_H\times n_H)} \tag{34}$$

where $\mathbf{x}_H = \left\{\mathbf{x}_{H,1}, \ldots, \mathbf{x}_{H,\mathrm{n}_H}\right\}$ is the vector of $n_H$high-fidelity sample locations. The moment matrix is defined by applying the sample locations to the basis vectors as

$$\mathbf{X}_\Delta = \begin{Bmatrix} \xi_\Delta^T(\mathbf{x}_{\Delta,1}) \\ \vdots \\ \xi_\Delta^T(\mathbf{x}_{\Delta,n_L}) \end{Bmatrix}_{(n_\Delta\times p)} \tag{35}$$

By substituting (32) and (33) into the exponential term of the likelihood function in (31) and taking $2/n_H^{th}$ root, the likelihood function is simplified as

$$p(\boldsymbol{\omega}_\Delta, \rho | \mathbf{y}_H, \mathbf{y}_L^c) \propto -\hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho) |\mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta)|^{-1/n_H} \tag{36}$$

Finally, $\boldsymbol{\omega}_\Delta$and ρ estimates are obtained by solving an optimization problem defined as

$$\underset{\boldsymbol{\omega}_\Delta, \rho}{\text{argmin}} \quad \hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho) |\mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta)|^{-1/n_H} \tag{37}$$

The objective function is the simplified negative maximum likelihood function of $\{\boldsymbol{\omega}_\Delta, \rho\}$. One interesting aspect of this problem is the presence of ρ. For a fixed ρ, (37) turns a problem of finding $\boldsymbol{\omega}_\Delta$for fitting$\mathbf{y}_H - \rho \mathbf{y}_L^c$ that is the same with finding hyperparameters for fitting a Kriging surrogate for $\mathbf{y}_H - \rho \mathbf{y}_L^c$ (Lophaven et al. 2002).

By assuming $\hat{\sigma}_\Delta^2(\rho)$ and $\mathbf{R}_\Delta$are the process standard deviation and correlation matrix for $\mathbf{y}_H - \rho \mathbf{y}_L^c$, (37) can be turned into

$$\underset{\rho}{\text{argmin}} \quad \hat{\sigma}_\Delta^2(\rho) |\mathbf{R}_\Delta|^{-1/n_H} \tag{38}$$

The objective function if the minimized negative likelihood function of the Kriging surrogate for $\mathbf{y}_H - \rho \mathbf{y}_L^c$. Equation (38) can be interpreted as a problem for finding ρ minimizing the negative Kriging likelihood function.

Equation (38) is identical to (5). With the fact that $\hat{\sigma}_\Delta^2(\rho)$ and $|\mathbf{R}_\Delta|$represent the variance and waviness estimated based on $\mathbf{y}_H - \rho \mathbf{y}_L^c$, (38) is to find ρ that minimizes the bumpiness of discrepancy data $\mathbf{y}_H - \rho \mathbf{y}_L^c$.

High fidelity function prediction

When all the hyperparameters are estimated, MFS prediction at a chosen point $\mathbf{x}$ is made by updating the GP model defined with the estimated hyperparameters. Posterior distribution of the prediction is obtained with (17).

The predictor and prediction uncertainty estimate are the mean and variance of the posterior distribution, respectively, expressed as

$$\begin{aligned} \hat{y}_H(\mathbf{x}) &= \boldsymbol{\xi}(\mathbf{x})^T \hat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) \\ \hat{\sigma}_{y_H}^2(\mathbf{x}) &= \rho^2 \hat{\sigma}_L^2(\boldsymbol{\omega}_L) + \hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho) - \mathbf{t}(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \mathbf{t}(\mathbf{x}) \\ &+ \left(\boldsymbol{\xi}(\mathbf{x}) - \mathbf{t}(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^T \left(\mathbf{X}^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right)^{-1} \left(\boldsymbol{\xi}(\mathbf{x}) - \mathbf{t}(\mathbf{x})^T \boldsymbol{\Sigma}^{-1} \mathbf{X}\right) \end{aligned} \tag{39}$$

where $\boldsymbol{\Sigma}^{-1}$is the inverse matrix of the covariance matrix expressed in (42).

The vectors and matrixes used in (39) are defined as follows.

$$\hat{\boldsymbol{\beta}}^T = \left\{ \hat{\boldsymbol{\beta}}_L^T \quad \hat{\boldsymbol{\beta}}_\Delta^T \right\} \tag{40}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_L & \mathbf{0} \\ \rho \mathbf{X}_L & \mathbf{X}_\Delta \end{bmatrix} \tag{41}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \hat{\sigma}_L^2(\boldsymbol{\omega}_L) \mathbf{R}_L(\boldsymbol{\omega}_L) & \rho \hat{\sigma}_L^2(\boldsymbol{\omega}_L) \mathbf{R}_{LH}(\boldsymbol{\omega}_L)^T \\ \rho \hat{\sigma}_L^2(\boldsymbol{\omega}_L) \mathbf{R}_{LH}(\boldsymbol{\omega}_L) & \rho^2 \hat{\sigma}_L^2(\boldsymbol{\omega}_L) \mathbf{R}_L(\boldsymbol{\omega}_L) + \hat{\sigma}_\Delta^2(\boldsymbol{\omega}_\Delta, \rho) \mathbf{R}_\Delta(\boldsymbol{\omega}_\Delta) \end{bmatrix} \tag{42}$$

where the correlation matrix between the high and low fidelity GP models $\mathbf{R}_{LH}(\boldsymbol{\omega}_L)$is expressed as

$$\mathbf{R}_{LH}(\boldsymbol{\omega}_L) = \begin{bmatrix} 1 & \text{L} & corr\left(Y_L(\mathbf{x}_{L,n_L}), Y_L(\mathbf{x}_{H,1})\right) \\ \text{M} & \text{O} & \text{M} \\ corr\left(Y_L(\mathbf{x}_{L,1}), Y_L(\mathbf{x}_{H,n_H})\right) & \text{L} & 1 \end{bmatrix}_{(n_L \times n_H)} \tag{43}$$

$$\boldsymbol{\xi}(\mathbf{x}) = \begin{bmatrix} \boldsymbol{\xi}_L(\mathbf{x}) \\ \boldsymbol{\xi}_\Delta(\mathbf{x}) \end{bmatrix}_{(n_L + n_H) \times 1} \tag{44}$$

$\mathbf{t}(\mathbf{x})$is expressed as

$$\mathbf{t}(\mathbf{x}) = \begin{bmatrix} \text{cov}\left(Y_H(\mathbf{x}), Y_L(\mathbf{x}_{L,1})\right) \\ \text{M} \\ \text{cov}\left(Y_H(\mathbf{x}), Y_L(\mathbf{x}_{L,n_L})\right) \\ \text{cov}\left(Y_H(\mathbf{x}), Y_H(\mathbf{x}_{H,1})\right) \\ \text{M} \\ \text{cov}\left(Y_H(\mathbf{x}), Y_H(\mathbf{x}_{H,n_H})\right) \end{bmatrix}_{(n_L + n_H) \times 1} \tag{45}$$

where $\text{cov}(Y_H(\mathbf{x}), Y_L(\mathbf{x}'))and\text{cov}(Y_H(\mathbf{x}), Y_H(\mathbf{x}'))$are defined in (20) and (21).

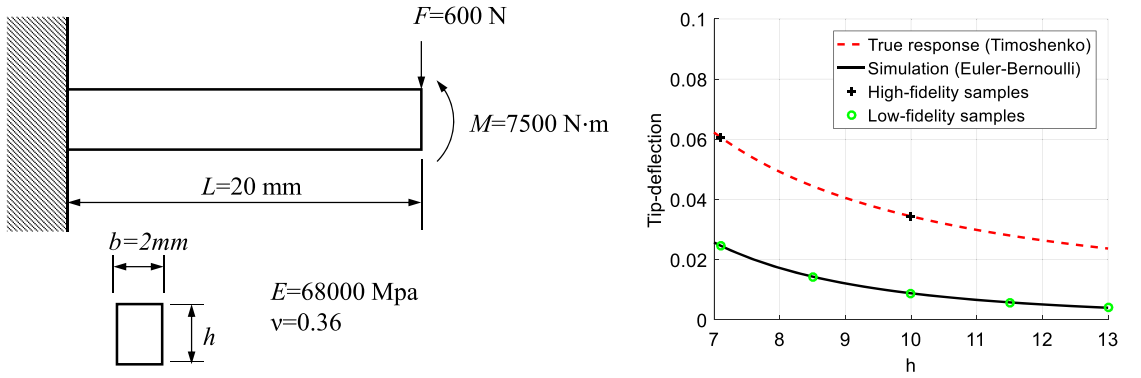
Appendix B: Cantilever beam example

A cantilever beam example is chosen as a structural example to show how ρ simplifies the discrepancy function to maximize MFS prediction accuracy. Figure 10(a) shows the geometry of a cantilever beam with a rectangular section under a concentrated force and moment. By choosing the height of the beam as a variable, it is easy to visualize the discrepancy function. Since shear deformation is not ignorable for a thick beam, the Timoshenko beam theory is used as a high-fidelity function to calculate the tip-deflection. The high-fidelity function is expressed as

$$f_H(h) = \frac{4FL^3}{Ebh^3} + \frac{6ML^2}{Ebh^3} + \frac{FL(4+5\upsilon)}{2Ebh} \tag{46}$$

The low-fidelity function is based on the Euler-Bernoulli beam theory, where the shear deformation is ignored, as

$$f_L(h) = \frac{4FL^3}{Ebh^3} + \frac{6ML^2}{Ebh^3} \tag{47}$$



a Cantilever beam under concentrated force and moment

b Tip-deflections of the high- and low-fidelity models

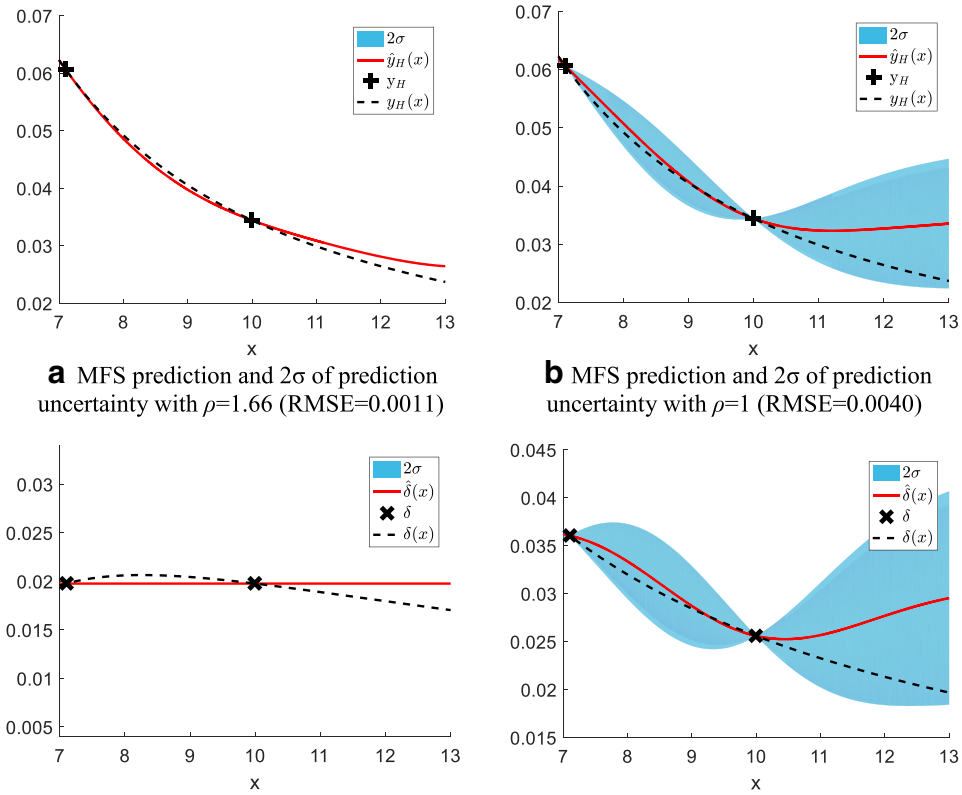
Fig. 10 Cantilever beam example and the tip-deflection of the high- and low-fidelity models with respect to section height

The error in the low-fidelity function comes from the third term of (46). Since the error is linear in the reciprocal of h, it cannot be captured perfectly with a polynomial trend function.

Figure 10(b) compares the high- and low-fidelity functions. Two samples at $x = \{7.1,10\}$ are used from the high-fidelity function and five samples at $x = \{7.1,8.5,10,11.5,13\}$ from the low-fidelity function.

Figure 11(a) and (b) show the predictions using the Bayesian MFS frameworks with and without including ρ (or with $\rho = 1$). The red solid curves are the MFS prediction, while the black dashed curves are the true high-fidelity function. Figure 11(c) and (d) show the corresponding discrepancy predictions and the true discrepancy functions. Note that the true discrepancy functions are different since one is determined by ρ and the other is the difference between the high- and low-fidelity functions. The figures also show the two-sigma prediction uncertainty. In the case of Fig. 11(a) and (c), the prediction uncertainty is too small to show in the figure.

Figure 11(c) shows that ρ is determined such that the true discrepancy function has a small variation, which allows an



a MFS prediction and 2σ of prediction uncertainty with ρ=1.66 (RMSE=0.0011)

b MFS prediction and 2σ of prediction uncertainty with ρ=1 (RMSE=0.0040)

c Discrepancy prediction and 2σ of prediction uncertainty with ρ=1.66 (RMSE=0.0011)

d Discrepancy prediction and 2σ of prediction uncertainty with fixed ρ=1 (RMSE=0.0040)

Fig. 11 The MFS predictions and the discrepancy predictions for cantilever beam example

accurate prediction with two samples. On the other hand, the true discrepancy function without ρ has much larger variation that leads larger bumpiness. The prediction error of the discrepancy function with smaller bumpiness is 0.0011 in terms of RMSE in Fig. 11(c), while its counterpart is 0.0040 in Fig. 11(d), which is almost four times larger.

References

Balabanov V, Haftka RT, Grossman B, Mason WH, Watson LT (1998) Multifidelity response surface model for HSCT wing bending material weight. In *Proceedings of 7th AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization* (pp. 778–788)

Cressie N (2015) *Statistics for spatial data*. John Wiley & Sons

Duchon J (1977) Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables,* W. Schempp and K. Zeller, eds. *Lecture Notes in Mathematics,* No. 571. Springer, Berlin, 85–100

Fernández-Godino MG, Park C, Kim NH, and Haftka RT (2016), Review of Multi-fidelity Models. arXIV preprint arXiv:1609.07196. September, 2016

Forrester AI, Sóbester A, Keane AJ (2007) Multi-fidelity optimization via surrogate modelling. Proceedings of the royal society A: mathematical, physical and engineering science 463(2088):3251–3269

Gano SE, Renaud JE, Martin JD, Simpson TW (2006) Update strategies for kriging models used in variable fidelity optimization. Struct Multidiscip Optim 32(4):287–298

Gutmann HM (2001) A radial basis function method for global optimization. J Glob Optim 19(3):201–227

Jin R, Chen W, Sudjianto A (2005) An efficient algorithm for constructing optimal design of computer experiments. Journal of Statistical Planning and Inference 134(1):268–287

Johnson RA, Wichern DW (2007) Applied multivariate statistical analysis. Sixth edition. Prentice–Hall, Englewood Cliffs, New Jersey

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

Le Gratiet L (2013) Multi-fidelity Gaussian process regression for computer experiments (Doctoral dissertation, Université Paris-Diderot-Paris VII)

Lophaven SN, Nielsen HB, Søndergaard J (2002). *DACE-A Matlab Kriging toolbox, version 2.0*

Mason BH, Haftka RT, Johnson ER, Farley GL (1998) Variable complexity design of composite fuselage frames by response surface techniques. Thin-Walled Struct 32(4):235–261

Matsumura T, Haftka RT, Kim NH (2015) Accurate predictions from noisy data; replication versus exploration with application to structural failure. Struct Multidiscip Optim 51(1):23–40

Morris MD, Mitchell TJ, Ylvisaker D (1993) Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. Technometrics 35(3):243–255

Park C, Haftka RT, Kim NH (2017). Remarks on multi-fidelity surrogates. *Structural and Multidisciplinary Optimization*, 55(3), 1029-1050

Qian PZ, Wu CJ (2008) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics 50(2): 192–204

Rasmussen CE (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning* (pp. 63–71). Springer Berlin Heidelberg

Reddon JR, Jackson DN, Schopflocher D (1985) Distribution of the determinant of the sample correlation matrix: Monte Carlo type one error rates. J Educ Stat 10(4):384–388

Ripley BD (1981) Spatial statistics. Wiley, New York. https://doi.org/10.1002/0471725218

Ben Salem M, Tomaso L (2018) Automatic selection for general surrogate models, Structural and multidisciplinary optimization. Retrieved from https://doi.org/10.1007/s00158-018-1925-3

Xiong S, Qian PZ, Wu CJ (2013) Sequential design and analysis of high accuracy and low-accuracy computer codes. Technometrics 55(1): 37–46

Zhou Q, Wang Y, Choi SK, Jiang P, Shao X, Hu J, Shu L (2017). A robust optimization approach based on multi-fidelity metamodel. Structural and Multidisciplinary Optimization, 57(2), 775-797.