Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Practical Multi-fidelity Heteroscedastic Gaussian Process Modeling Approach for Small-Sample Stochastic Experiments

Zecong Liu
National University of Defense Technology
Jiawei Liu
National University of Defense Technology
Liang Yan
National University of Defense Technology
Binglin Wang
National University of Defense Technology
Xiaojun Duan
xjduan@nudt.edu.cn

National University of Defense Technology

Research Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11799925/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Practical Multi-fidelity Heteroscedastic Gaussian Process Modeling Approach for Small-Sample Stochastic Experiments

Zecong Liu[1†], Jiawei Liu[1†], Liang Yan[1], Binglin Wang[2], Xiaojun Duan[1*]

[1*]College of Science, National University of Defense Technology, Changsha, 410073, Hunan, China.
[2]College of Computer Science and Technology, National University of Defense Technology, Changsha, 410073, Hunan, China.

*Corresponding author(s). E-mail(s): xjduan@nudt.edu.cn;
Contributing authors: liuzecong23@nudt.edu.cn;
liujiawei22a@nudt.edu.cn; yanliang@nudt.edu.cn;
wangbinglin14@nudt.edu.cn;
†Zecong Liu and Jiawei Liu contributed equally to this work and should be considered co-first authors.

Abstract

Multi-fidelity simulation experiments jointly model abundant cheap low-fidelity data with a few expensive but accurate high-fidelity data to approximate engineering response surfaces. Existing multi-fidelity Gaussian process methods, however, handle heterogeneous noise poorly when high-fidelity samples are scarce. This work proposes a multi-fidelity heteroscedastic Gaussian process (MF-hetGP). MF-hetGP retains the autoregressive structure while modeling multi-fidelity heterogeneous noise in a correlated low-to-high affine form to obtain the models' uncertainty quantification practicality. Unlike stochastic co-kriging (SCK), which still requires many high-fidelity samples to build its empirical noise field, MF-hetGP transfers the low-fidelity latent noise field to the high-fidelity level through the affine link, so that accurate prediction and reliable uncertainty quantification are achieved with only a few high-fidelity samples. By assigning Laplace priors to the high-fidelity parameters and employing the Woodbury identity, this work provides a closed-form expression for the likelihood function. Furthermore, to improve the efficiency of hyperparameter estimation, we derive closed-form hyperparameter derivatives and solve them efficiently using

a quasi-Newton algorithm. In addition, the MF-hetGP is extended to the case of multiple fidelity levels, and numerical results verify that prediction accuracy increases with the number of fidelity levels. Several numerical test functions and a radar seeker performance evaluation case are used to verify the advantages of the proposed method. These results demonstrate that MF-hetGP obtains better prediction accuracy than heteroscedastic Gaussian process and SCK.

Keywords: Stochastic simulation experiments, Multi-fidelity surrogate model, Heteroscedastic Gaussian process, Latent-variance inference, Radar seeker performance evaluation

Article Highlights

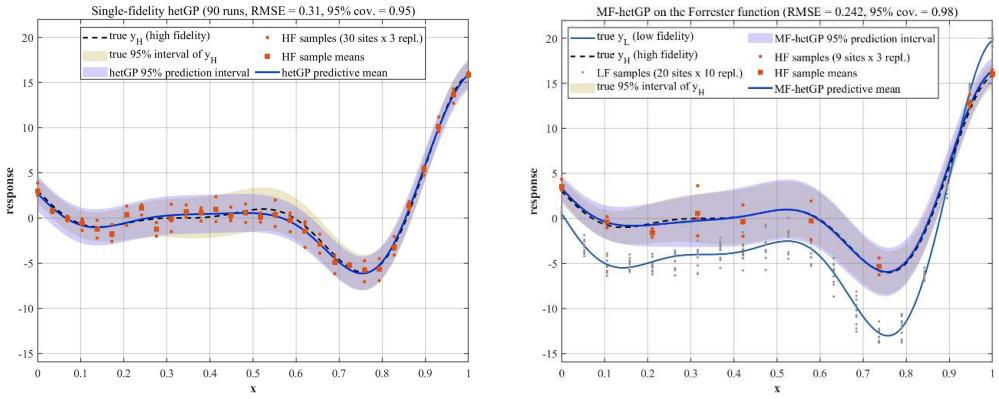
- A practical modeling approach needs only a few high-fidelity experiments to deliver accurate predictions.
- Closed-form likelihood and gradient functions make training fast, stable, and cheap.
- It extends to more than two fidelity levels, with total cost falling as level increasing.
- Several numerical studies show better accuracy and trustworthy uncertainty at less cost.

1 Introduction

Surrogate models, also called metamodels or emulators are a useful tool for studying complex physical systems and expensive computer simulations. Popular surrogate models include random forests, kriging, support vector machines, and neural networks[1–5]. Among them, Gaussian process regression can provide accurate prediction mean and reliable prediction confidence intervals under small sample size, which is suitable for expensive evaluation functions[6–9]. As the objects of study become more complex, however, the intrinsic noise of stochastic systems cannot be ignored. On the one hand, the same input can then produce different outputs. On the other hand, the intrinsic noise variance varies across the design space, that is the so-called heteroscedasticity. Ignoring this heteroscedastic noise, or handling it only with a homoscedastic GP[10, 11], the prediction mean and confidence intervals can be unreliable[12–14].

To address this problem, heteroscedastic Gaussian process regression (HetGP) has been developed, which models the observed data response as the sum of two independent Gaussian processes: a mean-field GP representing the denoised true response function, and a noise-field GP representing the noise function. According to how the noise-field GP is modeled, HetGPs can be categorized into two families: stochastic kriging (SK)[12] and DualGP models[15]. Stochastic kriging is an empirical-process treatment, which requires every design site to be replicated, obtains empirical variances from the replicated observations, and then trains the noise-field GP on these pointwise empirical variances. When the system does not allow replication, or allows only limited replication, the accuracy of the resulting uncertainty quantification deteriorates[13]. The DualGP model was initiated by Goldberg et al.[15]: two independent Gaussian

processes model the true response function and the log noise variance, respectively, where the GP on the log-variance acts in effect as a latent-variable process. Naturally, the estimation accuracy of this latent process governs the predictive accuracy of DualGP. Inference schemes developed so far include MCMC sampling[15, 16], EM acceleration[17, 18], and fully Bayesian approximations[19]. Practical hetGP[13] exploits the Woodbury identity to obtain a closed-form likelihood expression that is maximized rapidly with a quasi-Newton algorithm, and it does not force replication. Beyond Gaussian process models, heterogeneous-noise surrogate models have also been constructed by polynomial regression, where weighted least squares accommodates nonconstant error variance[14, 20]; by Bayesian support vector regression, which estimates the predictive mean and the input-dependent noise variance together[21]; and by heteroscedastic kernel ridge regression[22]; in the neural-network literature, the mean and variance of the target distribution are estimated jointly[23], and modern deep learning quantifies such input-dependent aleatoric uncertainty through heteroscedastic loss functions[16, 24].



(a) Single-fidelity hetGP (b) MF-hetGP

Fig. 1: Single-fidelity versus multi-fidelity heteroscedastic Gaussian process regression on the Forrester two-fidelity benchmark. (a) Single-fidelity hetGP[13] uses 90 high-fidelity sample points (30 locations with 3 replications each), with RMSE 0.31 and 95% prediction confidence-interval coverage 95%; (b) the proposed MF-hetGP uses 200 low-fidelity points (20 locations with 10 replications each) and 27 high-fidelity points (9 locations with 3 replications each). Assuming a high-to-low fidelity cost ratio of about 5:1, the high-fidelity cost is reduced by 70% and the total cost by 25%. At the reduced cost, the multi-fidelity method nevertheless outperforms the single-fidelity method in both curve prediction accuracy and prediction confidence-interval coverage.

HetGPs, however, all rely on a large number of samples, which conflicts with expensive physical simulation experiments. As shown in Figure 1(a), single-fidelity hetGP needs about 100 high-fidelity sample points to model even a one-dimensional heteroscedastic problem well. In the radar seeker performance evaluation problem of

Section 4 such a high-fidelity budget is infeasible. To relax the computational pressure, multi-fidelity(MF) modeling method is employed. MF surrogate modeling methods fuse a large amount of low-fidelity data and a few but accurate high-fidelity data under some model assumptions. For example, Co-kriging fusing multi-fidelity data by auto regression(AR1) assumption[25]. Hierarchical Kriging(H-kriging) assumes the data obeying hierarchical structure aiming to relax correlation matrices calculation complexity[26]. Even the multi-level Gaussian process(MLGP) assumes both high- and low-fidelity data independent[27]. Related multi-fidelity Gaussian processes(MFGPs) include RNA models[28, 29], bayesian hierarchical models[30], TWW models[31] even local transfer GP models[32]. Some tricks for enhancing MFGPs including adaptive sampling[33], active learning[28, 34, 35], experiment designs[36, 37] and fully Bayesian estimations[38]. The first extension of MFGPs to heteroscedastic noise situation is stochastic co-kriging (SCK)[39], however, it still depend on a large number of high-fidelity samples and does not essentially save high-fidelity samples.

This work proposes a practical multi-fidelity heteroscedastic Gaussian process method (MF-hetGP): we retain the autoregressive multi-fidelity structure of SCK, upgrade the intrinsic variances of the two levels to DualGP-style latent-variable joint inference, and derive an explicit closed-form likelihood expression via the Woodbury identity. To save high-fidelity samples, we introduce an affine noise link, that is, the latent noise Gaussian process inferred at the low-fidelity level serves as a prior process to the high-fidelity latent process, so that the corresponding parameters of high-fidelity noise-GP can be easily inferred by a few data. Figure 1(b) shows the modeling effect: compared with practical hetGP[13], the number of high-fidelity sample points is reduced by more than 70% at essentially unchanged predictive accuracy. In addition, we extend MF-hetGP to more than two level settings, where fusing more fidelity levels further reduces the total cost.

The remainder of this paper is organized as follows. Section 2 first reviews the foundations of HetGP modeling (prediction and parameter estimation), then proposes the MF-hetGP model and discusses in detail the setting and estimation of its parameters. Section 3 demonstrates the advantages of the method on several numerical simulation cases, including a detailed discussion of one complete example and the setup of multi-level-fidelity and higher-dimensional cases. Section 4 presents the radar seeker performance evaluation case and its modeling results, meeting the basic requirements of weapon performance assessment. Section 5 gives the conclusions and discussion.

2 The MF-hetGP methodology

2.1 Overview of heteroscedastic Gaussian processes

2.1.1 Gaussian process prediction under heteroscedasticity

Consider a stochastic simulation experiment: N noisy observations are collected in the input space $\mathcal{X} \subseteq \mathbb{R}^d$, with observation vector $\boldsymbol{y}_N$ and input set $\boldsymbol{X}_N$. Heteroscedastic Gaussian process regression models the j-th noisy output as the superposition of a

Gaussian process prior and input-dependent noise,

$$y_j = \boldsymbol{f}(x_j)^\top \boldsymbol{\beta} + M(x_j) + \varepsilon_j, \qquad j = 1, \ldots, N \tag{1}$$

where $\boldsymbol{f}(x_j)$ and $\boldsymbol{\beta}$ are the trend basis and the regression coefficients at the j-th observation. $M \sim \mathrm{GP}\big(0, \tau^2 R(\cdot\,;\theta)\big)$ is the zero-mean stationary Gaussian process prior modeling the true response surface, whose covariance function $\tau^2 R(\cdot\,;\theta)$ measures the correlation of the responses at any two inputs, with field variance τ^2 and correlation length θ, and $\varepsilon_j \sim \mathcal{N}\big(0, V(x_j)\big)$ is the intrinsic noise of the j-th observation, independent of the mean field. GP models assume that the noise variance is independent of the input, i.e., $V(x) \equiv c$, but the intrinsic variance of real engineering data force the response data heteroscedatics. The core of heteroscedastic GP modeling is therefore to make $V(\cdot)$ an explicit component of the model without introducing overly strong structural assumptions.

Given the covariance model and the noise model, prediction under a heteroscedastic GP follows directly from the closure of the Gaussian distribution under conditioning: the training outputs $\boldsymbol{y}$ and the output $Y(x_0)$ at a new input x_0 are jointly Gaussian,

$$\begin{pmatrix} \boldsymbol{y} \\ Y(x_0) \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{F}\boldsymbol{\beta} \\ \boldsymbol{f}(x_0)^\top \boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{c}(x_0) \\ \boldsymbol{c}(x_0)^\top & \tau^2 + V(x_0) \end{pmatrix} \right) \tag{2}$$

where $\mathbf{F} = \big[\boldsymbol{f}(x_j)^\top\big]_{j=1}^{N}$ is the trend matrix, $\boldsymbol{\Sigma}_y = \tau^2 \mathbf{R}(D, D) + \mathrm{diag}\big(V(x_j)\big)$ is the covariance of the N training outputs, composed of the field covariance and the pointwise noise and $\boldsymbol{c}(x_0) = \tau^2 \mathbf{R}(x_0, D)$ is the covariance vector between x_0 and the training inputs. Conditioning Eq. (2) on $\boldsymbol{y}$ yields the posterior predictive distribution $Y(x_0) \mid \boldsymbol{y} \sim \mathcal{N}\big(\hat{Y}(x_0),\, \mathrm{Var}[Y(x_0) \mid \boldsymbol{y}]\big)$, whose mean and variance are

$$\hat{Y}(x_0) = \boldsymbol{f}(x_0)^\top \boldsymbol{\beta} + \boldsymbol{c}(x_0)^\top \boldsymbol{\Sigma}_y^{-1} \left(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta}\right) \tag{3}$$

$$\mathrm{Var}\left[Y(x_0) \mid \boldsymbol{y}\right] = \tau^2 - \boldsymbol{c}(x_0)^\top \boldsymbol{\Sigma}_y^{-1} \boldsymbol{c}(x_0) + V(x_0). \tag{4}$$

In practice $\boldsymbol{\beta}$ is replaced by the generalized least squares (GLS) estimate $\hat{\boldsymbol{\beta}} = \big(\mathbf{F}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{F}\big)^{-1} \mathbf{F}^\top \boldsymbol{\Sigma}_y^{-1} \boldsymbol{y}$, and its estimation error is accounted for in the variance. Equation (4) decomposes the predictive variance explicitly into two components. The first is the *epistemic* component $\tau^2 - \boldsymbol{c}^\top \boldsymbol{\Sigma}_y^{-1} \boldsymbol{c}$, which stems from the unknown response surface itself. The prior uncertainty τ^2 can be estimated by $(\boldsymbol{y}_N - F\hat{\boldsymbol{\beta}}) \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y}_N - F\hat{\boldsymbol{\beta}})/N$ and the epistemic uncertainty is smaller the more similar the new input is to the training points, vanishing as data accumulate. The second is the *aleatoric* component $V(x_0)$, which stems from the intrinsic noise of the data-generating process; it does not depend on the training data, can not decrease as more observations arrive, and must be estimated accurately. The modeled noise response function must therefore vary with the input and must not be confounded with the mean field—which also explains why forcing a homoscedastic assumption onto heteroscedastic data misleads both components simultaneously [14]. SK and DualGP share all prediction equations

(1)–(4) mean that the two families differ only in the quality of $V(\cdot)$, which is analyzed concretely in Section 2.1.2.

2.1.2 Parameter Estimation based on Likelihood Function

SK estimates the noise field in two steps from empirical variances. Since the sampling locations are the "unique-n" set $X_n = \{x_1, \ldots, x_n\}$ with site x_i's replication $a_i (N = \sum a_i)$, denote their response means by $\bar{\boldsymbol{y}}_n = \{\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_n\}$, where

$$\bar{y}_i = \frac{1}{a_i} \sum_{j=1}^{a_i} y_j(x_i). \tag{5}$$

Noting that the data noise affects Eq. (3) only through the correlation matrix, the hyperparameters can be trained directly on $(X_n, \bar{\boldsymbol{y}}_n)$, giving the SK predictive mean curve

$$\hat{y}(x) = \boldsymbol{f}(x)^\top \hat{\boldsymbol{\beta}} + \boldsymbol{c}(x)^\top \boldsymbol{\Sigma}_y^{-1} \left(\bar{\boldsymbol{y}}_n - \mathbf{F} \hat{\boldsymbol{\beta}} \right). \tag{6}$$

The heteroscedastic noise function can then be modeled by empirical variances: the empirical variance at location x_i is estimated directly from the replications as

$$s^2(x_i) = \frac{1}{a_i - 1} \sum_{j=1}^{a_i} \left[y_j(x_i) - \bar{y}(x_i) \right]^2. \tag{7}$$

Using the data $(X_n, \boldsymbol{s}_n)$ with $\boldsymbol{s}_n = [s^2(x_1), ..., s^2(x_n)]$ on the unique-n locations, another Gaussian process can be trained directly to approximate $V(x)$. Using $\hat{V}(x)$ to replace that in Eq. (4) can obtain SK prediction confidence interval. The robustness of kriging under heterogeneous variances in stochastic simulation is studied in [39], and the replication-versus-exploration trade-off in [40].

Note that the likelihood over all "total-N" data is actually

$$\log \mathcal{L}_N = -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \ln \left| \mathbf{C}_N + \boldsymbol{\Lambda}_N \right| - \frac{1}{2} \boldsymbol{y}^\top \left(\mathbf{C}_N + \boldsymbol{\Lambda}_N \right)^{-1} \boldsymbol{y}, \tag{8}$$

but because SK works on the replications of the unique-n locations, its actual likelihood is

$$\log \mathcal{L}_n = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln \left| \boldsymbol{\Upsilon}_n \right| - \frac{1}{2} \bar{\boldsymbol{y}}^\top \boldsymbol{\Upsilon}_n^{-1} \bar{\boldsymbol{y}}, \tag{9}$$

where $\boldsymbol{\Upsilon} = R(X_n, X_n) + \mathrm{diag}(\lambda_i / a_i)$. This likelihood (9) loses information, therefore Practical hetGP[13] uses the Woodbury identity to give the complete likelihood:

$$\log \mathcal{L} = -\frac{N}{2} \ln \left(2\pi \hat{\nu} \right) - \frac{N}{2} - \frac{1}{2} \sum_{i=1}^{n} \left[(a_i - 1) \ln \lambda_i + \ln a_i \right] - \frac{1}{2} \ln \left| \boldsymbol{\Upsilon}_n \right|, \tag{10}$$

where $\hat{\nu} = \bar{\boldsymbol{y}}^\top \boldsymbol{\Upsilon}_n^{-1} \bar{\boldsymbol{y}} / n$ is the concentrated estimate of the field scale τ^2; $\lambda_i = V(x_i)/\tau^2$ is the noise-to-field variance ratio; and $\boldsymbol{\Upsilon}_n = \mathbf{R}(D, D) + \mathrm{diag}\left(\lambda_i / a_i\right)$ is the Woodbury-reduced covariance. The summation term $(a_i - 1) \ln \lambda_i$ together with the within-site

sums of squares enters the likelihood pointwise, so that the information of all observations is retained without loss; hetGP maximizes Eq. (10) over (θ, φ, g) with closed-form gradients and a quasi-Newton algorithm, the latent variables being regularized through their Gaussian prior with a smoothing link. In this article, we suggest the "nugget" g trick to avoid overfitting. Denote $A_n = \mathrm{diag}(a_1, a_2, ..., a_n)$, the noise-GP is

$$\boldsymbol{\Delta}_n \sim \mathcal{N}_n\left(0,\ \nu_g \boldsymbol{\Upsilon}^{(g)}\right), \tag{11}$$

where

$$\boldsymbol{\Upsilon}^{(g)} = \boldsymbol{\Upsilon} + g\mathbf{A}_n^{-1}. \tag{12}$$

2.2 Multi-fidelity heteroscedastic Gaussian process model

2.2.1 Multi-fidelity latent variance-field modeling

Generalizing the single-fidelity model of Section 2.1.1 to two levels, we employ the AR1 structure:

$$y_{Lj}(x_i) = \boldsymbol{f}_L(x_i)^\top \boldsymbol{\beta}_L + M_L(x_i) + \varepsilon_{Lj}(x_i), \tag{13}$$

$$y_{Hj}(x_i) = \rho\, y_L(x_i) + \boldsymbol{f}_H(x_i)^\top \boldsymbol{\beta}_H + M_H(x_i) + \varepsilon_{Hj}(x_i), \tag{14}$$

where $M_L \sim \mathrm{GP}\big(0, \tau_L^2 R_L(\cdot\,; \theta_L^M)\big)$ and $M_H \sim \mathrm{GP}\big(0, \tau_H^2 R_H(\cdot\,; \theta_H^M)\big)$ are mutually independent Gaussian processes; ρ is a constant fidelity scaling factor; and $\varepsilon_{Lj}(x_i) \sim \mathcal{N}\big(0, V_L(x_i)\big)$ and $\varepsilon_{Hj}(x_i) \sim \mathcal{N}\big(0, V_H(x_i)\big)$ are the aleatoric noises of the two levels. A common treatment of the regression basis functions is to set the trend terms of both levels to constants, because the trends of the high- and low-fidelity levels are highly correlated; we assume $\boldsymbol{f}_L \equiv \boldsymbol{f}_H \equiv 1$, so that β_L and β_H are the constant means of the two levels. Noting that when $X_H \subseteq X_L$, the difference statistic follows directly from Eq. (14):

$$\delta_i \equiv y_{H,i} - \rho\, y_{L,i}, \qquad i = 1, \ldots, k_H, \tag{15}$$

whose mean is $\rho\beta_{L,1} + \beta_{H,1} - \rho\beta_{L,1} = \beta_{H,1}$. $\boldsymbol{\delta} = (\delta_1, ..., \delta_{K_H})$ can thus be viewed as observations of a mean GP plus noise. Taking $\delta \sim \mathrm{GP}\big(\beta_D, \tau_H^2 R_H\big)$ with $\beta_D = \beta_{H,1}$, the effective noise of δ_i follows directly from Eqs. (13)–(14):

$$\mathrm{Var}\,(\delta_i) = \frac{V_H(x_{H,i})}{a_{H,i}} + \rho^2 \frac{V_L(x_{H,i})}{a_{L,i}}. \tag{16}$$

Equation (16) is the core mechanism by which MF-hetGP saves samples: the high-fidelity level does not need to infer a complete noise field of its own—the "shape" of its noise field is obtained directly from the low level, and the high-fidelity level only needs to determine a few scaling parameters. This stems from the physical insight that the noise fields across fidelity levels remain strongly correlated.

The low-level variance V_L follows the latent-variable GP structure of Section 2.1.2:

$$V_L(x) = \nu_L \exp\big(\Lambda_L(x)\big), \tag{17}$$

where

$$\log \boldsymbol{\Lambda}_L = \mathbf{C}_{g,L}\left(\mathbf{C}_{g,L} + g_L\, \mathrm{diag}(a_L)^{-1}\right)^{-1} \boldsymbol{\Delta}_L, \tag{18}$$

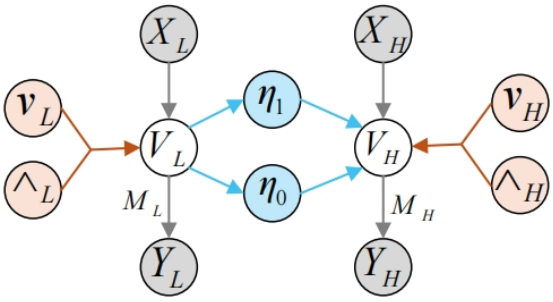


Fig. 2: Schematic illustration of latent-variable multi-fidelity Gaussian process modeling. White circles denote latent variables, gray circles denote observed data, khaki circles represent model parameters, and blue circles indicate affine link parameters.

where $\mathbf{\Delta}_L$ is similar to Eq. (11) and $\mathbf{C}_{g,L} = \mathbf{R}(X_L, X_L; \varphi_L)$; g_L is the smoothness constant. To keep the spatial correlation between V_H and V_L well behaved, an affine noise link is introduced,

$$V_H(x) = \eta_0 + \eta_1 V_L(x), \tag{19}$$

where η_0 and η_1 are two constants to be estimated. Under the setting that the true noise satisfies $V_H = \eta_0^\star + \eta_1^\star V_L$, (η_0, η_1) can be identified from the within-site sums of squares at a small number of high-fidelity sites alone, without resorting to an independent latent-variable GP field. If no simple affine relation exists between V_L and V_H, Eq. (19) can be replaced by an independent latent-variable GP field (the *free* mode in this paper), but this paper mainly discusses the *affine* mode. The free model may overfit the data; when the data are abundant the free MF-hetGP is more accurate, but since this paper studies the small-sample regime, the model is kept simple to guard against overfitting.

Substituting Eq. (19) into Eq. (16) gives the effective noise of the difference statistic,

$$\sigma^2_{H,i}(\rho, \eta_0, \eta_1) = \frac{\eta_0 + \eta_1 V_L(x_{H,i})}{a_{H,i}} + \rho^2 \frac{V_L(x_{H,i})}{a_{L,i}}, \tag{20}$$

and hence the covariance matrix of the difference field,

$$\mathbf{W}_{n,H} = \tau_H^2 \mathbf{C}_H + \mathrm{diag}\left(\sigma^2_{H,1}, \ldots, \sigma^2_{H,k_H}\right), \tag{21}$$

where

$$\mathbf{C}_H = \mathbf{R}(X_H, X_H; \theta_H). \tag{22}$$

Considering that k_H is usually much smaller than k_L, this paper fixes the mean-field length scale of the difference field to $\theta_H \equiv \theta_H^{\mathrm{fix}}$ (a prior-dominated constant), to avoid overfitting the mean field to the small amount of high-fidelity data. Figure 2 concludes the multi-fidelity latent variance modeling of main variables' relationships.

2.2.2 Posterior prediction and interpretability

Stacking the sample means of the two levels into $\boldsymbol{y} = \left[\bar{\boldsymbol{y}}_L^\top, \bar{\boldsymbol{y}}_H^\top\right]^\top \in \mathbb{R}^{k_L+k_H}$ and assembling the joint covariance following the AR1 structure,

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^\top & \boldsymbol{\Sigma}_{22} \end{pmatrix} + \operatorname{diag}\left(\frac{V_L(X_L)}{a_L}, \frac{V_H(X_H)}{a_H}\right), \tag{23}$$

where

$$\begin{aligned} &\boldsymbol{\Sigma}_{11} = \tau_L^2 \mathbf{R}_L(X_L, X_L), \qquad \boldsymbol{\Sigma}_{12} = \rho\tau_L^2 \mathbf{R}_L(X_L, X_H), \\ &\boldsymbol{\Sigma}_{22} = \rho^2\tau_L^2 \mathbf{R}_L(X_H, X_H) + \tau_H^2 \mathbf{R}_H(X_H, X_H). \end{aligned} \tag{24}$$

The regression matrix is

$$\mathbf{F} = \begin{pmatrix} \mathbf{1}_{k_L} & \mathbf{0}_{k_L} \\ \rho\mathbf{1}_{k_H} & \mathbf{1}_{k_H} \end{pmatrix}, \tag{25}$$

and the generalized least squares estimate is

$$\hat{\boldsymbol{\beta}} = \left[\hat{\beta}_1, \hat{\beta}_2\right]^T = \left(\mathbf{F}^T \boldsymbol{\Sigma}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}. \tag{26}$$

For any test point $x_0 \in \Omega$, the predictive mean is

$$\hat{y}(x_0) = \rho\hat{\beta}_1 + \hat{\beta}_2 + \boldsymbol{c}(x_0)^\top \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right), \tag{27}$$

and the predictive model MSE is

$$s^2(x_0) = \underbrace{\rho^2\tau_L^2 + \tau_H^2 - \boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{c}}_{\text{epistemic / model uncertainty}} + \underbrace{\boldsymbol{\eta}^\top \left(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1} \mathbf{F}\right)^{-1} \boldsymbol{\eta}}_{\text{parameter uncertainty}} + \underbrace{V_H(x_0)}_{\text{aleatoric uncertainty}}, \tag{28}$$

where $\boldsymbol{\eta} = [\rho, 1]^\top - \boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \mathbf{F}$, and

$$\boldsymbol{c}(x_0) = \begin{pmatrix} \rho\tau_L^2 \mathbf{R}_L(X_L, x_0) \\ \rho^2\tau_L^2 \mathbf{R}_L(X_H, x_0) + \tau_H^2 \mathbf{R}_H(X_H, x_0) \end{pmatrix} \in \mathbb{R}^{(k_L+k_H)\times 1}. \tag{29}$$

The predictive MSE of this model decomposes into three terms with clear interpretations, corresponding respectively to:

- **Epistemic uncertainty** $\rho^2\tau_L^2 + \tau_H^2 - \boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{c}$: it stems from the finiteness of the design points. When x_0 is far from the design points, the cross kernel $\boldsymbol{c}$ becomes small and this term tends to the prior variance $\rho^2\tau_L^2 + \tau_H^2$; when x_0 coincides with a design point, $\boldsymbol{c}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{c}$ becomes large and the term tends to zero. This is the posterior variance of the mean-field GP, which decreases as N_L, N_H increase.
- **Model uncertainty** $\boldsymbol{\eta}^\top \left(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1} \mathbf{F}\right)^{-1} \boldsymbol{\eta}$: it stems from the estimation error of the regression coefficients $\boldsymbol{\beta}$. Even if x_0 lies exactly on a design point, the error of $\boldsymbol{\beta}$ is still transmitted to the predictive variance through $\boldsymbol{\eta}$; this component shrinks at the rate $O(N_l^{-1})$ as N_l increases.

- **Aleatoric uncertainty** $V_H(x_0)$: the high-fidelity intrinsic noise field given by Eq. (19), representing the inherent randomness of a single high-fidelity observation ε_H; it does not vanish as the sample size increases, and it is the basis on which the prediction band can cover new observation points.

MF-hetGP extends Eqs. (1)–(4) to the multi-fidelity setting, in which ρ measures the correlation between the high- and low-fidelity levels, and V_H measures the degree to which the low-fidelity latent variance function contributes to the high-fidelity level.

2.3 Hyperparameter estimation

There are two classes of hyperparameters to be estimated: (i) for the low-fidelity level—the mean-field, variance-field, and scale parameters θ_L, $\varphi_L, g_L, \mathbf{\Delta}_L$, ν_L; (ii) for the high-fidelity level—the fidelity scaling ρ, the affine link coefficients η_0, η_1, the difference-field variance τ_H^2, and the scale parameter β_D.

This paper adopts a hierarchical two-stage estimation: the latent-variance joint inference is first performed on the low level, and the high-level parameters are then estimated on the difference statistic. This hierarchy is not an approximation: it is dictated by the autoregressive structure of Eqs. (13)–(14)—the low-level noise field V_L is the input of the high-level effective noise (20), so V_L must be estimated first; that is, the low-fidelity likelihood is maximized first, and the high-fidelity likelihood second.

2.3.1 Likelihood functions

The low-fidelity likelihood is exactly the single-fidelity hetGP of Section 2.1.2. Let $\nu_L \equiv \tau_L^2$; let $\mathbf{W}_{n,L}$ and $\mathbf{L}_{n,L}$ denote the Woodbury-reduced covariance matrix and its Cholesky factor; let $\mathbf{L}_{g,L}$ be the Cholesky factor of $\mathbf{W}_{g,L} = \mathbf{C}_{g,L} + g_L \operatorname{diag}(a_L^{-1})$, $\nu_{G,L}$ is the scale of the latent prior, and $\mathbf{\Delta}_L$ the latent variables; the concentrated log-likelihood of the low level is

$$\begin{aligned}\log \mathcal{L}_L\left(\theta_L, \varphi_L, g_L, \mathbf{\Delta}_L\right) = &-\frac{N_L}{2}\log(2\pi) - \frac{N_L}{2}\log\nu_L - \frac{N_L}{2} - \frac{1}{2}\sum_{i=1}^{k_L}(a_{L,i}-1)\log\Lambda_{L,i} \\ &- \sum_{i=1}^{k_L}\log a_{L,i} - \sum_{i=1}^{k_L}\log\left[\mathbf{L}_{n,L}\right]_{ii} - \frac{k_L}{2}\log\left(2\pi\nu_{G,L}\right) \\ &- \frac{\mathbf{\Delta}_L^\top \mathbf{W}_{g,L}^{-1}\mathbf{\Delta}_L}{2\nu_{G,L}} - \sum_{i=1}^{k_L}\log\left[\mathbf{L}_{g,L}\right]_{ii},\end{aligned} \tag{30}$$

where ν_L and $\Lambda_{L,i}$ follow from the concentrated estimates (derived in Appendix A). Equation (30) is optimized over $(\theta_L, \varphi_L, g_L, \mathbf{\Delta}_L)$ to obtain the low-level estimates, from which ν_L and $V_L(\cdot)$ are recovered via Eqs. (17)–(18). To stabilize the optimization, this paper further reparameterizes φ_L through $e^w = \varphi_L/\theta_L$ jointly with θ_L, and imposes the lower bounds $\varphi_L \geq \varphi_{\min} = 0.05$ and $g_L \geq g_{\min} = 10^{-3}$ on the latent variance field.

Let $\boldsymbol{\delta} = [\delta_1 \dots \delta_{k_H}]^\top$; from the difference-field model of Eqs. (20)–(21), the concentrated estimate of the difference-field mean is

$$\hat{\beta}_D = \frac{\mathbf{1}^\top \mathbf{W}_{n,H}^{-1} \boldsymbol{\delta}}{\mathbf{1}^\top \mathbf{W}_{n,H}^{-1} \mathbf{1}}, \qquad \boldsymbol{r} = \boldsymbol{\delta} - \hat{\beta}_D \mathbf{1}. \tag{31}$$

The difference-field covariance and its Cholesky factor are given by Eq. (21). The log-likelihood of the high level is

$$\begin{aligned} \log \mathcal{L}_H \left(\rho, \eta_0, \eta_1, \tau_H^2\right) = &- \frac{k_H}{2} \log(2\pi) - \frac{1}{2} \boldsymbol{r}^\top \mathbf{W}_{n,H}^{-1} \boldsymbol{r} - \sum_{i=1}^{k_H} \log \left[\mathbf{L}_{n,H}\right]_{ii} \\ &- \frac{1}{2} \sum_{i=1}^{k_H} \left[(a_{H,i} - 1) \log V_{H,i} + \frac{SS_{w,H,i}}{V_{H,i}} \right], \end{aligned} \tag{32}$$

where $\mathrm{SS}_{w,H,i} = (a_{H,i} - 1)\, S_{H,i}^2$ is the within-site sum of squares of the high-fidelity replications and $V_{H,i} = \eta_0 + \eta_1 V_L(x_{H,i})$ follows from Eq. (19). Term by term, Eq. (32) reads as follows. The first two terms of the first line, $-\frac{k_H}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{r}^\top \mathbf{W}_{n,H}^{-1}\boldsymbol{r}$, form the k_H-dimensional Gaussian likelihood of the difference field: the quadratic form measures how well the centered difference residuals $\boldsymbol{r}$ are explained by the difference-field covariance $\mathbf{W}_{n,H}$. The third term, $-\sum_{i=1}^{k_H} \log \left[\mathbf{L}_{n,H}\right]_{ii}$, is the log-determinant (complexity) penalty of that covariance, which regularizes the fitted scale of the difference field. The second line is the within-site term contributed by the high-fidelity replications: for each site it couples $(a_{H,i} - 1)\log V_{H,i}$ with the ratio $\mathrm{SS}_{w,H,i}/V_{H,i}$, i.e., the log-density of a chi-square statistic with $a_{H,i} - 1$ degrees of freedom. This within-site term supplies information about $V_{H,i}$ that is independent of ρ whenever $a_{H,i} \geq 2$, pinning the fitted noise level to the observed within-site scatter; when $a_{H,i} = 1$ it vanishes, and V_H is then determined entirely by the inherited low-level field V_L through the affine link (19).

When k_H is small, Eq. (32) may exhibit multimodality or boundary behavior with respect to ρ [13], a problem common in maximum likelihood estimation of Gaussian processes. To this end, this paper first obtains an initial value by least squares,

$$\hat{\rho}_0 = \mathrm{clip}\left(\frac{\bar{\boldsymbol{y}}_L(X_H)^\top \bar{\boldsymbol{y}}_H}{\bar{\boldsymbol{y}}_L(X_H)^\top \bar{\boldsymbol{y}}_L(X_H)} \right), \tag{33}$$

where $\mathrm{clip}(\cdot)$ restricts the slope to a bounded interval; then, on a grid of candidate values built around $\hat{\rho}_0$, it optimizes $\left(\eta_0, \eta_1, \tau_H^2\right)$ for each candidate ρ and takes the ρ with the largest likelihood as the estimate,

$$\hat{\rho} = \arg\max_{\rho} \max_{\eta_0, \eta_1, \tau_H^2} \log \mathcal{L}_H \left(\rho, \eta_0, \eta_1, \tau_H^2\right), \tag{34}$$

followed by a local refinement in the neighborhood of the grid winner.

2.3.2 Gradient-assisted solution

Both the low and the high level use a quasi-Newton method with analytic gradients to accelerate the solution. The low-level gradients have been given by Eqs. (16)–(22) of hetGP [13]; for completeness, the core chain relations are given here. Differentiating Eq. (30) with respect to the latent variance field $\Lambda_{L,i}$:

$$\frac{\partial \log \mathcal{L}_L}{\partial \Lambda_{L,i}} = \frac{a_{L,i} s^2_{b,L,i}}{2\nu_L \Lambda^2_{L,i}} + \frac{p^2_{L,i}}{2\nu_L a_{L,i}} - \frac{a_{L,i}-1}{2\Lambda_{L,i}} - \frac{\left[\mathbf{W}^{-1}_{n,L}\right]_{ii}}{2a_{L,i}}, \tag{35}$$

where $\boldsymbol{p}_L = \mathbf{W}^{-1}_{n,L}\bar{\boldsymbol{y}}_L$. The soft-threshold chain rule $\partial\boldsymbol{\Lambda}/\partial\boldsymbol{z} = \mathrm{soft}'(\boldsymbol{z})\boldsymbol{\Lambda}$ then propagates Eq. (35) back to the latent variables $\boldsymbol{z}$; finally, for the length scales θ_L, φ_L, the regularization term g_L, and the latent variables $\boldsymbol{\Delta}_L$, the kernel derivatives $\partial\mathbf{C}/\partial\theta_h = -\mathbf{C}\odot\mathbf{D}^2_h$ ($\mathbf{D}_h$ being the pairwise distance matrix of the h-th dimension) are back-propagated, giving the gradient

$$\begin{aligned}\frac{\partial \log \mathcal{L}_L}{\partial \boldsymbol{u}_L} = \big[&\theta_L \nabla_{\theta_L} \log \mathcal{L}_L;\ (\varphi_L - \varphi_{\min})\nabla_{\varphi_L}\log\mathcal{L}_L;\\ &(g_L - g_{\min})\,\partial_{g_L}\log\mathcal{L}_L;\ \nabla_{\boldsymbol{\Delta}_L}\log\mathcal{L}_L\big],\end{aligned} \tag{36}$$

where $\boldsymbol{u}_L = [\log\theta_L,\ w,\ \log(g_L - g_{\min}),\ \boldsymbol{\Delta}_L]$.

For the high level, let $\boldsymbol{p}_H = \mathbf{W}^{-1}_{n,H}\boldsymbol{r}$; the gradients with respect to the effective noise variance $\sigma^2_{H,i}$ and τ^2_H are

$$\frac{\partial \log \mathcal{L}_H}{\partial \sigma^2_{H,i}} = \frac{p^2_{H,i}}{2} - \frac{\left[\mathbf{W}^{-1}_{n,H}\right]_{ii}}{2}, \qquad \frac{\partial \log \mathcal{L}_H}{\partial \tau^2_H} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{W}^{-1}_{n,H}\mathbf{C}_H\right) + \frac{1}{2}\boldsymbol{p}_H^\top \mathbf{C}_H \boldsymbol{p}_H, \tag{37}$$

and, since by Eqs. (20)–(21) $V_{H,i} = \eta_0 + \eta_1 V_{L,i}$ and $\sigma^2_{H,i} = V_{H,i}/a_{H,i} + \rho^2 V_{L,i}/a_{L,i}$, the chain rule propagates them to (η_0, η_1):

$$\frac{\partial \log \mathcal{L}_H}{\partial \eta_0} = \sum_{i=1}^{k_H} \frac{1}{a_{H,i}} \frac{\partial \log \mathcal{L}_H}{\partial \sigma^2_{H,i}}, \qquad \frac{\partial \log \mathcal{L}_H}{\partial \log \eta_1} = \eta_1 \sum_{i=1}^{k_H} \frac{V_{L,i}}{a_{H,i}} \frac{\partial \log \mathcal{L}_H}{\partial \sigma^2_{H,i}}. \tag{38}$$

Combining these with the direct derivative of the within-site term in Eq. (32), $\partial(\cdot)/\partial V_{H,i} = -(a_{H,i}-1)/(2V_{H,i}) + SS_{w,H,i}/\left(2V^2_{H,i}\right)$, yields the complete gradient with respect to $(\eta_0, \log\eta_1)$. To handle potentially non-positive-definite covariance matrices and ill-conditioned situations, the optimization uses a multi-start strategy.

Since ρ is obtained by the grid profile of Eq. (34), ρ is fixed in the corresponding optimization, so the gradient only needs to be computed with respect to $\left(\eta_0, \eta_1, \tau^2_H\right)$. The optimization of τ^2_H is parameterized as $u_3 = \log\left(\tau^2_H - \tau_{\min}\right)$ with lower bound $\tau_{\min} = 0.5$, to prevent the likelihood from collapsing at the boundary when k_H is small (i.e., the difference field degenerating into structureless white noise), thereby preserving the residual structure under mismatch.

2.3.3 The Laplace prior setting

When $a_{H,i}$ is small, the within-site estimate $S_{H,i}^2$ of $V_{H,i}$ has only $a_{H,i} - 1$ degrees of freedom and its sampling noise is large. If Eq. (32) alone were used to estimate (η_0, η_1), the estimator would chase the noise and overfit. A Laplacian shrinkage prior is therefore introduced for the affine link, anchoring (η_0, η_1) near the "noise inheritance point" (i.e., $V_H = V_L$, $\eta_0 = 0, \eta_1 = 1$):

$$p(\eta_0, \eta_1) \propto \exp\left(-\frac{|\eta_0|}{s_0} - \frac{|\eta_1 - 1|}{s_1}\right), \tag{39}$$

with prior scales

$$s_0 = \max\left(\text{median}\left(V_L(X_H)\right),\ 10^{-3}\right), \qquad s_1 = 2. \tag{40}$$

Taking s_0 as the median of $V_L(X_H)$ makes the prior meaningful on the "noise scale". $s_1 = 2$ is a relatively wide dimensionless scale, which shrinks the prior near $\eta_1 = 1$ without pinning η_1 to 1—the latter would flatten the likelihood of ρ and make it unidentifiable.

Because the Laplace distribution is non-differentiable at $\eta_0 = 0, \eta_1 = 1$, to cooperate with gradient-type optimizers the absolute values are smoothed:

$$p(\eta_0, \eta_1) \propto \exp\left(-\frac{\sqrt{\eta_0^2 + \epsilon^2}}{s_0} - \frac{\sqrt{(\eta_1 - 1)^2 + \epsilon^2}}{s_1}\right). \tag{41}$$

When ϵ^2 is taken too small, the curvature at $\eta_1 = 1$ is too large and stalls the optimization; this paper recommends $\epsilon^2 = 10^{-4}$. The high-level maximum a posteriori (MAP) estimate with the prior is

$$\left(\hat{\eta}_0, \hat{\eta}_1, \hat{\tau}_H^2\right) = \arg\max_{\eta_0, \eta_1, \tau_H^2} \log \mathcal{L}_H\left(\rho, \eta_0, \eta_1, \tau_H^2\right) + \log p(\eta_0, \eta_1). \tag{42}$$

Correspondingly, when differentiating the objective (32), the gradient terms of the prior must be supplemented,

$$\frac{\partial \log p}{\partial \eta_0} = -\frac{\eta_0}{s_0\sqrt{\eta_0^2 + \epsilon^2}}, \qquad \frac{\partial \log p}{\partial \eta_1} = -\frac{\eta_1 - 1}{s_1\sqrt{(\eta_1 - 1)^2 + \epsilon^2}}. \tag{43}$$

Merging Eq. (43) with Eqs. (37)–(38) gives the complete high-level gradient. The overall workflow is summarized in Algorithm 1.

Algorithm 1 MF-hetGP ($M = 2$): training and prediction

Input: low-level data $D_L = \left(X_L, \bar{\boldsymbol{y}}_L, S_L^2, a_L\right)$ and high-level data $D_H = \left(X_H, \bar{\boldsymbol{y}}_H, S_H^2, a_H\right)$ with $X_H \subseteq X_L$; noise link $\in \{\text{affine},\ \text{free}\}$; test set X_{test}.
Output: fitted model; $\left(\hat{y}(x_0),\ s^2(x_0)\right)$ for every $x_0 \in X_{\text{test}}$.
1: Normalize the inputs as $x \mapsto (x - x_{\min})/\text{span}$ (shared by both levels); compute SS_L and $s^2_{b,L}$ from $\left(\bar{\boldsymbol{y}}_L, S_L^2, a_L\right)$, warm-started by a homoscedastic GP fit.
2: *Level 1 (low-fidelity hetGP):* $\boldsymbol{u}_L^* \leftarrow \arg\min_{\boldsymbol{u}} \text{hetObjL}(\boldsymbol{u})$ by multi-start quasi-Newton on (30) with the gradient (35).
3: Recover $\hat{\theta}_L$, $\hat{\varphi}_L$, $\hat{g}_L$, $\hat{\boldsymbol{\Delta}}_L$ from $\boldsymbol{u}_L^*$ through the parameterization (36); $\hat{\nu}_L \leftarrow$ concentrated estimate of (30); $\hat{\tau}_L^2 \leftarrow \hat{\nu}_L$.
4: Compute $\hat{V}_L(x)$ from (17)–(18), and set $\hat{V}_{L,H} \leftarrow \hat{V}_L(X_H)$.
5: *Level 2 (profile ρ on $\boldsymbol{\delta} = \bar{\boldsymbol{y}}_H - \rho \cdot \bar{\boldsymbol{y}}_L(X_H)$, cf. (15)):* $\rho_0 \leftarrow$ least-squares slope (33); **for** each candidate ρ on the profile grid of (34) **do**: obtain $\left(\hat{\eta}_0(\rho), \hat{\eta}_1(\rho), \hat{\tau}_H^2(\rho)\right)$ by Quasi-Newton method maximization of (32) with the Laplace prior (42).
6: $\left(\hat{\rho}, \hat{\eta}_0, \hat{\eta}_1, \hat{\tau}_H^2\right) \leftarrow \arg\max_\rho \log \mathcal{L}_H(\rho)$ as in (34), with a local refinement around the winner.
7: *Assembly and prediction:* assemble $\boldsymbol{\Sigma}$ via (23)–(24) and $\mathbf{F}$ via (25); $\hat{\boldsymbol{\beta}} \leftarrow \left(\mathbf{F}^T\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{y}}$ (26).
8: **for** $x_0 \in X_{\text{test}}$ **do**: compute $\boldsymbol{c}(x_0)$ from (29); $\hat{y}(x_0) \leftarrow \rho\hat{\beta}_1 + \hat{\beta}_2 + \boldsymbol{c}(x_0)^T\boldsymbol{\Sigma}^{-1}\left(\bar{\boldsymbol{y}} - \mathbf{F}\hat{\boldsymbol{\beta}}\right)$ (27); $\hat{V}_H(x_0) \leftarrow \hat{\eta}_0 + \hat{\eta}_1\hat{V}_L(x_0)$ (19); $\text{mse}(x_0)$ from (28).
9: **return** the model and $\left(\hat{y}, s^2\right)$.

2.4 Generalization to M fidelity levels

This section generalizes the two-level model of Sections 2.2–2.3 uniformly to M fidelity levels, following the co-kriging formulation.

2.4.1 Recursive data model and difference statistics

Generalizing Eqs. (13)–(14) to a recursive autoregressive structure:

$$\begin{aligned} y_{1,j}(x_i) &= \beta_1 + M_1(x_i) + \varepsilon_{1,j}(x_i), \\ y_{l,j}(x_i) &= \rho_l\, y_{l-1}(x_i) + \beta_l + M_l(x_i) + \varepsilon_{l,j}(x_i), \quad l \geq 2, \end{aligned} \tag{44}$$

where the levels are mutually independent, $M_l \sim \text{GP}\left(0, \tau_l^2 R_l(\cdot;\theta_l)\right)$, $\varepsilon_{l,j}(x_i) \sim \mathcal{N}\left(0, V_l(x_i)\right)$, $\rho_1 \equiv 1$, and the basis functions are constants. Recursive expansion gives

$$y_l(x) = \sum_{k=1}^{l} \alpha_{l,k}\left(\beta_k + M_k(x) + \varepsilon_k(x)\right), \qquad \alpha_{l,k} \equiv \prod_{j=k+1}^{l} \rho_j, \tag{45}$$

where empty products are defined as 1. Applied to the sample means used in the sequel, $\varepsilon_k(x)$ in Eq. (45) is understood as the averaged noise $\bar{\zeta}_k(x)$. For $l \geq 2$, the

nested design yields the difference statistic

$$\delta_{l,i} \equiv y_{l,i} - \rho_l \, y_{l-1}(x_{l,i}), \tag{46}$$

whose variance is

$$\mathrm{Var}\left(\delta_{l,i}\right) = \frac{V_l(x_{l,i})}{a_{l,i}} + \rho_l^2 \frac{V_{l-1}(x_{l,i})}{a_{l-1}(x_{l,i})}. \tag{47}$$

2.4.2 Affine noise chain

The noise of the lowest fidelity level V_1 follows the DualGP of Section 2.1.2:

$$V_1(x) = \nu_1 \exp\left(\mathbf{C}_{g,1}(x, X_1)\mathbf{W}_{g,1}^{-1}\boldsymbol{\Delta}_1\right). \tag{48}$$

Levels $2, \ldots, M$ inherit bottom-up through the affine chain:

$$V_l(x) = \eta_{l,0} + \eta_{l,1} V_{l-1}(x), \tag{49}$$

so that

$$V_M(x) = \eta_{M,0}^{\text{eff}} + \eta_{M,1}^{\text{eff}} V_1(x), \qquad \eta_{M,0}^{\text{eff}} = \sum_{k=2}^{M} \eta_{k,0} \prod_{j=k+1}^{M} \eta_{j,1}, \qquad \eta_{M,1}^{\text{eff}} = \prod_{j=2}^{M} \eta_{j,1}. \tag{50}$$

That is, no matter how large M is, the noise function of level M is characterized by V_1 plus $2M-1$ scaling constants: the number of parameters grows linearly with M rather than with $\sum_l k_l$—the core structural advantage of multi-level MF-hetGP, which saves computational cost.

2.4.3 Joint covariance and regression matrix

From Eqs. (44)–(45), the cross-covariance of levels $l \le m$ is

$$\boldsymbol{\Sigma}_{l,m} = \sum_{k=1}^{l} \alpha_{l,k}\alpha_{m,k}\tau_k^2 \mathbf{R}_k(X_l, X_m), \qquad l, m = 1, \ldots, M, \tag{51}$$

and the joint covariance and regression matrix are assembled blockwise:

$$\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{l,m}]_{l,m=1}^{M} + \mathrm{diag}\left(\frac{V_1(X_1)}{a_1}, \ldots, \frac{V_M(X_M)}{a_M}\right), \tag{52}$$

$$\mathbf{F}_{l,k} = \alpha_{l,k}, \; k \le l; \qquad 0, \; k > l. \tag{53}$$

Generalized least squares gives $\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top \boldsymbol{\Sigma}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{y}$, where $\boldsymbol{y} = \left[\boldsymbol{y}_1^\top \ldots \boldsymbol{y}_M^\top\right]^\top$.

2.4.4 Prediction and uncertainty decomposition

The cross-covariance of level M is given blockwise by $\boldsymbol{c}_l(x_0) = \sum_{k=1}^{l} \alpha_{l,k}\alpha_{M,k}\tau_k^2 \mathbf{R}_k(X_l, x_0)$, stacked into $\boldsymbol{c}(x_0)$; the prior mean is $\boldsymbol{f}_0 = [\alpha_{M,1} \dots \alpha_{M,M}]$. The predictive mean is

$$\hat{y}_M(x_0) = \boldsymbol{f}_0\hat{\boldsymbol{\beta}} + \boldsymbol{c}(x_0)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right), \tag{54}$$

and the total predictive variance for a new observation $y_M(x_0)$ decomposes as

$$s^2(x_0) = \underbrace{\sum_{k=1}^{M} \alpha_{M,k}^2\tau_k^2 - \boldsymbol{c}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{c}}_{\text{epistemic}} + \underbrace{\boldsymbol{\eta}^\top\left(\mathbf{F}^\top\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)^{-1}\boldsymbol{\eta}}_{\text{parameter / model}} + \underbrace{V_M(x_0)}_{\text{aleatoric}}, \tag{55}$$

The three terms have the same meaning as in Eq. (28): the epistemic term decreases as $\sum_l N_l$ increases and tends to the prior variance far from the design points; the parameter term shrinks at $O(N_l^{-1})$; the aleatoric term given by Eqs. (48)–(50) does not vanish with the sample size and is the basis on which the prediction band covers new observations. The $1-\alpha$ prediction band is $\hat{y}_M \pm z_{1-\alpha/2}\sqrt{\text{mse}}$.

2.4.5 Hierarchical estimation

Level $l \geq 2$ only needs to fit its own $\left(\rho_l, \eta_{l,0}, \eta_{l,1}, \tau_l^2\right)$ on the difference statistic. The log-likelihood is the per-level generalization of Eq. (32):

$$\begin{aligned}\log\mathcal{L}_l = &-\frac{k_l}{2}\log(2\pi) - \frac{1}{2}\boldsymbol{r}_l^\top\mathbf{W}_{n,l}^{-1}\boldsymbol{r}_l - \sum_i \log\left[\mathbf{L}_{n,l}\right]_{ii} \\ &- \frac{1}{2}\sum_i\left[(a_{l,i}-1)\log V_{l,i} + \frac{SS_{w,l,i}}{V_{l,i}}\right],\end{aligned} \tag{56}$$

where $\boldsymbol{r}_l = \boldsymbol{\delta}_l - \beta_{D,l}\mathbf{1}$, $\beta_{D,l} = \left(\mathbf{1}^\top\mathbf{W}_{n,l}^{-1}\boldsymbol{\delta}_l\right) / \left(\mathbf{1}^\top\mathbf{W}_{n,l}^{-1}\mathbf{1}\right)$, $V_{l,i} = \eta_{l,0} + \eta_{l,1}V_{l-1}(x_{l,i})$, and

$$\mathbf{W}_{n,l} = \tau_l^2\mathbf{C}_l + \text{diag}\left(\frac{V_{l,i}}{a_{l,i}} + \rho_l^2\frac{V_{l-1}(x_{l,i})}{a_{l-1}(x_{l,i})}\right). \tag{57}$$

ρ_l is profiled as in Eqs. (33)–(34), and the Laplacian inheritance prior follows Eq. (39):

$$p\left(\eta_{l,0}, \eta_{l,1}\right) \propto \exp\left(-\frac{\sqrt{\eta_{l,0}^2 + \epsilon^2}}{s_{l,0}} - \frac{\sqrt{\left(\eta_{l,1}-1\right)^2 + \epsilon^2}}{s_{l,1}}\right). \tag{58}$$

The training and prediction workflow of the M-level model is summarized in Algorithm 2.

Algorithm 2 MF-hetGP (M levels): training and prediction

Input: per-level data $\left(X_l, \bar{\boldsymbol{y}}_l, S_l^2, a_l\right)$, $l = 1, \ldots, M$, with nested designs $X_1 \supseteq X_2 \supseteq \cdots \supseteq X_M$; prior scales $s_{l,0}$, $s_{l,1}$; test set X_{test}.

Output: fitted model; $\left(\hat{y}_M(x_0),\, s^2(x_0)\right)$ for every $x_0 \in X_{\text{test}}$.

1: *Level 1:* run Steps 1–4 of Algorithm 1 to obtain $\hat{V}_1(x)$ from (48).

2: **for** $l = 2, \ldots, M$ **do**:

3: Construct the difference statistic $\boldsymbol{\delta}_l = \bar{\boldsymbol{y}}_l - \rho_l\, \bar{\boldsymbol{y}}_{l-1}(X_l)$ (46) and evaluate the inherited field $\hat{V}_{l-1}(X_l)$.

4: **for** each candidate ρ_l on the profile grid of (34) **do**: maximize (56) with the Laplace prior (58) over $\left(\eta_{l,0}, \eta_{l,1}, \tau_l^2\right)$ by Quasi-Newton method.

5: Keep $\left(\hat{\rho}_l, \hat{\eta}_{l,0}, \hat{\eta}_{l,1}, \hat{\tau}_l^2\right)$ with the largest likelihood, and locally refine the winner; $\hat{V}_l(x) \leftarrow \hat{\eta}_{l,0} + \hat{\eta}_{l,1}\hat{V}_{l-1}(x)$ (49).

6: *Assembly and prediction:* assemble $\boldsymbol{\Sigma}$ via (51)–(52) and $\mathbf{F}$ via (53); $\hat{\boldsymbol{\beta}} \leftarrow \left(\mathbf{F}^T\boldsymbol{\Sigma}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}$.

7: **for** $x_0 \in X_{\text{test}}$ **do**: compute the blocks $\boldsymbol{c}_l(x_0)$ of (51) and stack them into $\boldsymbol{c}(x_0)$; $\hat{y}_M(x_0)$ from (54); $\text{mse}(x_0)$ from (55).

8: **return** the model and $\left(\hat{y}_M, s^2\right)$.

3 Numerical validation

3.1 An illustration for MF-hetGP

3.1.1 Forrester function and basic modeling results

MF-hetGP is demonstrated and tested on a one-dimensional two-fidelity benchmark with heteroscedastic noise. The high-fidelity surface is the Forrester function [41], the most classical test case of the multi-fidelity literature:

$$y_H(x) = (6x - 2)^2 \sin(12x - 4), \qquad x \in [0, 1], \tag{59}$$

and the low-fidelity surface is its scale-and-bias approximation,

$$y_L(x) = 1.5\, y_H(x) - 4, \tag{60}$$

so that the autoregressive model is exactly well specified: the truth is $\rho^* = 1/1.5 = 0.667$ and $\delta^*(x) = 4/1.5 = 2.667$; the closeness of $\hat{\rho}$ to 0.667 is therefore the most direct indicator of the fit quality. The intrinsic variances of both levels are input-dependent functions, and these two functions are correlated, consistent with the physical mechanism:

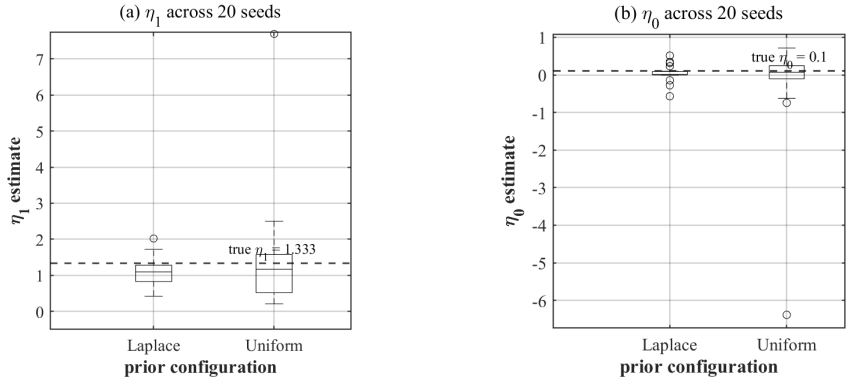
$$V_H(x) = 0.5 + 2.5 \sin^2(\pi x), \qquad V_L(x) = 0.5 + 2.0 \sin^2(\pi x). \tag{61}$$

The noise is strongest at $x = 0.5$ and weakest at the two ends (0.5). The simulation outputs are generated as $y_{Lj}(x_i) = y_L(x_i) + \varepsilon_{Lj}$ and $y_{Hj}(x_i) = y_H(x_i) + \varepsilon_{Hj}$ with $\varepsilon_{Lj} \sim \mathcal{N}(0, V_L(x_i))$ and $\varepsilon_{Hj} \sim \mathcal{N}(0, V_H(x_i))$.

Figure 1 of Section 1 has already shown the basic modeling effect of this function, in comparison with single-fidelity modeling. Figure 1(a) is the single high-fidelity modeling result: SF-hetGP uses 90 high-fidelity sample points, with RMSE 0.31 and 95% prediction confidence-interval coverage 95%. Considering that the low-fidelity level is cheap—its cost ratio to the high-fidelity level is 1:5—the proposed MF-hetGP uses 200 low-fidelity points at 20 locations with 10 replications each, to estimate the low-fidelity heteroscedastic noise function more accurately for transfer to the high-fidelity noise function, plus 27 high-fidelity points. The joint modeling attains predictive RMSE 0.242 and confidence-interval coverage 98%. At a reduced total cost, the multi-fidelity method outperforms the single-fidelity method in both curve prediction accuracy and prediction confidence-interval coverage. To further illustrate the characteristics and potential of the method, the next subsection gives the setup explanation of the Laplace prior and the prediction accuracy tests of the modeling.

3.1.2 Ablation study of the Laplace prior

Since the high-fidelity level has only $k_H = 9$ locations with $a_H = 3$ replications each, the information on the two scalars of the affine link $V_H = \eta_0 + \eta_1 V_L$ is weak. An ablation study of the prior information is therefore conducted, comparing it with an uninformative prior. To demonstrate the reliability of the results, the experiments are repeated over 20 random seeds with different initial designs, and the parameter estimation results are shown in Figure 3. The variance fitting curves are shown in Figure 4.



(a) Laplace informative prior (b) Uninformative prior

Fig. 3: Parameter estimation results of the affine link coefficients over 20 random seeds under the two priors. The black dashed lines mark the true values.

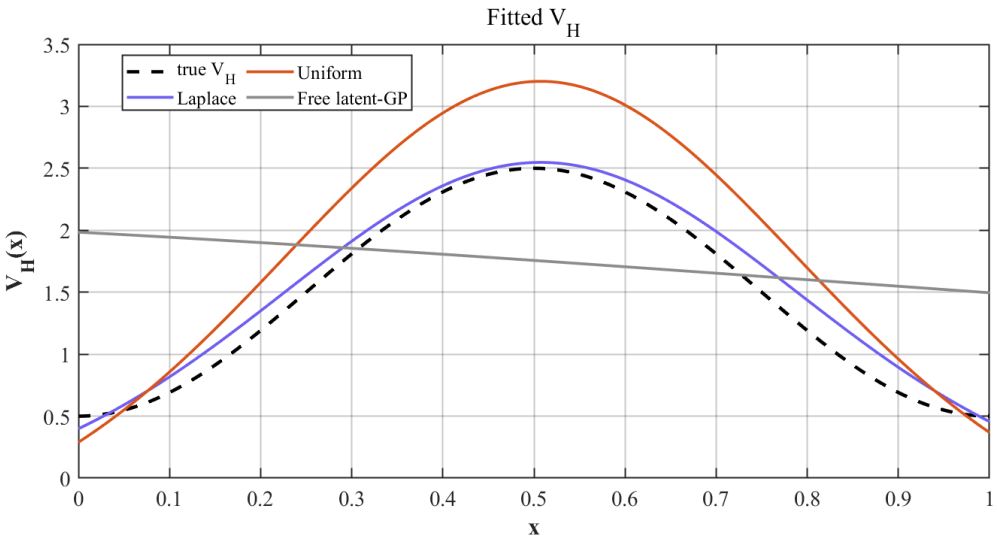


Fig. 4: Fitted high-fidelity variance functions of the three prior configurations versus the true field (black dashed line).

Figure 3 compares, over 20 random seeds, the box plots of the affine link coefficient estimates η_1 (left panel) and η_0 (right panel) under the Laplace prior and the uninformative (uniform) prior, with the black dashed lines marking the true values. For η_1 (true value 1.333), the Laplace-prior estimates are highly concentrated and close to the true value with only tiny fluctuation, whereas the uninformative-prior estimates have clearly larger variance: not only does the median deviate more, but an extreme outlier as high as about 7.5 appears. For η_0, the Laplace-prior estimates cluster tightly and stably near 0, whereas the uninformative prior produces an extreme negative outlier below -6, which is completely unacceptable in the physical sense of a variance intercept. The figure shows directly that, with very few high-fidelity samples, the Laplace prior suppresses abnormal fluctuations of the parameters, guaranteeing the robustness and interpretability of the noise model across different random seeds.

Figure 4 shows the fitting effect of three different prior configurations on the high-fidelity variance field $V_H(x)$. The black dashed line is the true variance field (a bell-shaped distribution, high in the middle and low at both ends, with a peak of about 2.5). The Laplace prior (blue line) fits the true curve almost perfectly, accurately capturing the shape and the peak of the heteroscedasticity; the uninformative prior (orange line) recovers the overall shape but, lacking regularization, clearly overestimates the variance level (peak above 3); the free latent-GP (gray line) fails to capture the heteroscedastic feature at all, collapsing into a nearly flat line. This comparison directly proves that, with very few high-fidelity samples, the Laplace inheritance prior effectively prevents parameter drift and overfitting—it is the only configuration that accurately reconstructs the complex variance field—whereas the lack of prior constraints or fully independent modeling leads to systematic bias or severely degenerate solutions.

3.2 Multi-fidelity case: cost analysis

3.2.1 Setup: the five-fidelity six-hump camel case

This section evaluates the cost–accuracy trade-off of the multi-level extension on a two-dimensional five-fidelity benchmark whose top level is the six-hump camel back function, a classical multimodal two-dimensional test function with six local minima and two global minima[42]. The five truth levels share one common surface shape through affine rescalings:

$$
\begin{aligned}
&y_5(x) = \left(4 - 2.1x_1^2 + \tfrac{x_1^4}{3}\right) x_1^2 + x_1x_2 + 4x_2^2\left(1 - x_2^2\right), \qquad x_1 \in [-3,3],\ x_2 \in [-2,2],\\
&y_4(x) = 1.5y_5(x) + 4 + 0.2g(x), \qquad y_3(x) = 2.0y_5(x) + 8 + 0.4g(x),\\
&y_2(x) = 2.5y_5(x) + 12 + 0.8g(x), \qquad y_1(x) = 3.0y_5(x) + 16 + 1.6g(x),\\
&g(x) = \exp\left(-1.4\tfrac{x_1+2}{4}\cos\left(3.5\pi\tfrac{x_1+1}{2}\right)\right).
\end{aligned}
\tag{62}
$$

The intrinsic noises follow the affine chain of Eq. (49) on the normalized inputs $u \in [0,1]^2$:

$$
\begin{aligned}
V_1(u) &= 4\left(0.02 + 0.5\left(\sin^2(\pi u_1)\sin^2(\pi u_2) + 0.3u_1u_2\right)\right),\\
V_l(u) &= 4b_l + \eta_{l,1}V_{l-1}(u), \quad l = 2,\dots,5,
\end{aligned}
\tag{63}
$$

with $b = (0.03, 0.05, 0.08, 0.10)$ and true link slopes $\eta^*_{l,1} = (1.2, 1.3, 1.5, 1.8)$—the most expensive levels are also the noisiest, which is exactly what makes the cheap levels valuable. Figure 5 shows the six-hump response surface and the top-level noise variance field of this case.

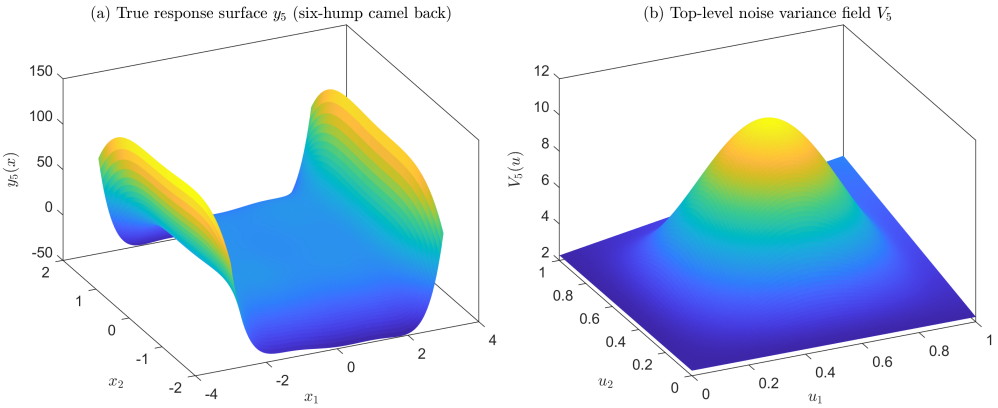


Fig. 5: The five-fidelity six-hump camel case of Section 3.2.1: (a) the true top-level response surface y_5 (six-hump camel back) and (b) the true top-level heteroscedastic noise variance field V_5.

Each evaluation of fidelity level l costs c_l, with per-run cost ratio $c_1 : c_2 : c_3 : c_4 : c_5 = 1 : 4 : 16 : 64 : 256$. Five nested designs are compared in Table 1. All designs are nested so that the low-level trend information transfers upward.

3.2.2 Results

Table 1: Sample sizes of the experimental designs of each fidelity level.

Scheme	A	B	C	D	E
F5	50×2	8×3	8×3	8×3	8×3
F4	0	64×3	24×3	16×3	16×3
F3	0	0	48×4	32×4	32×4
F2	0	0	0	96×5	64×4
F1	0	0	0	0	96×5
Total cost	100	72	54	51.5	49.9

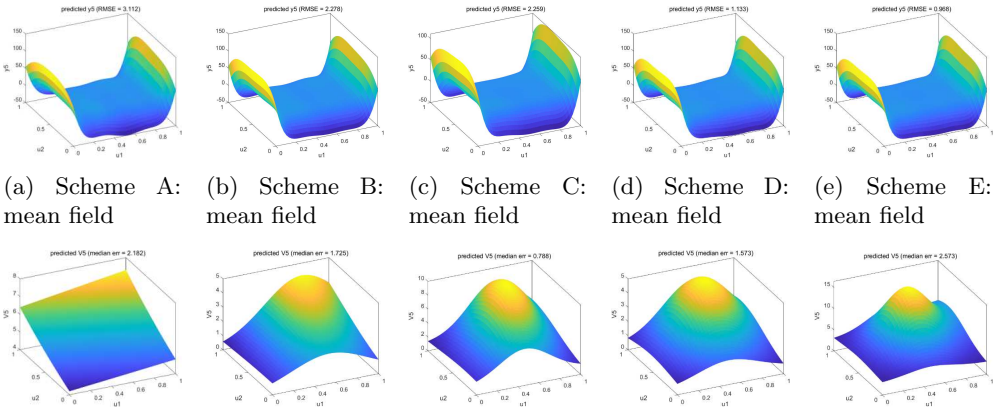
Note: entries are sites × replications. The per-run cost ratio is $F1{:}F2{:}F3{:}F4{:}F5 = 1{:}4{:}16{:}64{:}256$, and the total cost is reported in units of one F5 run; F5 is the six-hump top level and F1 the cheapest level.

Table 2: Modeling metrics of the five schemes.

Scheme	A	B	C	D	E
RMSE	3.12	2.278	2.259	1.13	0.968
Interval coverage	—*	87.6%	95.6%	95.4%	96.8%

*Interval underfitting.

Table 2 lists the five scheme modeling results, and Figure 6 the fitted mean and top-level variance fields of a representative trial. First, fusing additional cheap fidelity levels improves the accuracy monotonically while the total budget decreases: the RMSE falls from 3.12 for Scheme A to 2.278 for Scheme B, 2.259 for Scheme C, 1.13 for Scheme D, and 0.968 for Scheme E, while the total cost drops from 100 to 49.9. Second, the interval calibration improves as soon as the top-level noise shape can be inherited from densely replicated lower levels: Scheme A, built on only 2 replications per top-level site, cannot identify the top-level noise field and its intervals underfit (marked by *), and Scheme B still attains only 87.6% coverage; in contrast, Schemes C–E, whose top-level noise shapes are inherited from the replicated low-level latent fields, reach 95.4%–96.8% coverage, close to the nominal 95%. Third, Figure 6 shows the surface fits sharpening level by level: the mean surface of Scheme A is visibly pulled by the strong top-level noise and its fitted variance field is uninformative, whereas Schemes D and E reproduce the six-hump shape, and recover the four-lobed structure of V_5—the inherited noise shape, rather than the number of top-level replications, is the main driver of this improvement.



(a) Scheme A: mean field (b) Scheme B: mean field (c) Scheme C: mean field (d) Scheme D: mean field (e) Scheme E: mean field

(f) Scheme A: variance field (g) Scheme B: variance field (h) Scheme C: variance field (i) Scheme D: variance field (j) Scheme E: variance field

Fig. 6: Modeling effect of the five schemes: (a)–(e) predicted mean fields and (f)–(j) predicted variance fields.

3.3 Hartmann-4: comparison with other methods

3.3.1 Setup

To examine the behavior of the method in higher dimensions (the engineering case of Section 4 is a four-dimensional problem), this section pits the method against all comparison methods on the classical Hartmann-4 two-fidelity benchmark. The high-fidelity level takes

$$y_H(x) = \sum_{i=1}^{4} \alpha_i \exp\left(-\sum_{j=1}^{4} A_{ij} \left(x_j - P_{ij}\right)^2 \right), \qquad x \in [0,1]^4, \tag{64}$$

where $\alpha = (1, 1.2, 3, 3.2)$ and A, P are the classical parameter tables; the low-fidelity level is its closely correlated approximation,

$$y_L(x) = 0.85\, y_H(x) + 0.6 \sin^2(x_1 + x_3) + 0.4 \cos(3x_2) - 1.5. \tag{65}$$

The intrinsic variances of the two levels take the spatially interpretable "corner-enhanced" pattern,

$$V_H(x) = 0.5 + 4x_1x_2 + 2x_4, \qquad V_L(x) = 0.3 + 1.5x_1x_2 + x_4. \tag{66}$$

The design is a nested Latin hypercube with an enlarged budget: $k_L = 160$ low-fidelity points with $a_L = 8$ replications each, and $k_H = 64$ high-fidelity points (a subset of D_L) with $a_H = 8$ replications each, the replication counts being allocated by minimizing the effective noise of Eq. (47) under an equal-cost constraint. In high

dimensions the correlation length is no longer isotropic; this section switches each level to **per-dimension correlation lengths** $\theta \in \mathbb{R}^d$ (Gaussian kernel $R(x, x') = \exp\left(-\sum_{h=1}^{d} \theta_h \left(x_h - x'_h\right)^2\right)$), with the estimation of Section 2.3.1 and the gradients of Section 2.3.2 generalized accordingly to per-dimension form. Comparison methods: single-fidelity hetGP [13] (high-level data only, also with per-dimension correlation lengths; R package implementation [43]), SCK, and MCMC dual co-kriging (both levels fitted with a response GP and a log-noise GP, coupled by $z_H = \gamma z_L + \eta$, 10000 Gibbs sweeps, sample means used as single observations). The evaluation metrics are the RMSE on a 500-point Latin hypercube test grid, the empirical coverage CBI of the 95% prediction band, and the median error of the high-fidelity noise field on the grid, all averaged over 10 Monte Carlo runs.

3.3.2 Results

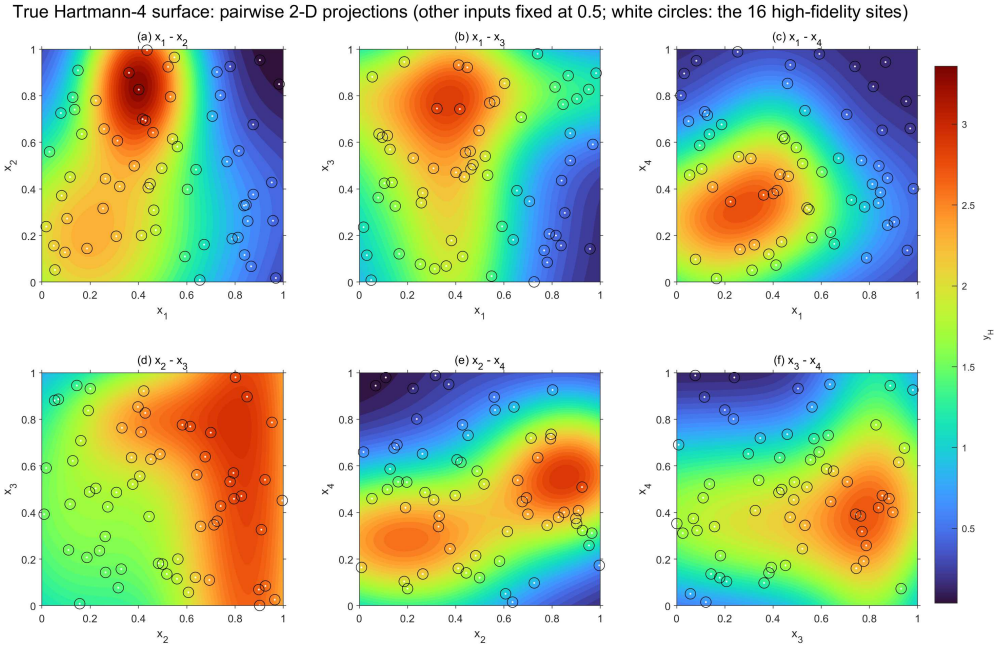


Fig. 7: Pairwise two-dimensional projections of the true four-dimensional Hartmann-4 high-fidelity surface (other inputs fixed at 0.5) with the 64 high-fidelity design sites (white circles).

Table 3 reports the comparison of the four methods at the enlarged budget. MF-hetGP dominates on all three metrics: the RMSE is reduced by 23.0% relative to single-fidelity hetGP and by 22.3% relative to SK, while the median error of the noise field is also the smallest. The MCMC dual co-kriging cannot be re-run at the enlarged budget: its per-sweep $O\left((k_L + k_H)^3\right)$ factorization grows cubically with the

design, and it was already two orders of magnitude slower than the two-step decoupled estimation at the original small design while remaining less accurate (Table 3, note)—further evidence of the cost-effectiveness of the Woodbury exact likelihood in high dimensions. Figure 7 visualizes the modeling effect of a representative run.

Table 3: Comparison of the four methods on the Hartmann-4 benchmark (enlarged budget $k_L = 160$, $a_L = 8$, $k_H = 64$, $a_H = 8$; 10 Monte Carlo runs, mean $\pm$ standard deviation).

Method	RMSE	Median error of noise field	CBI
MF-hetGP	0.370 ± 0.052	0.284 ± 0.073	1.00
hetGP (high level only)	0.480 ± 0.055	0.414 ± 0.130	1.00
SCK	0.476 ± 0.057	0.645 ± 0.108	1.00
MCMC dual co-kriging*	1.252 ± 0.068	1.233 ± 0.161	0.91

*Reference values from the original 16-site $\times$ 2-replication design: the per-sweep $O((k_L+k_H)^3)$ factorization of the MCMC sampler becomes prohibitive at the enlarged budget, so it could not be re-run here.

Figure 7 shows the pairwise two-dimensional projections of the true four-dimensional surface, which explain the ranking in Table 3. The response concentrates its variation in the sections involving x_1 (panels (a)–(c)): sharp ridges and multimodal troughs run across the x_1–x_2 and x_1–x_3 sections, while the sections spanned only by x_2, x_3, x_4 (panels (d)–(f)) are considerably smoother. It is on these sharply varying regions that the difference between the methods concentrates: even with 64 high-fidelity sites (white circles) each replicated 8 times, single-fidelity hetGP must resolve this structure from its own 64 sites alone, and the chi-square-corrupted pointwise variances of SCK further degrade its fit—hence their RMSE of 0.480 and 0.476 in Table 3. MF-hetGP instead borrows the global shape from the 160 replicated low-fidelity sites through the autoregressive structure, which is why its RMSE of 0.370 is 23.0% and 22.3% lower than the two baselines, and why its fitted noise field attains the smallest median error (0.284) while keeping the prediction bands at the nominal coverage.

Figure 8 shows the MF-hetGP fitted surface on the same six sections, with the same color scale as Figure 7 so that the two are directly comparable. The fitted surface reproduces the ridges, the multimodal troughs, and the locations of the peaks almost exactly: over the union of the six section grids the RMSE of the fit is 0.313, consistent with the 0.370 reported in Table 3. The only visible deviations are confined to the sharpest ridges of the x_1–x_2 and x_1–x_3 sections (panels (a) and (b)), where the 64 high-fidelity sites are sparsest relative to the local curvature and the fitted contours are slightly rounded—exactly the regions that limit the accuracy of all four methods. Everywhere else the fitted surface is visually indistinguishable from the truth, which explains why MF-hetGP attains the smallest RMSE and the smallest noise-field error in Table 3.

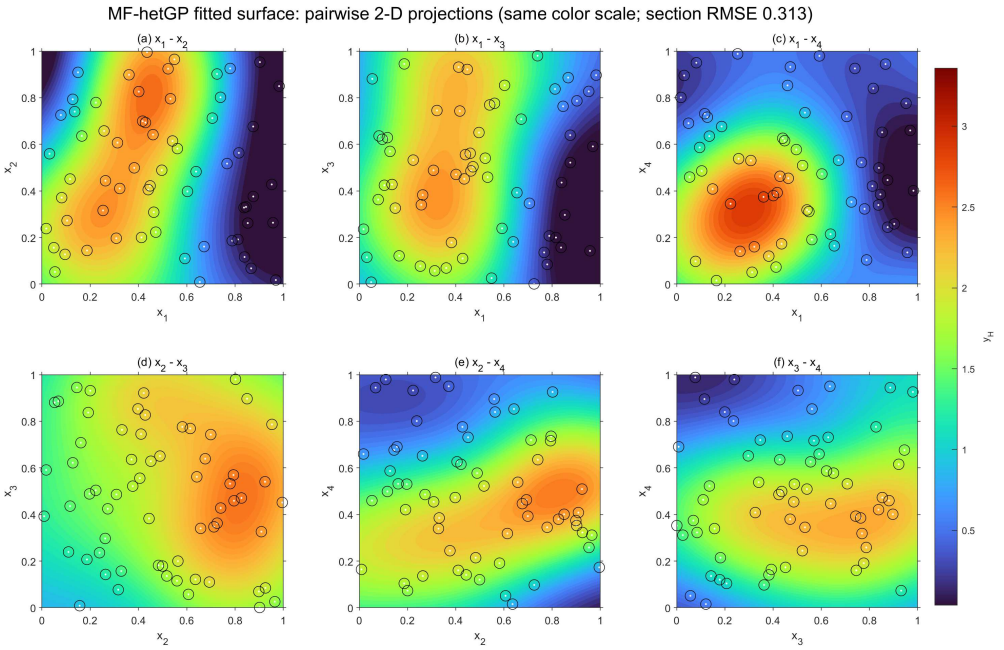


Fig. 8: The MF-hetGP fitted surface on the same six sections and with the same color scale as Figure 7.

4 Engineering case study: radar seeker performance evaluation

4.1 Background and research objectives

The radar seeker is a key equipment for a missile to achieve terminal guidance. With the development of electronic and information warfare, jamming and anti-jamming have become important means of confrontation between the attacking and defending sides; the electromagnetic combat environment faced by the seeker is increasingly complex, and anti-jamming performance evaluation has become an important subject in equipment development, qualification, and tactical use, among which the performance response surface characterizes the performance level of the seeker under various situations and is a core indicator of battlefield applicability.

Range gate pull-off (RGPO) jamming belongs to off-board active angle deception jamming. The target releases a decoy that generates jamming signals, making the radar simultaneously receive signals from multiple directions, deceiving the radar angle measurement and preventing the true target from being locked on. In the experiment the radar is fixed, the target is at range R, the distance between the target and the towed decoy is l, both move at velocity v, and the bandwidth of the jamming signal released by the decoy is β; the experiment scenario is shown in Figure 9[44]. The entire search, identification, and tracking process is completed by the built-in algorithm of the seeker, which presents itself externally as a black-box system.

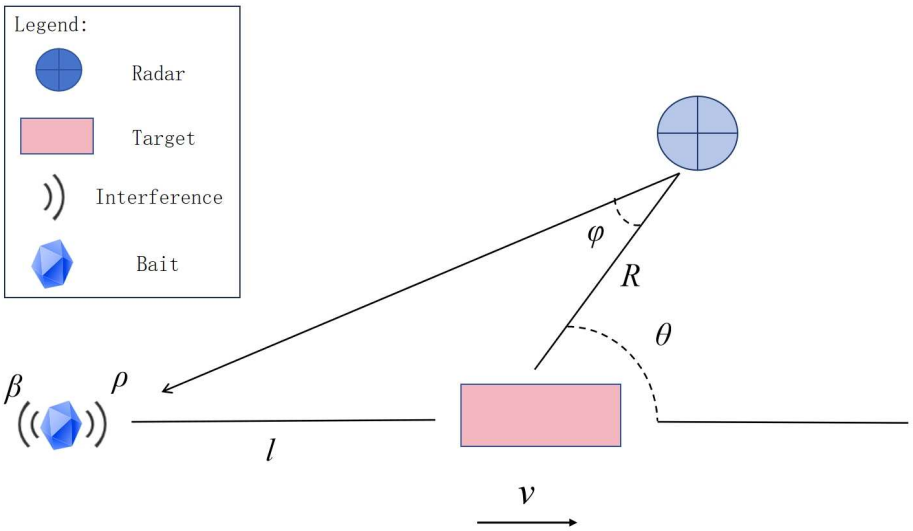


Fig. 9: Experiment setting of the radar seeker performance evaluation experiment.[44]

This experiment takes the radar stable angular tracking error φ as the response. The definitions and ranges of the experimental factors are shown in Table 4. Radar seeker anti-jamming performance evaluation usually relies on a multi-fidelity experiment system composed of all-digital simulation, hardware-in-the-loop (HWIL) simulation, and real tests; the characteristics of the three experiment means are shown in Table 5.

Table 4: Definitions and ranges of the experimental factors.

Factor	Symbol	Name	Unit	Range
x_1	v	Target velocity	kts	$[0, 1]$
x_2	l	Deception distance	m	$[0, 1]$
x_3	R	Target range	km	$[0, 1]$
x_4	β	Jamming bandwidth	MHz	$[0, 1$

Note: the ranges are the actual values of the case data; all factors are normalized to $[0, 1]$ for modeling.

Multi-fidelity experimental data exhibit the typical contradiction between fidelity and sample size: the higher the fidelity, the fewer the samples. Modeling with only the scarce high-fidelity data cannot capture the details of the response surface; modeling with only the abundant low-fidelity data transfers its systematic bias directly to the final evaluation conclusions. In addition, this case involves heteroscedastic noise variance, whose neglect would also make the response surface model inaccurate.

Table 5: The three fidelity experiment means and their characteristics.

Experiment means	Realization	Sample size	Cost	Noise characteristics
All-digital simulation	Reproduce target scattering, jamming signals, and the guidance loop in a software environment	Thousands	Low	Dominated by Monte Carlo randomness and numerical errors; low noise level but with obvious systematic bias
HWIL simulation	The seeker is connected to the simulation loop and jamming is injected by a signal simulator or in an anechoic chamber	Hundreds	Higher	Superimposed hardware thermal noise, signal simulation errors, and equipment state drift; noise varies with the experimental conditions
Real test	Equipment test in a realistic field electromagnetic environment	Tens	High	Affected by random factors such as weather and electromagnetic environment; noise variance is large and varies markedly with the scenario

Note: the sample sizes are typical orders of magnitude; the samples actually obtainable in real tests are usually fewer.

4.2 MF-hetGP modeling accuracy verify

To examine the modeling effect of the MF-hetGP model in this case, a full-space approximate set obtained from extensive multi-fidelity experiments is used as the benchmark data. In the experimental design space, a nested quasi-uniform design is adopted: 64 quasi-uniformly distributed digital-simulation points are first generated in the four-dimensional experimental space, from which 32 HWIL simulation points and 16 real-test points are then nested. Figure 10 shows the six pairwise two-dimensional projections of this design; the points of the three levels all cover the whole experimental space and strictly satisfy the nesting relation real $\subset$ HWIL $\subset$ digital, guaranteeing the effective transfer of the low-level trend information to the higher levels.

The comparison results of the three modeling schemes are shown in Table 6 (means and standard deviations of 10 Monte Carlo runs; the RMSE and coverage are referenced to the noise-free truth, and the 95% interval is taken as ± 1.96 times the total predictive variance).

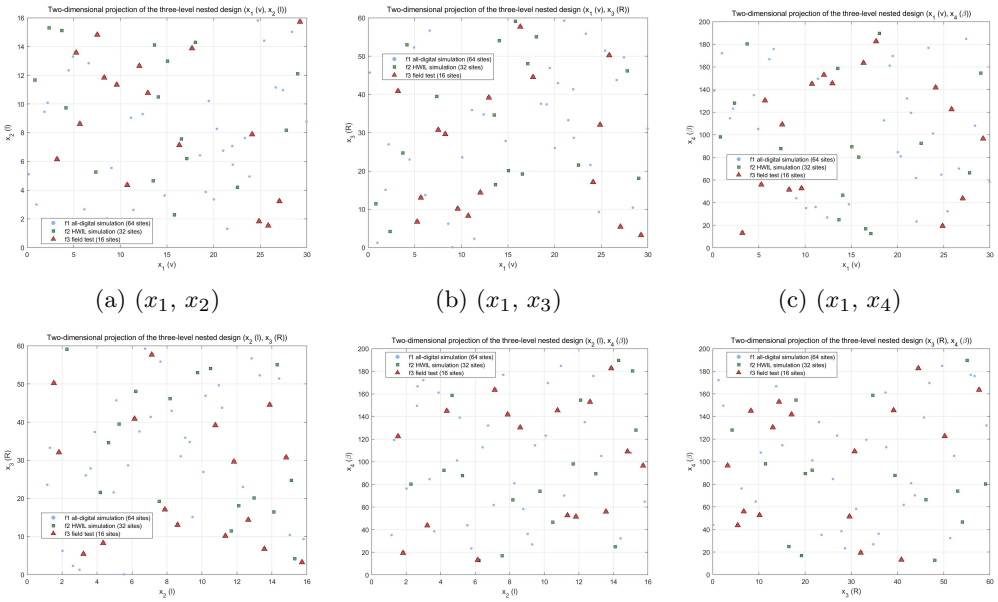
Table 6: Accuracy and coverage comparison of the three-fidelity modeling and the reference schemes.

Modeling scheme	RMSE (mean $\pm$ std)	95% coverage (%)	Median $\lvert V_3 \rvert$ error	Fitted ρ_2 / ρ_3
D3 (real data only)	2.338 ± 0.617	86.7 ± 8.6	1.184 ± 0.369	—
D23 (HWIL + field)	1.898 ± 0.134	93.8 ± 1.8	1.015 ± 0.325	—
D123 (three-fidelity)	1.567 ± 0.147	96.3 ± 2.3	0.617 ± 0.200	1.06 / 1.08

Note: ρ_2 and ρ_3 are the fitted means of the inter-fidelity transfer coefficients, with true values of about 1.18 and 1.14.

From Table 6, the RMSE of the proposed method is 1.567, a reduction of 33.0% relative to the single-fidelity method and a further reduction of 17.4% relative to

the two-level fusion method, and the standard deviation over the 10 runs (0.147) is also clearly smaller—the modeling effect is stable. Moreover, the 95% interval coverage of the proposed method is 96.3%, closest to the nominal level, whereas the single-fidelity method attains only 86.7% with a fluctuation of ±8.6%—the uncertainty quantification is more reliable.



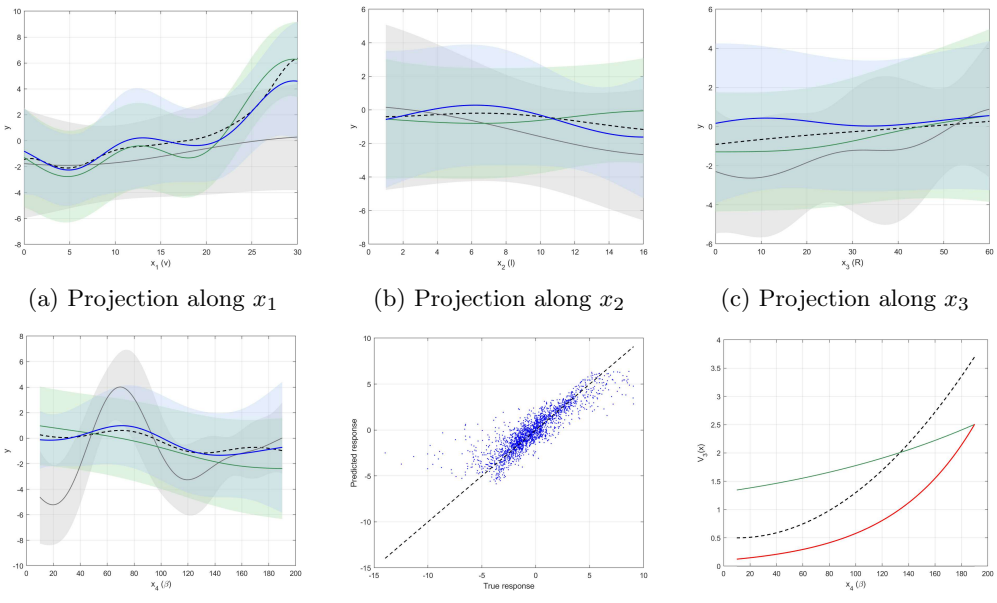
(a) (x_1, x_2) (b) (x_1, x_3) (c) (x_1, x_4)

(d) (x_2, x_3) (e) (x_2, x_4) (f) (x_3, x_4)

Fig. 10: Two-dimensional projections of the three-level nested design: (a) (x_1, x_2), (b) (x_1, x_3), (c) (x_1, x_4), (d) (x_2, x_3), (e) (x_2, x_4), (f) (x_3, x_4). Blue circles: 64 all-digital simulation points; green squares: 32 HWIL simulation points; red triangles: 16 real-test points.

Figure 11 shows the modeling effect, where (a)–(d) are the projection plots on the four dimensions, from which it can be seen that the D123 scheme fits best, with essentially full coverage on the profiles; (e) and (f) show the fitting effect of the model—both the response surface and the noise function are fitted well.

In summary, the advantages of the proposed method over conventional schemes lie in three respects. (1) Multi-fidelity information fusion: the low-cost data of digital and HWIL simulation provide the global trend information of the response, making up for the scarcity of real-test samples, and the prediction accuracy at equal budget is improved by 33% relative to the single-fidelity method. (2) Reliable heteroscedastic uncertainty quantification: the heteroscedastic structure of the noise variance varying with the experimental conditions is modeled correctly, the 95% interval coverage is close to the nominal level, and the distortion of the evaluation conclusions caused by overly wide or overly narrow intervals under the homoscedastic assumption is



(a) Projection along x_1 (b) Projection along x_2 (c) Projection along x_3

(d) Projection along x_4 (e) Data fitting (f) Noise variance estimate

Fig. 11: Modeling effect visualization of the seeker case: (a)–(d) profile fits with 95% intervals of the three schemes along x_1–x_4 versus the truth (black dashed; gray: D3, green: D23, blue: D123); (e) the prediction–truth scatter of D123 (RMSE 1.567, coverage 96.3%); (f) the estimated top-level noise variance $V_3(x)$ versus the truth.

avoided. (3) Efficient use of the top-level samples: the affine noise inheritance lets the top fidelity identify its noise shape without replicated experiments, concentrating the scarce real-test resources on acquiring response information. The above results validate the reliability of the proposed method in practical engineering applications.

5 Conclusions

This paper proposed a multi-fidelity heteroscedastic Gaussian process (MF-hetGP) for stochastic simulation experiments. The method infers the low-fidelity noise variance jointly with the mean field through a latent-variable Gaussian process and transfers the fitted noise shape to the high-fidelity level through an affine noise link, so that the high-fidelity noise structure can be identified from only a few high-fidelity samples. A closed-form full-data likelihood with closed-form gradients, derived via the Woodbury identity, keeps the training efficient, and the model extends to any number of fidelity levels with a parameter cost growing linearly in the number of levels. Numerically, MF-hetGP reduces the high-fidelity sampling requirement by more than 70% on the Forrester benchmark, lowers the RMSE by 23.0% and 22.3% relative to single-fidelity hetGP and stochastic kriging on the Hartmann-4 benchmark, and improves accuracy while halving the total cost in the five-fidelity study. In the radar seeker

performance evaluation case, three-fidelity fusion reduces the RMSE by 33.0% compared with single-fidelity modeling and achieves 96.3% coverage of the nominal 95% prediction intervals.

Future work will extend the method to higher-dimensional problems and larger fidelity hierarchies, convert the structural lower bounds and the Laplace prior scales into data-driven criteria, and develop adaptive sequential designs combined with multi-fidelity Bayesian optimization. Besides, a new approach will be developed together with asymptotic guarantees for the parameter estimators and the prediction intervals, and acceleration by sparse methods [45].

Declarations

- **Funding** This work was supported by the NSAF Joint Fund of the National Natural Science Foundation of China (Grant No. U2230208), the National Defense Key Laboratory Fund (Grant No. SFY-GH-Q7-20240865) and Innovation Research Foundation of National University of Defense Technology (ZK25-10).

- **Conflict of interest** The authors declare that they have no conflict of interest.
- **Ethics approval and consent to participate** Not applicable.
- **Consent for publication** Not applicable.
- **Data availability** The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.
- **Code availability** The code used in this study is available from the corresponding author on reasonable request.
- **Author contribution** All authors contributed to the study conception and design. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Appendix A Woodbury derivation of the low-level concentrated log-likelihood

To keep the main text coherent, the formulas of Sections 2.2 and 2.3 are all stated as definitions or conclusions there; the complete derivations are collected here. The notation is consistent with the main text. The full-data log-likelihood of all N_L raw low-fidelity observations $\boldsymbol{t}_L \in \mathbb{R}^{N_L}$ under the zero-mean Gaussian field and independent heteroscedastic noise is

$$\log \mathcal{L}_L^{\text{full}} = -\frac{1}{2}\ln\left|\mathbf{C}_N + \boldsymbol{\Lambda}_N\right| - \frac{1}{2}\boldsymbol{t}_L^{\top}\left(\mathbf{C}_N + \boldsymbol{\Lambda}_N\right)^{-1}\boldsymbol{t}_L - \frac{N_L}{2}\ln(2\pi), \tag{A1}$$

where $\mathbf{C}_N$ is the field correlation matrix over the N_L observations (blocked by replications) and $\boldsymbol{\Lambda}_N = \operatorname{diag}\left(V_L(x_i)/\tau_L^2\right)$ is the corresponding noise-to-field variance ratio. Applying the Woodbury identity [46] block by block over the replication structure

(Lemmas 3.1–3.2 of [13]), the determinant term expands to

$$\frac{1}{2}\ln|\mathbf{C}_N+\mathbf{\Lambda}_N| = \frac{1}{2}\ln|\mathbf{\Upsilon}_{nL}| + \frac{1}{2}\sum_{i=1}^{k_L}\left[(a_{L,i}-1)\ln\Lambda_{L,i}+\ln a_{L,i}\right], \tag{A2}$$

where $\mathbf{\Upsilon}_{nL} = \mathbf{R}_L(X_L,X_L)+\mathbf{\Lambda}_L\mathbf{A}_L^{-1}$ is the covariance reduced to unique-point dimension k_L; the quadratic-form term expands through the inverse form of the same identity into

$$\boldsymbol{t}_L^\top\left(\mathbf{C}_N+\mathbf{\Lambda}_N\right)^{-1}\boldsymbol{t}_L = \sum_{i=1}^{k_L}\frac{SS_{L,i}}{\Lambda_{L,i}} - \sum_{i=1}^{k_L}\frac{a_{L,i}\,\bar{y}_L(x_i)^2}{\Lambda_{L,i}} + \bar{\boldsymbol{y}}_L^\top\mathbf{\Upsilon}_{nL}^{-1}\bar{\boldsymbol{y}}_L, \tag{A3}$$

where $\mathrm{SS}_{L,i}$ is the within-site sum of squares. Substituting Eqs. (A2)–(A3) into Eq. (A1), differentiating with respect to the field scale τ_L^2 and setting to zero gives the concentrated estimate (noting that $\mathbf{\Lambda}_N$ and $\mathbf{\Lambda}_L$ differ by the factor τ_L^2, the solution of $\partial(\cdot)/\partial\tau_L^2 = 0$ is the concentrated estimate), and substituting back gives the data part of Eq. (30); adding the latent prior $\log p(\mathbf{\Delta}_L) = -\frac{k_L}{2}\ln(2\pi\nu_{gL}) - \frac{1}{2\nu_{gL}}\mathbf{\Delta}_L^\top\left(\mathbf{\Upsilon}_L^{(g)}\right)^{-1}\mathbf{\Delta}_L - \frac{1}{2}\ln\left|\mathbf{\Upsilon}_L^{(g)}\right|$ (with $\frac{1}{2}\ln\left|\mathbf{\Upsilon}_L^{(g)}\right| = \sum_i \ln[\mathbf{L}_{gL}]_{ii}$) yields Eq. (30). Hence the low-level objective differs from the full-data likelihood only by constants.

Appendix B Derivation of the high-level likelihood

The high-level full data consist of two parts: one difference statistic $\delta(x_i)$ per high-level site, and the $a_{H,i}$ raw replications within each site. The difference part follows a k_H-dimensional Gaussian distribution (Eq. (21) is the covariance and the constant mean is β_H); after concentrating out β_H (Eq. (31)), its log-density is

$$-\frac{k_H}{2}\ln(2\pi) - \frac{1}{2}\boldsymbol{r}^\top\mathbf{W}_{n,H}^{-1}\boldsymbol{r} - \sum_{i=1}^{k_H}\ln[\mathbf{L}_{nH}]_{ii}\,. \tag{B4}$$

The within-site part: given $V_{H,i}$, the within-site sum of squares of the $a_{H,i}$ i.i.d. Gaussian replications at site i satisfies $\mathrm{SS}_{w,H,i}/V_{H,i} \sim \chi^2_{a_{H,i}-1}$, whose log-density (up to constants) is

$$-\frac{1}{2}\sum_{i=1}^{k_H}\left[(a_{H,i}-1)\ln V_{H,i} + \frac{SS_{w,H,i}}{V_{H,i}}\right]. \tag{B5}$$

Adding the two independent parts gives Eq. (32). By the Markov property of the autoregressive structure (13)–(14), the two-level fields share no cross information given $Y_L(X_H)$, so the sum of Eq. (30) and Eq. (32) is the joint full-data likelihood (up to constants). The within-site term is independent of ρ and of the difference field, and it is the key to the identifiability of ρ when $a_{H,i} \geq 2$: it anchors $V_{H,i}$ on the scale of the within-site sums of squares, preventing V_H from inflating freely to absorb the residuals of a mismatched ρ.

Appendix C Derivation of the closed-form gradients

Low level. Differentiating Eq. (30) with respect to $\Lambda_{L,i}$: the data part through the dependence of ν_L on $\Lambda_{L,i}$ in Eq. (30),

$$\frac{\partial \nu_L}{\partial \Lambda_{L,i}} = -\frac{SS_{L,i}}{N_L \Lambda_{L,i}^2} + \frac{a_{L,i}\bar{y}_L(x_i)^2}{N_L \Lambda_{L,i}^2} - \frac{p_{L,i}^2}{N_L a_{L,i}}, \tag{C6}$$

with $\boldsymbol{p}_L = \boldsymbol{\Upsilon}_{nL}^{-1}\bar{\boldsymbol{y}}_L$, uses $\partial \boldsymbol{\Upsilon}_{nL}/\partial \Lambda_{L,i} = a_{L,i}^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^\top$ and the matrix differentiation identities; merging with the derivatives of the explicit terms $-\frac{1}{2}(a_{L,i}-1)\ln \Lambda_{L,i}$ and $-\ln[\mathbf{L}_{nL}]_{ii}$, namely $-\frac{1}{2a_{L,i}}\left[\boldsymbol{\Upsilon}_{nL}^{-1}\right]_{ii}$, gives Eq. (35) (simplified with $s_{b,L,i}^2 = S_{L,i}^2(a_{L,i}-1)/a_{L,i}$ and $\mathrm{SS}_{L,i} = a_{L,i}\bar{y}_L^2 + (a_{L,i}-1)S_{L,i}^2$).

Latent-variable chain: from Eqs. (18)–(17), $\partial \boldsymbol{\Lambda}_L/\partial \boldsymbol{z}_L = \mathrm{soft}'(\boldsymbol{z}_L)\odot \boldsymbol{\Lambda}_L$ and $\partial \boldsymbol{z}_L/\partial \boldsymbol{\Delta}_L = \mathbf{C}_L^{(g)}\left(\boldsymbol{\Upsilon}_L^{(g)}\right)^{-1}$, hence

$$\frac{\partial \log \mathcal{L}_L}{\partial \boldsymbol{\Delta}_L} = \left(\mathbf{C}_L^{(g)}\left(\boldsymbol{\Upsilon}_L^{(g)}\right)^{-1}\right)^\top \left(\mathrm{soft}'(\boldsymbol{z}_L)\odot \boldsymbol{\Lambda}_L \odot \frac{\partial \log \mathcal{L}_L}{\partial \boldsymbol{\Lambda}_L}\right) - \frac{1}{\nu_{gL}}\left(\boldsymbol{\Upsilon}_L^{(g)}\right)^{-1}\boldsymbol{\Delta}_L, \tag{C7}$$

the last term coming from the gradient of the prior term $-\boldsymbol{\Delta}_L^\top\left(\boldsymbol{\Upsilon}_L^{(g)}\right)^{-1}\boldsymbol{\Delta}_L/(2\nu_{gL})$. For θ_L: the mean-field part is differentiated through the trace and quadratic forms of $\boldsymbol{\Upsilon}_{nL}$ (using $\partial \mathbf{R}/\partial \theta_h = -\mathbf{R}\odot \mathbf{D}_h^2$, with $\mathbf{D}_h$ the pairwise distance matrix of the h-th dimension); for φ_L, g_L: the traces, quadratic forms of $\boldsymbol{\Upsilon}_L^{(g)}$, and the chain through $\boldsymbol{z}_L$—the structure of Eqs. (19)–(22) of [13], omitted here. Under the parameterization (36), the chain factors such as $\partial \varphi_L/\partial w = \varphi_L - \varphi_{\min}$ are multiplied in term by term following Eq. (36).

High level. For Eq. (32): from $\partial\left(\boldsymbol{r}^\top \mathbf{W}_{n,H}^{-1}\boldsymbol{r}\right)/\partial \sigma_{H,i}^2 = -\left(\mathbf{W}_{n,H}^{-1}\boldsymbol{r}\right)_i^2 = -p_{H,i}^2$ (the envelope theorem eliminates the chain term of β_D, since $\boldsymbol{r}$ and $\mathbf{1}$ are orthogonal with respect to $\mathbf{W}_{n,H}^{-1}$) and $\partial \ln|\mathbf{W}_{n,H}|/\partial \sigma_{H,i}^2 = \left[\mathbf{W}_{n,H}^{-1}\right]_{ii}$, the first equation of Eq. (37) follows; from $\partial \boldsymbol{r}^\top \mathbf{W}_{n,H}^{-1}\boldsymbol{r}/\partial \tau_H^2 = -\boldsymbol{p}_H^\top \mathbf{C}_H \boldsymbol{p}_H$ and $\partial \ln|\mathbf{W}_{n,H}|/\partial \tau_H^2 = \mathrm{tr}\left(\mathbf{W}_{n,H}^{-1}\mathbf{C}_H\right)$, the second. The derivative of the within-site term $g_i(V_{H,i}) = -\frac{1}{2}\left[(a_{H,i}-1)\ln V_{H,i} + \mathrm{SS}_{w,H,i}/V_{H,i}\right]$ is

$$\frac{\partial g_i}{\partial V_{H,i}} = -\frac{a_{H,i}-1}{2V_{H,i}} + \frac{\mathrm{SS}_{w,H,i}}{2V_{H,i}^2}. \tag{C8}$$

Then, through the chain relations of Eq. (20), $\partial \sigma_{H,i}^2/\partial \eta_0 = 1/a_{H,i}$, $\partial \sigma_{H,i}^2/\partial \log \eta_1 = \eta_1 V_L(x_{H,i})/a_{H,i}$, $\partial V_{H,i}/\partial \eta_0 = 1$, $\partial V_{H,i}/\partial \log \eta_1 = \eta_1 V_L(x_{H,i})$, merging gives Eq. (38) (the contribution of the within-site term is included as $\sum_i (\partial g_i/\partial V_{H,i})(\partial V_{H,i}/\partial \cdot)$). Under the M-level generalization the gradients of each level are isomorphic, with only the subscripts replaced.

References

[1] Bhosekar, A., Ierapetritou, M.: Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Industrial & Engineering Chemistry Research **57**(15), 5432–5470 (2018)

[2] Liu, H., Ong, Y.-S., Cai, J.: A survey of adaptive sampling for global metamodeling in support of simulation-based engineering applications. ACM Computing Surveys **51**(6), 1–36 (2018)

[3] Zhang, Y., Kim, C.W., Tee, K.F.: Application of artificial neural networks in structural engineering: A review. Journal of Civil Structural & Environmental Engineering **6**(3), 231–247 (2016)

[4] Kumar, A., Sridhara, S., Suresh, K.: A novel surrogate model for optimizing lattice for desired force-displacement response. Engineering With Computers **42**(5) (2026)

[5] Pinheiro, N.C., Guo, D., Menke, H.P., Joshi, A.C., Heaney, C.E., Elsheikh, A.H., Pain, C.C.: Surrogate models for rock–fluid interaction: a grid-size-invariant approach. Engineering With Computers **42**(4) (2026)

[6] Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P.: Design and analysis of computer experiments. Statistical Science **4**(4), 409–423 (1989)

[7] Santner, T.J., Williams, B.J., Notz, W.I.: The Design and Analysis of Computer Experiments. Springer, New York (2003)

[8] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)

[9] Yao, Z., Wu, W., Huang, H.: Non-stationary multi-fidelity surrogate model based on spatial warping. Engineering With Computers **41**(5), 3347–3364 (2025)

[10] Peng, C., Wu, C.F.J.: On the choice of nugget in kriging modeling for deterministic computer experiments. Journal of Computational and Graphical Statistics **23**(1), 151–168 (2012)

[11] Wang, W., Haaland, B.A.: Controlling sources of inaccuracy in stochastic kriging. Technometrics **61**(3), 309–321 (2018)

[12] Ankenman, B.E., Nelson, B.L., Staum, J.: Stochastic kriging for simulation metamodeling. Operations Research **58**(2), 371–382 (2010)

[13] Binois, M., Gramacy, R.B., Ludkovski, M.: Practical heteroscedastic Gaussian process modeling for large simulation experiments. Journal of Computational and Graphical Statistics **27**(4), 808–821 (2018)

[14] Ozbayram, O., Olivier, A., Graham-Brady, L.: Heteroscedastic Gaussian process regression for material structure–property relationship modeling. Computer Methods in Applied Mechanics and Engineering **431**, 117326 (2024)

[15] Goldberg, P.W., Williams, C.K.I., Bishop, C.M.: Regression with input-dependent noise: A Gaussian process treatment. In: Advances in Neural Information Processing Systems, vol. 10, pp. 493–499 (1997)

[16] Bishop, C.M., Qazaz, C.S.: Regression with input-dependent noise: A Bayesian treatment. In: Advances in Neural Information Processing Systems, vol. 9, pp. 347–354 (1996)

[17] Boukouvalas, A., Cornford, D.: Learning heteroscedastic Gaussian processes for complex datasets. Technical report, Aston University, Neural Computing Research Group (2009)

[18] Kersting, K., Plagemann, C., Pfaff, P., Burgard, W.: Most likely heteroscedastic Gaussian process regression. In: Proceedings of the 24th International Conference on Machine Learning, pp. 393–400 (2007)

[19] Lázaro-Gredilla, M., Titsias, M.K.: Variational heteroscedastic Gaussian process regression. In: Proceedings of the 28th International Conference on Machine Learning, pp. 841–848 (2011)

[20] Montgomery, D.C., Peck, E.A., Vining, G.G.: Introduction to Linear Regression Analysis, 5th edn. Wiley, New York (2012)

[21] Chu, W., Keerthi, S.S., Ong, C.J.: Bayesian support vector regression using a unified loss function. IEEE Transactions on Neural Networks **15**(1), 29–44 (2004)

[22] Cawley, G.C., Talbot, N.L.C., Foxall, R.J., Dorling, S.R., Mandic, D.P.: Heteroscedastic kernel ridge regression. Neurocomputing **57**, 105–124 (2004)

[23] Nix, D.A., Weigend, A.S.: Estimating the mean and variance of the target probability distribution. In: Proceedings of the IEEE International Conference on Neural Networks, pp. 55–60 (1994)

[24] Kendall, A., Gal, Y.: What uncertainties do we need in Bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems, vol. 30, pp. 5574–5584 (2017)

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

[26] Han, Z.-H., Görtz, S.: Hierarchical kriging model for variable-fidelity surrogate modeling. AIAA Journal **50**(9), 1885–1896 (2012)

[27] Ehara, A., Guillas, S.: AN ADAPTIVE STRATEGY FOR SEQUENTIAL

DESIGNS OF MULTILEVEL COMPUTER EXPERIMENTS. International Journal for Uncertainty Quantification **13**(4), 61–98 (2023)

[28] Heo, J., Sung, C.-L.: Active learning for a recursive non-additive emulator for multi-fidelity computer experiments. Technometrics **67**(1), 58–72 (2025)

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

[30] Qian, P.Z.G., Wu, C.F.J.: Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics **50**(2), 192–204 (2008)

[31] Tuo, R., Wu, C.F.J., Yu, D.: Surrogate modeling of computer experiments with different mesh densities. Technometrics **56**(3), 372–378 (2014)

[32] Wang, X., Mak, S., Miller, J., Wu, J.: Local transfer learning Gaussian process modeling, with applications to surrogate modeling of expensive computer simulators. SIAM/ASA Journal on Uncertainty Quantification **14**(1), 256–286 (2026)

[33] Sung, C., Ji, Y., Mak, S., Wang, W., Tang, T.: Stacking designs: Designing multi-fidelity computer experiments with target predictive accuracy. SIAM/ASA Journal on Uncertainty Quantification **12**(1), 157–181 (2024)

[34] Kou, M., Dong, X., Liu, H., Zhou, Q., Lin, Q.: An adaptive multi-fidelity surrogate-based robust optimization approach considering the combined effect of various uncertainties. Engineering With Computers **42**(2) (2026)

[35] Chen, Y.-P., Wang, L., Comlek, Y., Chen, W.: A latent variable approach for non-hierarchical multi-fidelity adaptive sampling. Computer Methods in Applied Mechanics and Engineering **421**, 116773 (2024)

[36] Baillie, N., Kerleguer, B., Feau, C., Garnier, J.: Efficient multi-fidelity Gaussian process regression for noisy outputs and non-nested experimental designs. arXiv:2511.20183 (2025)

[37] Stroh, R., Bect, J., Demeyer, S., Fischer, N., Marquis, D., Vázquez, E.: Sequential design of Multi-Fidelity computer experiments: Maximizing the rate of stepwise uncertainty reduction. Technometrics **64**(2), 199–209 (2021)

[38] Huang, Y., Wong, S.W.K.: Fully bayesian sequential design for heteroscedastic stochastic simulations. Technometrics, 1–16 (2026)

[39] Chen, X., Hemmati, S., Yang, F.: Stochastic co-kriging for steady-state simulation metamodeling. In: Proceedings of the 2017 Winter Simulation Conference, pp. 1750–1761 (2017)

[40] Binois, M., Huang, J., Gramacy, R.B., Ludkovski, M.: Replication or exploration? Sequential design for stochastic simulation experiments. Technometrics **61**(1), 7–23 (2019)

[41] Forrester, A.I.J., Sóbester, A., Keane, A.J.: Multi-fidelity optimization via surrogate modelling. Proceedings of the Royal Society A **463**(2088), 3251–3269 (2007)

[42] Liu, Z., Yan, L., Duan, X., Xiao, Y., Liu, B., Chen, J.: A novel multi-fidelity sequential optimization method based on multi-level gaussian process. Engineering Applications of Artificial Intelligence **162**, 112678 (2025)

[43] Binois, M., Gramacy, R.B.: hetGP: Heteroscedastic Gaussian process modeling and sequential design under replication. R package. https://cran.r-project.org/package=hetGP (2021)

[44] Liu, J., Xiao, Y., Duan, X., Chen, X., Wang, Z., Liu, Z.: A novel batch-selection strategy for parallel global optimization. Engineering Optimization **57**(3), 599–623 (2025)

[45] Snelson, E., Ghahramani, Z.: Sparse Gaussian processes using pseudo-inputs. In: Advances in Neural Information Processing Systems, vol. 18, pp. 1257–1264 (2005)

[46] Harville, D.A.: Matrix Algebra from a Statistician's Perspective. Springer, New York (1997)