

Physics-Informed CoKriging: A Gaussian-Process-Regression-Based Multifidelity Method for Data-Model Convergence

Xiu Yang^{*1}, David Barajas-Solano^{†1}, Guzel Tartakovsky^{‡2}, and Alexandre M. Tartakovsky^{§1}

¹Advanced Computing, Mathematics and Data Division, Pacific Northwest National Laboratory, Richland, WA 99352

²Hydrology Group, Pacific Northwest National Laboratory, Richland, WA 99352

November 27, 2018

Abstract

In this work, we propose a new Gaussian process regression (GPR)-based multifidelity method: physics-informed CoKriging (CoPhIK). In CoKriging-based multifidelity methods, the quantities of interest are modeled as linear combinations of multiple parameterized stationary Gaussian processes (GPs), and the hyperparameters of these GPs are estimated from data via optimization. In CoPhIK, we construct a GP representing low-fidelity data using physics-informed Kriging (PhIK), and model the discrepancy between low- and high-fidelity data using a parameterized GP with hyperparameters identified via optimization. Our approach reduces the cost of optimization for inferring hyperparameters by incorporating partial physical knowledge. We prove that the physical constraints in the form of deterministic linear operators are satisfied up to an error bound. Furthermore, we combine CoPhIK with a greedy active learning algorithm for guiding the selection of additional observation locations. The efficiency and accuracy of CoPhIK are demonstrated for reconstructing the partially observed modified Branin function, reconstructing the sparsely observed state of a steady state heat transport problem, and learning a conservative tracer distribution from sparse tracer concentration measurements.

Keywords: physics-informed, Gaussian process regression, CoKriging, multifidelity, active learning, error bound.

1 Introduction

Gaussian processes (GPs) are a widely used tool in applied mathematics, statistics, and machine learning for regression, classification, and optimization [14, 40, 44]. GP regression (GPR), also known as *Kriging* in geostatistics, constructs a statistical model of a partially observed process by assuming that its observations are a realization of a GP. A GP is uniquely described by its mean and covariance function (also known as *kernel*). In standard (referred to here as *data-driven*) GPR, usually parameterized forms of mean and covariance functions are assumed, and

^{*}xiu.yang@pnnl.gov

[†]david.barajas-solano@pnnl.gov

[‡]guzel.tartakovsky@pnnl.gov

[§]alexandre.tartakovsky@pnnl.gov

the hyperparameters of these functions (e.g., variance and correlation length) are estimated from data by maximizing the log marginal likelihood of the data. GPR is also closely related to kernel machines in machine learning, but it provides a richer characterization in the result, as it provides uncertainty estimates [48]. GP is also connected to infinite neural networks, that is, networks with an infinite number of hidden units [29].

There are several variants of GPR, including simple, ordinary, and universal Kriging [21]. Ordinary Kriging is the most widely used GPR method. It assumes stationarity of the random field, including constant mean and variance, and a prescribed stationary covariance function. The stationarity assumption reduces the number of hyperparameters and the model complexity. For example, in universal Kriging, the mean is modeled as a linear combination of basis functions [1], which increases the number of unknown parameters and may lead to non-convex optimization problems. Although the assumption of stationarity may not be suitable for some application problems, it is often necessary as there are usually not enough data to compute accurate estimates of non-stationary mean and covariance functions. Progress have been made at incorporating physical knowledge into kernels, e.g., [42, 18, 6, 7, 38, 39] by computing kernels for systems governed by linear and weakly nonlinear (allowing accurate linearization) ordinary and partial differential equations. Such kernels are computed by substituting a GPR approximation of the system’s state variables into the governing equation and obtaining a system of equations for the kernel hyperparameters. For complex linear systems, computing the kernel in such a way can become prohibitively expensive, and for strongly nonlinear systems, it may not be possible at all.

In our previous work [52], we proposed the physics-informed Kriging method (PhIK) that incorporates (partial) physical knowledge into GPRs. In modeling complex systems, it is common to treat unknown parameters and fields as random parameters and fields, and the resulting realizations of the state of the system are employed to study the uncertainty of the model or the system. The goal of PhIK is to exploit the information of the system provided by these realizations to assimilate observations. In PhIK, such random realizations are used to compute the prior mean and covariance. A similar idea is used in the ensemble Kalman filter (EnKF) [13] and the formula of the “filtering step” is equivalent to the PhIK prediction Eq. (2.13). Whereas EnKF introduces uncertainty mainly from the observation noise and evolves an ensemble of state variables drawn from the posterior distribution of the previous time step, PhIK utilizes the stochasticity in models and directly uses simulation outputs for prediction without redrawing the ensemble in each time step. Not only does PhIK provide prediction or reconstruction in the form of posterior mean, it also performs *uncertainty reduction* (UR) in the form of posterior variance. More importantly, PhIK posterior mean satisfies linear physical constraints with a bounded error [52], which is critical for guaranteeing the predictive value of the method. The main drawback of PhIK is that it is highly dependent on the physical model, because the prior mean and covariance are determined entirely by the model and are not informed by data. Therefore, convergence of PhIK to the true solution with the increasing number of available observations is slower than in the data-driven GPR if the physical model is incorrect.

In this work, we propose a physics-informed CoKriging (CoPhIK) method, an extension of the CoKriging-based multifidelity framework [20, 15] to physics-informed Kriging. In this context, the direct observations of a physical system are considered as high-fidelity data and the stochastic physical model outputs are treated as low-fidelity data. CoPhIK uses PhIK to construct a GP Y_L that regresses low-fidelity data, and uses another parameterized GP Y_d to model the discrepancy between low- and high-fidelity data by assuming a specific kernel; then it infers hyperparameters of the GP model for Y_d via optimization. Subsequently, CoPhIK uses a linear combination of Y_L and Y_d to represent high-fidelity data. The mean and covariance in CoPhIK integrate physical model outputs and observation data; therefore, CoPhIK is expected

to have better accuracy than PhIK in some applications (e.g., the first two numerical examples in Section 3). On the other hand, due to the introduction of the GP Y_d , CoPhIK may lose some capacity for satisfying physical constraints with respect to PhIK, as will be shown in the error estimate provided by Theorem 2.2.

This work is organized as follows: Section 2 summarizes the GPR framework and physics-informed Kriging (Sections 2.1 to 2.3), and introduces the CoPhIK method (Section 2.4). Section 3 provides three numerical examples to demonstrate the efficiency of the proposed method. Conclusions are presented in Section 4.

2 Methodology

We begin this section by reviewing the general GPR framework [48], the ordinary Kriging method based on the assumption of stationary GP [14], and the PhIK method [52]. Then, we introduce the modified PhIK and CoPhIK methods.

2.1 GPR framework

We consider the spatial dependence of a scalar state of a physical system. Let $\mathbb{D} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}$, be the spatial domain, $y : \mathbb{D} \rightarrow \mathbb{R}$ denote the state of interest, and let $y^{(1)}, y^{(2)}, \dots, y^{(N)}$, denote N observations of y collected at the observation locations $\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^N$, where $\mathbf{x}^{(i)} \in \mathbb{D} \subseteq \mathbb{R}^d$, $y^{(i)} \in \mathbb{R}$. The observations are arranged into the observation vector $\mathbf{y} = (y^{(1)}, y^{(2)}, \dots, y^{(N)})^\top$. We aim to predict y at any new location $\mathbf{x}^* \in \mathbb{D}$. The GPR method assumes that the observation vector \mathbf{y} is a realization of the N -dimensional random vector with multivariate Gaussian distribution

$$\mathbf{Y} = \left(Y(\mathbf{x}^{(1)}, \omega), Y(\mathbf{x}^{(2)}, \omega), \dots, Y(\mathbf{x}^{(N)}, \omega) \right)^\top,$$

where $Y(\cdot, \cdot) : \mathbb{D} \times \Omega \rightarrow \mathbb{R}$ is a GP defined on the probability space (Ω, \mathcal{F}, P) . Of note, the observation coordinates $\mathbf{x}^{(i)}$ can be considered as parameters for the GP Y such that $Y(\mathbf{x}^{(i)}, \cdot)$ is a Gaussian random variable for any $\mathbf{x}^{(i)} \in \mathbb{D}$. For brevity, we denote $Y(\mathbf{x}, \cdot)$ by $Y(\mathbf{x})$. The GP Y is usually represented using GP notation as

$$Y(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) , \quad (2.1)$$

where $\mu(\cdot) : \mathbb{D} \rightarrow \mathbb{R}$ and $k(\cdot, \cdot) : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$ are the mean and covariance functions

$$\mu(\mathbf{x}) = \mathbb{E}\{Y(\mathbf{x})\} , \quad (2.2)$$

$$k(\mathbf{x}, \mathbf{x}') = \text{Cov}\{Y(\mathbf{x}), Y(\mathbf{x}')\} = \mathbb{E}\{[Y(\mathbf{x}) - \mu(\mathbf{x})][Y(\mathbf{x}') - \mu(\mathbf{x}')]\} . \quad (2.3)$$

The variance of $Y(\mathbf{x})$ is $k(\mathbf{x}, \mathbf{x})$, and its standard deviation is $\sigma(\mathbf{x}) = \sqrt{k(\mathbf{x}, \mathbf{x})}$. The covariance matrix of the random vector \mathbf{Y} is then given by

$$\mathbf{C} = \begin{pmatrix} k(\mathbf{x}^{(1)}, \mathbf{x}^{(1)}) & \cdots & k(\mathbf{x}^{(1)}, \mathbf{x}^{(N)}) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}^{(N)}, \mathbf{x}^{(1)}) & \cdots & k(\mathbf{x}^{(N)}, \mathbf{x}^{(N)}) \end{pmatrix} . \quad (2.4)$$

When the functions $\mu(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ are parameterized, their hyperparameters are identified by maximizing the log marginal likelihood of the observations (see examples in Section 2.2) [48]

$$\ln L = -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu}) - \frac{1}{2} \ln |\mathbf{C}| - \frac{N}{2} \ln 2\pi , \quad (2.5)$$

where $\boldsymbol{\mu} = (\mu(\mathbf{x}^{(1)}), \dots, \mu(\mathbf{x}^{(N)}))^\top$.

The GPR prediction at \mathbf{x}^* consists of the posterior distribution $y(\mathbf{x}^*) \sim \mathcal{N}(\hat{y}(\mathbf{x}^*), \hat{s}^2(\mathbf{x}^*))$, with posterior mean and variance given by

$$\hat{y}(\mathbf{x}^*) = \mu(\mathbf{x}^*) + \mathbf{c}(\mathbf{x}^*)^\top \mathbf{C}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \quad (2.6)$$

$$\hat{s}^2(\mathbf{x}^*) = \sigma^2(\mathbf{x}^*) - \mathbf{c}(\mathbf{x}^*)^\top \mathbf{C}^{-1} \mathbf{c}(\mathbf{x}^*), \quad (2.7)$$

and $\mathbf{c}(\mathbf{x}^*)$ is the vector of covariances

$$\mathbf{c}(\mathbf{x}^*) = \left(k(\mathbf{x}^{(1)}, \mathbf{x}^*), k(\mathbf{x}^{(2)}, \mathbf{x}^*), \dots, k(\mathbf{x}^{(N)}, \mathbf{x}^*) \right)^\top. \quad (2.8)$$

In practice, it is common to employ the posterior mean $\hat{y}(\mathbf{x}^*)$ as the prediction. The variance $\hat{s}^2(\mathbf{x}^*)$ is often called the mean squared error (MSE) of the prediction because $\hat{s}^2(\mathbf{x}^*) = \mathbb{E}\{(\hat{y}(\mathbf{x}^*) - Y(\mathbf{x}^*))^2\}$ [14]. Consequently, $\hat{s}(\mathbf{x}^*)$ is called the root mean squared error (RMSE).

To account for observation noise, one can model the noises as independent and identically distributed (iid) Gaussian random variables with zero mean and variance δ^2 , and replace \mathbf{C} in Eqs. 2.5 to 2.8 with $\mathbf{C} + \delta^2 \mathbf{I}$. In this study, we assume that observations of \mathbf{y} are noiseless. If \mathbf{C} is not invertible or its condition number is very large, one can add a small regularization term $\alpha \mathbf{I}$, where α is a small positive real number, to \mathbf{C} such that it becomes full rank. Adding the regularization term is equivalent to assuming there is iid observation noise with variance α .

2.2 Stationary GPR

In the widely used ordinary Kriging method, a stationary GP is assumed. Specifically, μ is set as a constant $\mu(\mathbf{x}) \equiv \mu$, and $k(\mathbf{x}, \mathbf{x}') = k(\boldsymbol{\tau})$, where $\boldsymbol{\tau} = \mathbf{x} - \mathbf{x}'$. Consequently, $\sigma^2(\mathbf{x}) = k(\mathbf{x}, \mathbf{x}) = k(\mathbf{0}) = \sigma^2$ is a constant. Popular forms of kernels include polynomial, exponential, Gaussian (squared-exponential), and Matérn functions. For example, the Gaussian kernel can be written as $k(\boldsymbol{\tau}) = \sigma^2 \exp(-\frac{1}{2}\|\boldsymbol{\tau}\|_w^2)$, where the weighted norm is defined as $\|\boldsymbol{\tau}\|_w^2 = \sum_{i=1}^d (\tau_i/l_i)^2$. The constant σ and the correlation lengths along each direction, $l_i \in \mathbb{R}$, $i = 1, \dots, d$ are the hyperparameters of the Gaussian kernel.

For the stationary kernel, the covariance matrix of observations, \mathbf{C} , can be written as $\mathbf{C} = \sigma^2 \boldsymbol{\Psi}$, where $\boldsymbol{\Psi}_{ij} = \exp(-\frac{1}{2}\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_w^2)$ for the Gaussian kernel. In the maximum likelihood (MLE) framework, the estimators of μ and σ^2 , denoted as $\hat{\mu}$ and $\hat{\sigma}^2$, are

$$\hat{\mu} = \frac{\mathbf{1}^\top \boldsymbol{\Psi}^{-1} \mathbf{y}}{\mathbf{1}^\top \boldsymbol{\Psi}^{-1} \mathbf{1}}, \quad \hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{1}\hat{\mu})^\top \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu})}{N}, \quad (2.9)$$

where $\mathbf{1}$ is a vector of 1s. The hyperparameters l_i are estimated by maximizing the log marginal likelihood, Eq. (2.5). The prediction of y at location \mathbf{x}^* is

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \boldsymbol{\psi}^\top \boldsymbol{\Psi}^{-1} (\mathbf{y} - \mathbf{1}\hat{\mu}), \quad (2.10)$$

where $\boldsymbol{\psi}$ is a vector of correlations between the observed data and the prediction, given by

$$\boldsymbol{\psi} = \boldsymbol{\psi}(\mathbf{x}^*) = \frac{1}{\sigma^2} \left(k(\mathbf{x}^{(1)} - \mathbf{x}^*), \dots, k(\mathbf{x}^{(N)} - \mathbf{x}^*) \right)^\top,$$

and the MSE of the prediction is

$$\hat{s}^2(\mathbf{x}^*) = \hat{\sigma}^2 \left(1 - \boldsymbol{\psi}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\psi} \right). \quad (2.11)$$

A more general approach to GPR is to employ parameterized nonstationary covariance kernels. Nonstationary kernels can be obtained by modifying stationary covariance kernels,

e.g., [41, 27, 32, 37, 4, 28], or from neural networks with specific activation functions, e.g., [29, 34], among other approaches. Many of these approaches assume a specific functional form for the correlation function, chosen according to expert knowledge. The key computational challenge in these data-driven GPR is the optimization step of maximizing the (log marginal) likelihood. In many practical cases, this is a non-convex optimization problem, and the condition number of \mathbf{C} or Ψ can be quite large. Another fundamental challenge is that parameterized models for mean and covariance usually don't account for physical constraints, and therefore require a large amount of data to accurately model physics.

2.3 PhIK

The recently proposed PhIK method [52] takes advantage of existing expert knowledge in the form of stochastic physics-based models. These stochastic models for physical systems include random parameters or random fields to reflect the lack of understanding (of physical laws) or knowledge (of the coefficients, parameters, etc.) of the real system. Monte Carlo (MC) simulations of the stochastic physical model can be conducted to generate an ensemble of state variables, from which the mean and covariance are estimated. This mean and covariance estimates are then employed to construct a GP model of the state variables. As such, there is no need to assume a specific parameterized covariance kernel or solve an optimization problem for the hyperparameters of the kernel.

Given M realizations of a stochastic model $u(\mathbf{x}; \omega)$, $\mathbf{x} \in \mathbb{D}$, $\omega \in \Omega$, denoted as $\{Y^m(\mathbf{x})\}_{m=1}^M$, we build the following GP model:

$$Y(\mathbf{x}) \sim \mathcal{GP}(\mu_{\text{MC}}(\mathbf{x}), k_{\text{MC}}(\mathbf{x}, \mathbf{x}')), \quad (2.12)$$

where μ_{MC} and k_{MC} are the ensemble mean and covariance functions

$$\begin{aligned} \mu(\mathbf{x}) &\approx \mu_{\text{MC}}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M Y^m(\mathbf{x}), \\ k(\mathbf{x}, \mathbf{x}') &\approx k_{\text{MC}}(\mathbf{x}, \mathbf{x}') = \frac{1}{M-1} \sum_{m=1}^M (Y^m(\mathbf{x}) - \mu_{\text{MC}}(\mathbf{x})) (Y^m(\mathbf{x}') - \mu_{\text{MC}}(\mathbf{x}')). \end{aligned} \quad (2.13)$$

The covariance matrix of observations can be estimated as

$$\mathbf{C} \approx \mathbf{C}_{\text{MC}} = \frac{1}{M-1} \sum_{m=1}^M (\mathbf{Y}^m - \boldsymbol{\mu}_{\text{MC}}) (\mathbf{Y}^m - \boldsymbol{\mu}_{\text{MC}})^{\top}, \quad (2.14)$$

where $\mathbf{Y}^m = (Y^m(\mathbf{x}^{(1)}), \dots, Y^m(\mathbf{x}^{(N)}))^{\top}$ and $\boldsymbol{\mu}_{\text{MC}} = (\mu_{\text{MC}}(\mathbf{x}^{(1)}), \dots, \mu_{\text{MC}}(\mathbf{x}^{(N)}))^{\top}$. The prediction and MSE at location $\mathbf{x}^* \in \mathbb{D}$ are

$$\hat{y}(\mathbf{x}^*) = \mu_{\text{MC}}(\mathbf{x}^*) + \mathbf{c}_{\text{MC}}(\mathbf{x}^*)^{\top} \mathbf{C}_{\text{MC}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\text{MC}}), \quad (2.15)$$

$$\hat{s}^2(\mathbf{x}^*) = \hat{\sigma}_{\text{MC}}^2(\mathbf{x}^*) - \mathbf{c}_{\text{MC}}(\mathbf{x}^*)^{\top} \mathbf{C}_{\text{MC}}^{-1} \mathbf{c}_{\text{MC}}(\mathbf{x}^*), \quad (2.16)$$

where $\hat{\sigma}_{\text{MC}}^2(\mathbf{x}^*) = k_{\text{MC}}(\mathbf{x}^*, \mathbf{x}^*)$ is the variance of the set $\{Y^m(\mathbf{x}^*)\}_{m=1}^M$, and

$$\mathbf{c}_{\text{MC}}(\mathbf{x}^*) = (k_{\text{MC}}(\mathbf{x}^{(1)}, \mathbf{x}^*), \dots, k_{\text{MC}}(\mathbf{x}^{(N)}, \mathbf{x}^*))^{\top}.$$

It was demonstrated in [52] that PhIK predictions satisfy *linear* physical constraints up to an error bound that depends on the numerical error, the discrepancy between the physical model and real system, and the smallest eigenvalue of matrix \mathbf{C} . Linear physical constraints include periodic, Dirichlet or Neumann boundary condition, and linear equation $\mathcal{L}u = g$, where

\mathcal{L} is a linear differential or integral operator. For example, let $u(\mathbf{x}; \omega)$ be a stochastic model of the velocity potential for an incompressible flow, i.e., $\nabla \cdot (\nabla u(\mathbf{x}; \omega)) = 0$; then PhIK guarantees that $\nabla \hat{y}(\mathbf{x})$ is a divergence-free field.

In PhIK, MC simulation of the stochastic physical model for computing μ_{MC} and k_{MC} can be replaced by more efficient approaches, such as quasi-Monte Carlo [30], multi-level Monte Carlo (MLMC) [17], probabilistic collocation [49], Analysis Of Variance (ANOVA) [50], compressive sensing [51], the moment equation and PDF methods [45, 3], and the bi-fidelity method [53]. Linear physical constraints are preserved if $\mu_{\text{MC}}(\mathbf{x})$ and $k_{\text{MC}}(\mathbf{x}, \mathbf{x}')$ are computed using a *linear combination* of the realizations $\{Y^m(\mathbf{x})\}_{m=1}^M$. As an example, we present the MLMC-based PhIK [52] in Appendix A.

Further, the matrix \mathbf{C} and vector $\boldsymbol{\mu}$ are fixed in Eq. (2.5) for a given ensemble $\{Y^m\}_{m=1}^M$. Thus, the log marginal likelihood is fixed. We can modify PhIK by adding a correction term to $\mu_{\text{MC}}(\mathbf{x})$ to increase the likelihood. Specifically, we replace $\mu_{\text{MC}}(\mathbf{x})$ by $\mu_{\text{MC}}(\mathbf{x}) + \Delta\mu$, where $\Delta\mu$ is a constant. Then, taking the derivative of $\ln L$ with respect to $\Delta\mu$ and setting it to be zero yields

$$\Delta\mu = \frac{\mathbf{1}^\top \boldsymbol{\Psi}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{\mathbf{1}^\top \boldsymbol{\Psi}^{-1}\mathbf{1}}. \quad (2.17)$$

This modification has a potential to increase the accuracy of the prediction, but it may also violate some physical constraints, e.g., the Dirichlet boundary condition. We name this method *modified PhIK*, and provide the following theorem on how well it preserves linear physical constraints.

Theorem 2.1. *Assume that a stochastic model $u(\mathbf{x}; \omega)$ defined on $\mathbb{D} \times \Omega$ ($\mathbb{D} \subseteq \mathbb{R}^d$) satisfies $\|\mathcal{L}u(\mathbf{x}; \omega) - g(\mathbf{x}; \omega)\| \leq \epsilon$ for any $\omega \in \Omega$, where \mathcal{L} is a deterministic bounded linear operator, $g(\mathbf{x}; \omega)$ is a well-defined function on $\mathbb{R}^d \times \Omega$, and $\|\cdot\|$ is a well-defined function norm. $\{Y^m(\mathbf{x})\}_{m=1}^M$ are a finite number of realizations of $u(\mathbf{x}; \omega)$, i.e., $Y^m(\mathbf{x}) = u(\mathbf{x}; \omega^m)$. Then, the prediction $\hat{y}(\mathbf{x})$ from modified PhIK satisfies*

$$\begin{aligned} \left\| \mathcal{L}\hat{y}(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| &\leq \epsilon + \left[2\epsilon \sqrt{\frac{M}{M-1}} + \sigma(g(\mathbf{x}; \omega^m)) \right] \cdot \\ &\quad \left\| \mathbf{C}_{\text{MC}}^{-1} \right\|_2 \left\| \mathbf{y} - \boldsymbol{\mu}_{\text{MC}} - \Delta\mu \mathbf{1} \right\|_2 \sum_{i=1}^N \sigma(Y^m(\mathbf{x}^{(i)})) + \left\| \mathcal{L}\Delta\mu \right\|, \end{aligned} \quad (2.18)$$

where $\sigma(Y^m(\mathbf{x}^{(i)}))$ is the standard deviation of the data set $\{Y^m(\mathbf{x}^{(i)})\}_{m=1}^M$ for fixed $\mathbf{x}^{(i)}$, $\overline{g(\mathbf{x})} = \frac{1}{M} \sum_{m=1}^M g(\mathbf{x}; \omega^m)$, and $\sigma(g(\mathbf{x}; \omega^m)) = \left(\frac{1}{M-1} \sum_{m=1}^M \|g(\mathbf{x}; \omega^m) - \overline{g(\mathbf{x})}\|^2 \right)^{\frac{1}{2}}$.

Proof. The modified PhIK prediction can be written as

$$\hat{y}(\mathbf{x}) = \mu_{\text{MC}}(\mathbf{x}) + \Delta\mu + \sum_{i=1}^N \tilde{a}_i k_{\text{MC}}(\mathbf{x}, \mathbf{x}^{(i)}), \quad (2.19)$$

where \tilde{a}_i is the i -th entry of $\mathbf{C}_{\text{MC}}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\text{MC}} - \Delta\mu \mathbf{1})$. According to Theorem 2.1 and Corollary

2.2 in [52],

$$\begin{aligned}
\left\| \mathcal{L}\hat{y}(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| &\leq \left\| \mathcal{L}\left(\mu_{\text{MC}}(\mathbf{x}) + \sum_{i=1}^N \tilde{a}_i k_{\text{MC}}(\mathbf{x}, \mathbf{x}^{(i)}) \right) - \overline{g(\mathbf{x})} \right\| + \|\mathcal{L}\Delta\mu\| \\
&\leq \epsilon + \left[2\epsilon \sqrt{\frac{M}{M-1}} + \sigma(g(\mathbf{x}; \omega^m)) \right] \left\| \mathbf{C}_{\text{MC}}^{-1} \right\|_2 \left\| \mathbf{y} - \boldsymbol{\mu}_{\text{MC}} - \Delta\mu \mathbf{1} \right\|_2 \sum_{i=1}^N \sigma(Y^m(\mathbf{x}^{(i)})) \\
&\quad + \|\mathcal{L}\Delta\mu\|.
\end{aligned}$$

□

For $\Delta\mu = 0$, the bound (2.18) reverts to the bound in [52] and the modified PhIK method reverts to PhIK. In some cases, the term $\|\mathcal{L}\Delta\mu\| = 0$, e.g., when \mathcal{L} is a differential operator such as the Neumann boundary condition operator.

2.4 CoPhIK

CoKriging was originally formulated to compute predictions of sparsely observed states of physical systems by leveraging observations of other states or parameters of the system [43, 22]. Recently, it has been employed for constructing multi-fidelity models [20, 25, 35], and has been applied in various areas, e.g., [24, 5, 33]. In this work, we propose a novel multi-fidelity method, CoPhIK, that integrates PhIK and CoKriging by combining numerical simulations and high-fidelity observations. Our multi-fidelity method is based on Kennedy and O'Hagan's CoKriging framework presented in [20, 14].

We briefly review the formulation of CoKriging for two-level multi-fidelity modeling in [15]. Suppose that we have high-fidelity data (e.g., accurate measurements of states) $\mathbf{y}_H = (y_H^{(1)}, \dots, y_H^{(N_H)})^\top$ at locations $\mathbf{X}_H = \{\mathbf{x}_H^{(i)}\}_{i=1}^{N_H}$, and low-fidelity data (e.g., simulation results) $\mathbf{y}_L = (y_L^{(1)}, \dots, y_L^{(N_L)})^\top$ at locations $\mathbf{X}_L = \{\mathbf{x}_L^{(i)}\}_{i=1}^{N_L}$, where $y_H^{(i)}, y_L^{(i)} \in \mathbb{R}$ and $\mathbf{x}_H^{(i)}, \mathbf{x}_L^{(i)} \in \mathbb{D} \subseteq \mathbb{R}^d$. By concatenating the observation locations and data respectively, i.e., $\tilde{\mathbf{X}} = \{\mathbf{X}_L, \mathbf{X}_H\}$ and $\tilde{\mathbf{y}} = (\mathbf{y}_L^\top, \mathbf{y}_H^\top)^\top$, we can construct a *multivariate* GP via Kriging as detailed in [2]. Kennedy and O'Hagan proposed an alternative formulation of CoKriging based on the auto-regressive model for Y_H

$$Y_H(\mathbf{x}) = \rho Y_L(\mathbf{x}) + Y_d(\mathbf{x}), \quad (2.20)$$

where $\rho \in \mathbb{R}$ is a regression parameter and $Y_d(\mathbf{x})$ is a GP that models the difference between Y_H and ρY_L . This model assumes that

$$\text{Cov}\{Y_H(\mathbf{x}), Y_L(\mathbf{x}') \mid Y_L(\mathbf{x})\} = 0, \text{ for all } \mathbf{x}' \neq \mathbf{x}, \mathbf{x}, \mathbf{x}' \in \mathbb{D}. \quad (2.21)$$

It was shown in [31] that the assumption of Eq. (2.21) implies the auto-regressive model of Eq. (2.20). The covariance of observations, $\tilde{\mathbf{C}}$, is then given by

$$\tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{C}_L(\mathbf{X}_L, \mathbf{X}_L) & \rho \mathbf{C}_L(\mathbf{X}_L, \mathbf{X}_H) \\ \rho \mathbf{C}_L(\mathbf{X}_H, \mathbf{X}_L) & \rho^2 \mathbf{C}_L(\mathbf{X}_H, \mathbf{X}_H) + \mathbf{C}_d(\mathbf{X}_H, \mathbf{X}_H) \end{pmatrix} \quad (2.22)$$

where \mathbf{C}_L is the covariance matrix based on GP Y_L 's kernel $k_L(\cdot, \cdot)$, and \mathbf{C}_d is the covariance matrix based on GP Y_d 's kernel $k_d(\cdot, \cdot)$. One can assume parameterized forms for these kernels (e.g., Gaussian kernel) and then simultaneously identify their hyperparameters along with ρ by maximizing the following log marginal likelihood:

$$\ln \tilde{L} = -\frac{1}{2}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}})^\top \tilde{\mathbf{C}}^{-1}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}) - \frac{1}{2} \ln |\tilde{\mathbf{C}}| - \frac{N_H + N_L}{2} \ln 2\pi. \quad (2.23)$$

Alternatively, one can employ the following two-step approach [15, 14]:

1. Use Kriging to construct Y_L using $\{\mathbf{X}_L, \mathbf{y}_L\}$.
2. Denote $\mathbf{y}_d = \mathbf{y}_H - \rho \mathbf{y}_L(\mathbf{X}_H)$, where $\mathbf{y}_L(\mathbf{X}_H)$ are the values of \mathbf{y}_L at locations common to those of \mathbf{X}_H , then construct Y_d using $\{\mathbf{X}_H, \mathbf{y}_d\}$ via Kriging.

The posterior mean and variance of Y_H at $\mathbf{x}^* \in \mathbb{D}$ are given by

$$\hat{y}(\mathbf{x}^*) = \mu_H(\mathbf{x}^*) + \tilde{\mathbf{c}}(\mathbf{x}^*)^\top \tilde{\mathbf{C}}^{-1}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}), \quad (2.24)$$

$$\hat{s}^2(\mathbf{x}^*) = \rho^2 \sigma_L^2(\mathbf{x}^*) + \sigma_d^2(\mathbf{x}^*) - \tilde{\mathbf{c}}(\mathbf{x}^*)^\top \tilde{\mathbf{C}}^{-1} \tilde{\mathbf{c}}(\mathbf{x}^*), \quad (2.25)$$

where $\mu_H(\mathbf{x}^*) = \rho \mu_L(\mathbf{x}^*) + \mu_d(\mathbf{x}^*)$, $\mu_L(\mathbf{x})$ is the mean of $Y_L(\mathbf{x})$, $\mu_d(\mathbf{x})$ is the mean of $Y_d(\mathbf{x})$, $\sigma_L^2(\mathbf{x}^*) = k_L(\mathbf{x}^*, \mathbf{x}^*)$, $\sigma_d^2(\mathbf{x}^*) = k_d(\mathbf{x}^*, \mathbf{x}^*)$, and

$$\tilde{\boldsymbol{\mu}} = \begin{pmatrix} \boldsymbol{\mu}_L \\ \boldsymbol{\mu}_H \end{pmatrix} = \begin{pmatrix} (\mu_L(\mathbf{x}_L^{(1)}), \dots, \mu_L(\mathbf{x}_L^{(N_L)}))^\top \\ (\mu_H(\mathbf{x}_H^{(1)}), \dots, \mu_H(\mathbf{x}_H^{(N_H)}))^\top \end{pmatrix}, \quad (2.26)$$

$$\tilde{\mathbf{c}}(\mathbf{x}^*) = \begin{pmatrix} \rho \mathbf{c}_L(\mathbf{x}^*) \\ \mathbf{c}_H(\mathbf{x}^*) \end{pmatrix} = \begin{pmatrix} (\rho k_L(\mathbf{x}^*, \mathbf{x}_L^{(1)}), \dots, \rho k_L(\mathbf{x}^*, \mathbf{x}_L^{(N_L)}))^\top \\ (k_H(\mathbf{x}^*, \mathbf{x}_H^{(1)}), \dots, k_H(\mathbf{x}^*, \mathbf{x}_H^{(N_H)}))^\top \end{pmatrix}, \quad (2.27)$$

where $k_H(\mathbf{x}, \mathbf{x}') = \rho^2 k_L(\mathbf{x}, \mathbf{x}') + k_d(\mathbf{x}, \mathbf{x}')$. Here, we have neglected a small contribution to \hat{s}^2 (see [14]).

Now we describe the CoPhIK method. We set $\mathbf{X}_L = \mathbf{X}_H$ to simplify the formula and computing, and denote $N = N_H = N_L$. We employ PhIK to construct GP Y_L using realizations $\{Y^m(\mathbf{x})\}_{m=1}^M$ of a stochastic model $u(\mathbf{x}; \omega)$ on $\mathbb{D} \times \Omega$. Specifically, we set $\mu_L(\mathbf{x}) = \mu_{MC}(\mathbf{x})$ and $k_L(\mathbf{x}, \mathbf{x}') = k_{MC}(\mathbf{x}, \mathbf{x}')$, where μ_{MC} and k_{MC} are given by Eq. (2.13). The GP model Y_d is constructed using the same approach as in the second step of the Kennedy and O'Hagan CoKriging framework. In other words, CoPhIK replaces the first step of their framework with PhIK, and follows the same procedure for the second step.

We proceed to describe the construction of Y_d in more detail. First, we set $\mathbf{y}_d = \mathbf{y}_H - \rho \mu_L(\mathbf{X}_L)$. The reason for this choice is that $\mu_L(\mathbf{X}_H)$ is the most probable observation of the GP Y_L . Next, we need to assume a specific form of the kernel function. Without loss of generality, in the following theoretical analysis and computational examples, we use the stationary Gaussian kernel model and constant μ_d . Once \mathbf{y}_d is computed, and the form of $\mu_d(\cdot)$ and $k_d(\cdot, \cdot)$ are decided, Y_d can be constructed as in ordinary Kriging. Now that all components in $\ln \tilde{L}$ are specified except for the \mathbf{y}_L in $\tilde{\mathbf{y}}$. We set \mathbf{y}_L as the realization from the ensemble $\{Y^m\}_{m=1}^M$ that maximizes $\ln \tilde{L}$. The algorithm is summarized in Algorithm 1.

Next, we analyze the form of $\tilde{\mathbf{C}}$. Recalling the choice $\mathbf{X}_L = \mathbf{X}_H$ and introducing the notation $\mathbf{C}_1 = C_L(\mathbf{X}_L, \mathbf{X}_L)$ and $\mathbf{C}_2 = C_d(\mathbf{X}_H, \mathbf{X}_H)$ in Eq. (2.22), we can write the inverse of $\tilde{\mathbf{C}}$ as

$$\tilde{\mathbf{C}}^{-1} = \begin{pmatrix} \mathbf{C}_1^{-1} + \rho^2 \mathbf{C}_2^{-1} & -\rho \mathbf{C}_2^{-1} \\ -\rho \mathbf{C}_2^{-1} & \mathbf{C}_2^{-1} \end{pmatrix}. \quad (2.28)$$

Thus,

$$\begin{aligned} \tilde{\mathbf{C}}^{-1}(\tilde{\mathbf{y}} - \tilde{\boldsymbol{\mu}}) &= \begin{pmatrix} \mathbf{C}_1^{-1} + \rho^2 \mathbf{C}_2^{-1} & -\rho \mathbf{C}_2^{-1} \\ -\rho \mathbf{C}_2^{-1} & \mathbf{C}_2^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{y}_L - \boldsymbol{\mu}_L \\ \mathbf{y}_H - \boldsymbol{\mu}_H \end{pmatrix} \\ &= \begin{pmatrix} (\mathbf{C}_1^{-1} + \rho^2 \mathbf{C}_2^{-1})(\mathbf{y}_L - \boldsymbol{\mu}_L) - \rho \mathbf{C}_2^{-1}(\mathbf{y}_H - \boldsymbol{\mu}_H) \\ -\rho \mathbf{C}_2^{-1}(\mathbf{y}_L - \boldsymbol{\mu}_L) + \mathbf{C}_2^{-1}(\mathbf{y}_H - \boldsymbol{\mu}_H) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{C}_1^{-1}(\mathbf{y}_L - \boldsymbol{\mu}_L) - \rho \mathbf{C}_2^{-1}((\mathbf{y}_H - \boldsymbol{\mu}_H) - \rho(\mathbf{y}_L - \boldsymbol{\mu}_L)) \\ \mathbf{C}_2^{-1}((\mathbf{y}_H - \boldsymbol{\mu}_H) - \rho(\mathbf{y}_L - \boldsymbol{\mu}_L)) \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{C}_1^{-1}(\mathbf{y}_L - \boldsymbol{\mu}_L) - \rho \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho \mathbf{y}_L - \mathbf{1}\mu_d) \\ \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho \mathbf{y}_L - \mathbf{1}\mu_d) \end{pmatrix}, \end{aligned} \quad (2.29)$$

Algorithm 1 CoPhIK using stochastic simulation model $u(\mathbf{x}; \omega)$ on $\mathbb{D} \times \Omega$ ($\mathbb{D} \subseteq \mathbb{R}^d$), and high-fidelity observation $\mathbf{y}_H = (y_H^{(1)}, \dots, y_H^{(N)})^\top$ at locations $\mathbf{X}_H = \{\mathbf{x}_H^{(i)}\}_{i=1}^N$.

- 1: Conduct stochastic simulation, e.g., MC simulation, using $u(\mathbf{x}; \omega)$ to generate realizations $\{Y^m\}_{m=1}^M$ on the entire domain \mathbb{D} .
 - 2: Use PhIK to construct GP Y_L on $D \times \Omega$, i.e., $\mu_L(\cdot) = \mu_{MC}(\cdot)$ and $k_L(\cdot, \cdot) = k_{MC}(\cdot, \cdot)$ in Eq. (2.13). Compute $\mu_L(\mathbf{X}_L) = (\mu_L(\mathbf{x}_L^{(1)}), \dots, \mu_L(\mathbf{x}_L^{(N)}))^\top$, and $\mathbf{C}_L(\mathbf{X}_L, \mathbf{X}_L)$ whose ij -th element is $k_L(\mathbf{x}_H^{(i)}, \mathbf{x}_H^{(j)})$. Set $\mathbf{C}_L(\mathbf{X}_L, \mathbf{X}_H) = \mathbf{C}_L(\mathbf{X}_H, \mathbf{X}_L) = \mathbf{C}_L(\mathbf{X}_H, \mathbf{X}_H) = \mathbf{C}_L(\mathbf{X}_L, \mathbf{X}_L)$ (because $\mathbf{X}_L = \mathbf{X}_H$).
 - 3: Denote $\mathbf{y}_d = \mathbf{y}_H - \rho\mu_L(\mathbf{X}_L)$, choose a specific kernel function $k_d(\cdot, \cdot)$ (Gaussian kernel in this work) for the GP Y_d , and identify hyperparameters via maximizing the log marginal likelihood Eq. (2.5), where $\mathbf{y}, \boldsymbol{\mu}, \mathbf{C}$ are specified as $\mathbf{y}_d, \boldsymbol{\mu}_d, \mathbf{C}_d$, respectively. Then construct $\tilde{\boldsymbol{\mu}}$ in Eq. (2.26), and \mathbf{C}_d whose ij -th element is $k_d(\mathbf{x}_H^{(i)}, \mathbf{x}_H^{(j)})$.
 - 4: Iterate over the set $\{Y^m\}_{m=1}^M$ to identify Y^m that maximizes $\ln \tilde{L}$ in Eq. (2.23), where $\mathbf{y}_L = (Y^m(\mathbf{x}_H^{(1)}), \dots, Y^m(\mathbf{x}_H^{(N)}))^\top$ is used in $\tilde{\mathbf{y}}$.
 - 5: Compute the posterior mean using Eq. (2.24), and variance using Eq. (2.25) for any $\mathbf{x}^* \in \mathbb{D}$.
-

where $\boldsymbol{\mu}_d = (\mu_d(\mathbf{x}_H^{(1)}), \dots, \mu_d(\mathbf{x}_H^{(N)}))^\top = \mathbf{1}\mu_d$. Therefore, the posterior mean at $\mathbf{x}^* \in \mathbb{D}$, given by Eq. (2.24), can be rewritten as

$$\begin{aligned} \hat{y}(\mathbf{x}^*) &= \mu_H(\mathbf{x}^*) + (\rho\mathbf{c}_L(\mathbf{x}^*)^\top, \mathbf{c}_H(\mathbf{x}^*)^\top) \begin{pmatrix} \mathbf{C}_1^{-1}(\mathbf{y}_L - \boldsymbol{\mu}_L) - \rho\mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \\ \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \end{pmatrix} \\ &= \mu_H(\mathbf{x}^*) + \rho\mathbf{c}_L(\mathbf{x}^*)^\top \mathbf{C}_1^{-1}(\mathbf{y}_L - \boldsymbol{\mu}_L) + (\mathbf{c}_H(\mathbf{x}^*) - \rho^2\mathbf{c}_L(\mathbf{x}^*))^\top \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \\ &= \rho \left(\mu_L(\mathbf{x}^*) + \mathbf{c}_L(\mathbf{x}^*)^\top \mathbf{C}_1^{-1}(\mathbf{y}_H - \boldsymbol{\mu}_L) \right) - \rho\mathbf{c}_L(\mathbf{x}^*)^\top \mathbf{C}_1^{-1}(\mathbf{y}_H - \mathbf{y}_L) \\ &\quad + \mu_d + (\mathbf{c}_H(\mathbf{x}^*) - \rho^2\mathbf{c}_L(\mathbf{x}^*))^\top \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d), \end{aligned} \tag{2.30}$$

which indicates that the posterior mean is of the functional form

$$\hat{y}(\mathbf{x}) = \underbrace{\rho \left[\mu_L(\mathbf{x}) + \sum_{i=1}^N a_i k_L(\mathbf{x}, \mathbf{x}^{(i)}) \right]}_{S_1} - \underbrace{\rho \sum_{i=1}^N b_i k_L(\mathbf{x}, \mathbf{x}^{(i)})}_{S_2} + \underbrace{\mu_d + \sum_{i=1}^N q_i k_d(\mathbf{x}, \mathbf{x}^{(i)})}_{S_3}, \tag{2.31}$$

where a_i, b_i, q_i are the i th entries of $\mathbf{C}_1^{-1}(\mathbf{y}_H - \boldsymbol{\mu}_L), \mathbf{C}_1^{-1}(\mathbf{y}_H - \mathbf{y}_L), \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d)$, respectively. Here, S_1 is the PhIK prediction multiplied by ρ , and $(-S_2 + S_3)$ can be considered as the CoPhIK correction term. Furthermore, we note that

$$S_1 - S_2 = \rho \left[\mu_L(\mathbf{x}) + \sum_{i=1}^N (a_i - b_i) k_L(\mathbf{x}, \mathbf{x}^{(i)}) \right], \tag{2.32}$$

has the same form as the PhIK prediction multiplied by ρ . This indicates that the error bound in preserving the physical constraints for this part is similar to the PhIK's error bound. Therefore, using notations in Eq. (2.31), we extend Theorem 2.1 in [52] for CoPhIK as the following.

Theorem 2.2. Assume that a stochastic model $u(\mathbf{x}; \omega)$ defined on $\mathbb{D} \times \Omega$ ($\mathbb{D} \subseteq \mathbb{R}^d$) satisfies $\|\mathcal{L}u(\mathbf{x}; \omega) - g(\mathbf{x}; \omega)\| \leq \epsilon$ for any $\omega \in \Omega$, where \mathcal{L} is a deterministic bounded linear operator, $g(\mathbf{x}; \omega)$ is a well-defined function on $\mathbb{R}^d \times \Omega$, and $\|\cdot\|$ is a well-defined function norm. $\{Y^m(\mathbf{x})\}_{m=1}^M$ are a finite number of realizations of $u(\mathbf{x}; \omega)$, i.e., $Y^m(\mathbf{x}) = u(\mathbf{x}; \omega^m)$. Then, the

prediction $\hat{y}(\mathbf{x})$ from *CoPhIK* satisfies

$$\begin{aligned}
\left\| \mathcal{L}\hat{y}(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| &\leq \rho\epsilon + (1-\rho)\left\| \overline{g(\mathbf{x})} \right\| \\
&+ \rho \left[2\epsilon \sqrt{\frac{M}{M-1}} + \sigma(g(\mathbf{x}; \omega^m)) \right] \cdot \left\| \mathbf{C}_1^{-1} \right\|_2 \left\| \mathbf{y}_L - \boldsymbol{\mu}_L \right\|_2 \sum_{i=1}^N \sigma(Y^m(\mathbf{x}^{(i)})) \\
&+ \left\| \mathcal{L}\mu_d \right\| + \left\| \mathbf{C}_2^{-1} \right\|_2 \left\| \mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d \right\|_2 \sum_{i=1}^N \left\| \mathcal{L}k_d(\mathbf{x}, \mathbf{x}^{(i)}) \right\|,
\end{aligned} \tag{2.33}$$

where $\sigma(Y^m(\mathbf{x}^{(i)}))$, $\overline{g(\mathbf{x})}$, $\sigma(g(\mathbf{x}; \omega^m))$ are defined in Theorem 2.1.

Proof. According to Eq. (2.31) and Eq. (2.32),

$$\begin{aligned}
\left\| \mathcal{L}\hat{y}(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| &\leq \left\| \mathcal{L}(S_1 - S_2) - \overline{g(\mathbf{x})} \right\| + \left\| \mathcal{L}S_3 \right\| \\
&\leq \rho \left\| \mathcal{L}\mu_L(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| + (1-\rho) \left\| \overline{g(\mathbf{x})} \right\| + \rho \left\| \mathcal{L} \left(\sum_{i=1}^N \tilde{a}_i k_L(\mathbf{x}, \mathbf{x}^{(i)}) \right) \right\| + \left\| \mathcal{L}S_3 \right\|.
\end{aligned}$$

Following the same procedure outlined in Theorem 2.1 of [52], we have

$$\left\| \mathcal{L}\mu_L(\mathbf{x}) - \overline{g(\mathbf{x})} \right\| \leq \epsilon,$$

and

$$\left\| \mathcal{L} \left(\sum_{i=1}^N \tilde{a}_i k_L(\mathbf{x}, \mathbf{x}^{(i)}) \right) \right\| \leq \left[2\epsilon \sqrt{\frac{M}{M-1}} + \sigma(g(\mathbf{x}; \omega^m)) \right] \cdot \left\| \mathbf{C}_1^{-1} \right\|_2 \left\| \mathbf{y}_L - \boldsymbol{\mu}_L \right\|_2 \sum_{i=1}^N \sigma(Y^m(\mathbf{x}^{(i)})).$$

Moreover,

$$\left\| \mathcal{L}S_3 \right\| \leq \left\| \mathcal{L}\mu_d \right\| + \sum_{i=1}^N |q_i| \left\| \mathcal{L}k_d(\mathbf{x}, \mathbf{x}^{(i)}) \right\| \leq \left\| \mathcal{L}\mu_d \right\| + \max_i |q_i| \sum_{i=1}^N \left\| \mathcal{L}k_d(\mathbf{x}, \mathbf{x}^{(i)}) \right\|,$$

where q_i is defined in Eq. (2.31), and

$$\max_i |q_i| = \left\| \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \right\|_\infty \leq \left\| \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \right\|_2 \leq \left\| \mathbf{C}_2^{-1} \right\|_2 \left\| \mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d \right\|_2.$$

□

Similar to Theorem 2.1, the upper bound includes $\left\| \mathcal{L}\mu_d \right\|$, where μ_d is a constant. This term may disappear for some choices of \mathcal{L} , e.g., when \mathcal{L} is a derivative operator. The norm $\left\| \mathcal{L}k_d \right\|$ depends on the selection of k_d and on the properties of \mathcal{L} . Therefore, it follows that carefully selecting a kernel for Y_d according to the properties of the system would result in a smaller error in preserving the corresponding physical constraints. Of note, Kennedy and O'Hagan's framework can be improved by using a more general nonlinear form of relation between low- and high-fidelity data (see, e.g., [36]). Consequently, a nonlinear relation between low- and high-fidelity data will change the upper bound in Theorem 2.2.

Furthermore, in Step 4 of Algorithm 1, if we set $\mathbf{y}_L = \mu_L(\mathbf{X}_L)$, then the posterior mean will be

$$\begin{aligned}
\hat{y}(\mathbf{x}^*) &= \mu_H(\mathbf{x}^*) + (\rho\mathbf{c}_L(\mathbf{x}^*)^\top, \mathbf{c}_H(\mathbf{x}^*)^\top) \begin{pmatrix} -\rho\mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\boldsymbol{\mu}_L - \mathbf{1}\mu_d) \\ \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\boldsymbol{\mu}_L - \mathbf{1}\mu_d) \end{pmatrix} \\
&= \mu_H(\mathbf{x}^*) + (\mathbf{c}_H(\mathbf{x}^*) - \rho^2\mathbf{c}_L(\mathbf{x}^*))^\top \mathbf{C}_2^{-1}(\mathbf{y}_H - \rho\mathbf{y}_L - \mathbf{1}\mu_d) \\
&= \mu_H(\mathbf{x}^*) + \mathbf{c}_d(\mathbf{x}^*) \mathbf{C}_d(\mathbf{X}_H, \mathbf{X}_H)^{-1}(\mathbf{y}_d - \mathbf{1}\mu_d),
\end{aligned} \tag{2.34}$$

where $\mathbf{c}_d(\mathbf{x}^*) = \mathbf{c}_H(\mathbf{x}^*) - \rho^2 \mathbf{c}_L(\mathbf{x}^*)$. In this form, the model information is included into μ_H and \mathbf{y}_d , but not into the covariance matrix \mathbf{C}_d explicitly. The error bound of preserving linear physical constraints can be derived from Theorem 2.2 by setting $\tilde{a}_i = 0$. In practice, we can modify Step 4 of Algorithm 1 as identifying \mathbf{y}_L via iterating over the set $\{Y^m\}_{m=1}^M \cup \{\boldsymbol{\mu}_L\}$.

Finally, we summarize this section by presenting in Figure 1 a spectrum of GP-based data-driven and physics-driven methods. In general, methods closer to the physics-driven end result in predictions that better enforce physical constraints, while methods closer to the data-driven end may result in more accurate posterior mean (e.g., smaller difference between posterior mean and the ground truth). The standard kriging and cokring methods are at the data-driven end of the spectrum. PhIK, modified PhIK and CoPhIK assign different “weights” to data and model, and thus can be considered as placed along the spectrum between the data-driven and the physics-driven ends. The performance of the physics-driven methods depends on the selection of stochastic model, and so as the physics-informed methods. In general, we expect that adequate domain knowledge will lead to a good selection of stochastic models, which will result in good accuracy of physics-informed GP methods. Moreover, as we point out in the introduction, physical knowledge can also be incorporated directly into GP kernels, especially for linear or linearized deterministic systems, which is an alternative approach for data-model convergence under the GP framework.

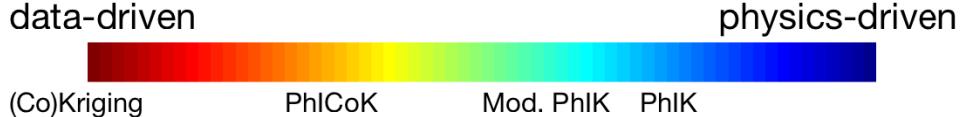


Figure 1: Spectrum of Physics-informed GP methods.

3 Numerical examples

We present three numerical examples to demonstrate the performance of CoPhIK for reconstructing spatially distributed states of physical systems from sparse observations and numerical simulations. All numerical examples are two-dimensional in physical space. We compare CoPhIK against ordinary Kriging (referred to as Kriging) and PhIK. For each example, the two-dimensional reference field and the reconstructed field (posterior mean) are discretized and presented in matrix form. We denote by \mathbf{F} the reference field in matrix form, and by \mathbf{F}_r the reconstructed field in matrix form. We employ the RMSE \hat{s} , the point-wise difference $\mathbf{F}_r - \mathbf{F}$ and the relative error $\|\mathbf{F}_r - \mathbf{F}\|_F / \|\mathbf{F}\|_F$ (where $\|\cdot\|_F$ is the Frobenius norm) to compare the performance of different methods. We use the Gaussian kernel in Kriging and CoPhIK because the fields in the examples are smooth. The kriging code employed in this work is based on the scripts in [14]. We also compare the CoPhIK, PhIK and Kriging performance for adaptively identifying new observations in order to reduce the uncertainty of the field reconstruction. For this purpose we employ the active learning algorithm outlined in Appendix B.

3.1 Branin function

Here, we reconstruct a function with unknown coefficients based on sparse measurements of the function. Specifically, we consider the modified Branin function [14]

$$f(\mathbf{x}) = a(\bar{y} - b\bar{x}^2 + c\bar{x} - r)^2 + g(1-p)\cos(\bar{x}) + g + qx, \quad (3.1)$$

where $\mathbf{x} = (x, y)^\top$,

$$\bar{x} = 15x - 5, \quad \bar{y} = 15y, \quad (x, y) \in \mathbb{D} = [0, 1] \times [0, 1],$$

and

$$a = 1, \quad b = \frac{5.1}{4\pi^2}, \quad c = \frac{5}{\pi}, \quad r = 6, \quad g = 10, \quad p = \frac{1}{8\pi}, \quad q = 5.$$

The contours of f , together with eight randomly chosen observation locations are presented in Figure 2. The function f is evaluated on 41×41 uniform grids, so that the resulting discrete field \mathbf{F} is a 41×41 matrix. In this example, the stochastic model $u(\mathbf{x}; \omega) : \mathbb{D} \times \Omega \rightarrow \mathbb{R}$ is

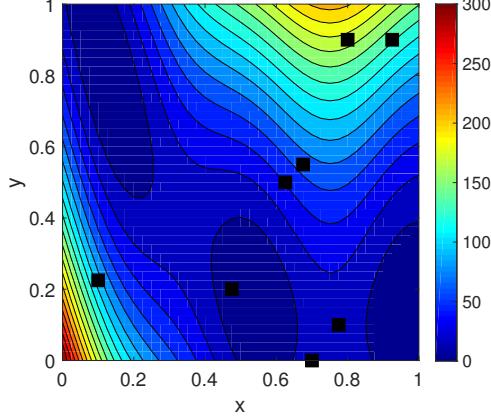


Figure 2: Contours of modified Branin function (on 41×41 uniform grids) and locations of eight observations (black squares).

obtained by modifying the second g in f , and treating the unknown coefficients b and q in $f(\mathbf{x})$ as random fields:

$$\hat{f}(\mathbf{x}; \omega) = a(\bar{y} - \hat{b}(\mathbf{x}; \omega)\bar{x}^2 + c\bar{x} - r)^2 + g(1 - p)\cos(\bar{x}) + \hat{g} + \hat{q}(\mathbf{x}; \omega)x, \quad (3.2)$$

where $\hat{g} = 20$,

$$\begin{aligned} \hat{b}(\mathbf{x}; \omega) &= b \left\{ 0.9 + \frac{0.2}{\pi} \sum_{i=1}^3 \left[\frac{1}{4i-1} \sin((2i-0.5)\pi x) \xi_{2i-1}(\omega) + \frac{1}{4i+1} \sin((2i+0.5)\pi y) \xi_{2i}(\omega) \right] \right\}, \\ \hat{q}(\mathbf{x}; \omega) &= q \left\{ 1.0 + \frac{0.6}{\pi} \sum_{i=1}^3 \left[\frac{1}{4i-3} \cos((2i-1.5)\pi x) \xi_{2i+5}(\omega) + \frac{1}{4i-1} \cos((2i-0.5)\pi y) \xi_{2i+6}(\omega) \right] \right\}, \end{aligned}$$

and $\{\xi_i(\omega)\}_{i=1}^{12}$ are iid Gaussian random variables with zero mean and unit variance. This stochastic model includes unkown ‘‘physics’’ (\hat{b} and \hat{q}) and incorrect ‘‘physics’’ (\hat{g}). We compute $M = 300$ realizations of $u(\mathbf{x}; \omega)$, denoted as $\{\hat{\mathbf{F}}^m\}_{m=1}^M$, by generating $M = 300$ samples of $\{\xi_i(\omega)\}_{i=1}^{12}$ and evaluating $u(\mathbf{x}; \omega)$ on the 41×41 uniform grids for each of them. Of note, function f is not a realization of $u(\mathbf{x}; \omega)$.

Figure 3 compares the purely data-driven reconstruction (i.e., Kriging) and purely ‘‘physics’’-based reconstruction (i.e., mean and variance obtained from the stochastic model Eq. (3.2) without conditioning on data). The first row in Figure 3 presents Kriging results obtained using the eight observations shown in Figure 2. The second row shows the ensemble mean (denoted as $\mu(\hat{\mathbf{F}}^m)$), standard deviation (denoted as $\sigma(\hat{\mathbf{F}}^m)$) of $\{\hat{\mathbf{F}}^m\}_{m=1}^M$, and $\mu(\hat{\mathbf{F}}^m) - \mathbf{F}$. The relative error is more than 50% for Kriging and 19% for the ensemble mean. In kriging, \hat{s} is large in the upper left subdomain where no observations are available. On the other hand, the standard

deviation of the realizations is large in the upper right region because of the randomness in the “physical” model. Similarly, $|\mathbf{F}_r - \mathbf{F}|$ is large in the upper left region, while $|\mu(\hat{\mathbf{F}}^m) - \mathbf{F}|$ is large in the upper right region. These results demonstrate that the data-driven and physics-based methods have significantly different accuracy in reconstructing the entire field and result in different estimates of uncertainty of the reconstruction.

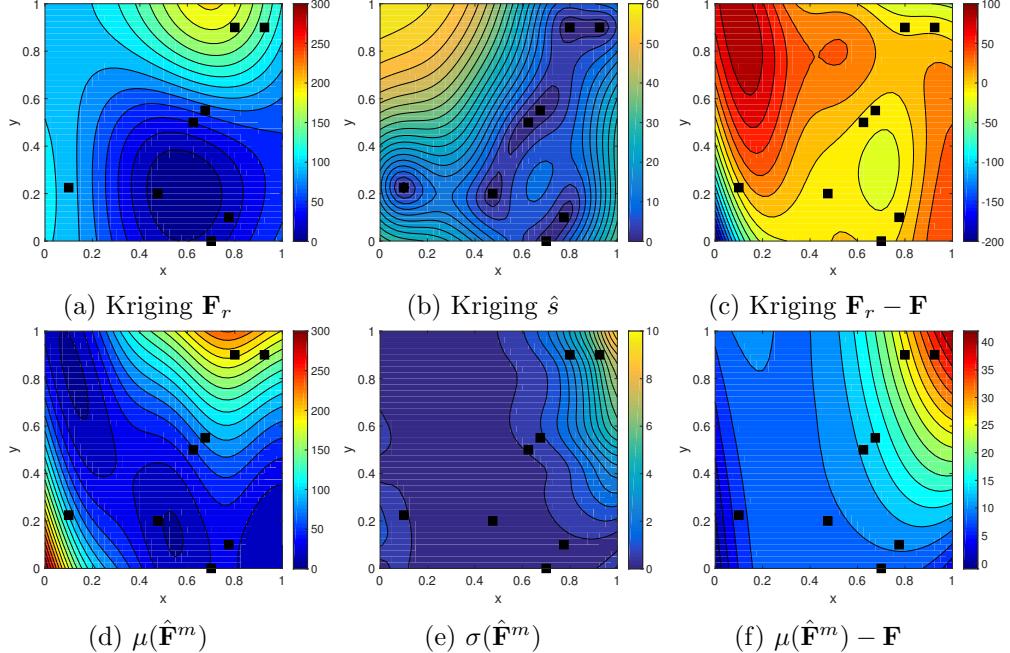


Figure 3: Reconstruction of the modified Branin function by Kriging (first row) and statistics of the ensemble $\{\hat{\mathbf{F}}^m\}_{m=1}^M$ (second row).

Figure 4 presents results obtained using PhIK and CoPhIK. In this case, CoPhIK outperforms PhIK in the accuracy, and both methods are significantly more accurate than Kriging. However, PhIK shows smaller \hat{s} than CoPhIK, and \hat{s} in both PhIK and CoPhIK is significantly smaller than in Kriging. This is because the prior covariance of PhIK is decided by the realizations, and it doesn't account for the observation. Therefore, the uncertainty of the PhIK result (\hat{s} of PhIK) is bounded by the uncertainty of the stochastic model ($\sigma(\hat{\mathbf{F}}^m)$). Also, the pattern of \hat{s} in PhIK is similar to that of $\sigma(\hat{\mathbf{F}}^m)$ in Figure 3(e), i.e., it is large in the right subdomain and small in the left subdomain. On the other hand, CoPhIK incorporates the observation in the posterior mean and kernel as we illustrate in Section 2, and its \hat{s} pattern is more similar to Kriging's in Figure 3(b) as we use the Gaussian kernel for both kriging and CoPhIK.

We use Algorithm 2 in combination with kriging, PhIK, and CoPhIK to perform active learning by adding one by one new observations of f at the global maximum of \hat{s} . Figure 5 presents the reconstructions of the modified Branin function with eight new observations added using active learning. In this figure, the first, second and third row corresponds to Kriging, PhIK and CoPhIK, respectively. The initial eight observations are marked by squares, and added observations are marked by stars. By comparing to Figures 3 and 4 it can be seen that reconstruction accuracy increases as more observations are added, and the uncertainty in the reconstruction is reduced. It can also be seen that the additional observation locations identified by CoPhIK are similar to that of kriging. In contrast, most of the additional observations identified by PhIK are on the right boundary.

Figure 6 compares the relative error $\|\mathbf{F}_r - \mathbf{F}\|_F/\|\mathbf{F}\|_F$ as a function of the total number of observations for active learning based on kriging, PhIK, modified PhIK and CoPhIK. For

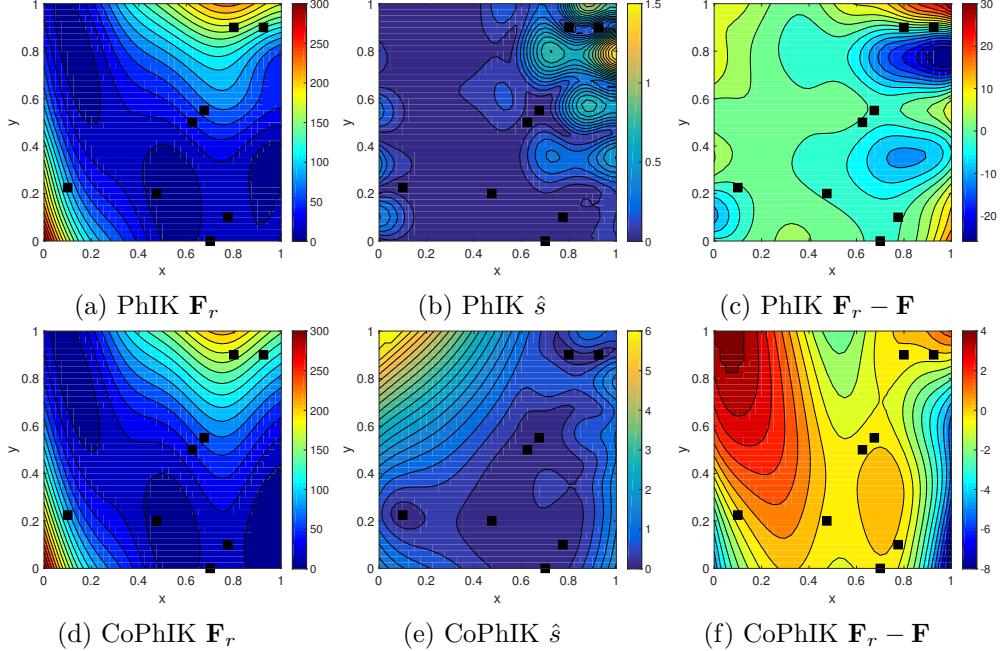


Figure 4: Reconstruction of the modified Branin function by PhIK (first row) and CoPhIK (second row).

the original eight observations, the largest error is in Kriging (over 50%) followed by PhIK and modified PhIK errors (about 8%), with the smallest error in CoPhIK (less than 3%). As more observations are added by the active learning algorithm, the error of Kriging decreases to approximately 1.5% at 24 observations. The error of PhIK is reduced from around 8% to approximately 4% at 12 observations. Adding 12 more observations doesn't improve the accuracy. The error of modified PhIK is reduced from 8% to 2% at 12 observations, then it changes very slowly with additional observations. With 24 observations, the accuracy of Kriging and modified PhIK is approximately the same. CoPhIK has the best accuracy among all the methods and, in general, its error decreases with additional measurements. The error in CoPhIK reduces to less than 0.1% with 24 observations in total, which is more than one order of magnitude better than the other three methods.

Finally, this example also illustrates that smaller uncertainty (\hat{s}) doesn't lead to smaller error in the posterior mean. In particular, in this case, PhIK has the smallest \hat{s} , but CoPhIK posterior mean is the most accurate.

3.2 Heat transfer

In the second example, we consider the steady state of a heat transfer problem. The dimensionless heat equation is given as

$$\frac{\partial T}{\partial t} - \nabla \cdot (\kappa(T) \nabla T) = 0, \quad \mathbf{x} \in \mathbb{D}, \quad (3.3)$$

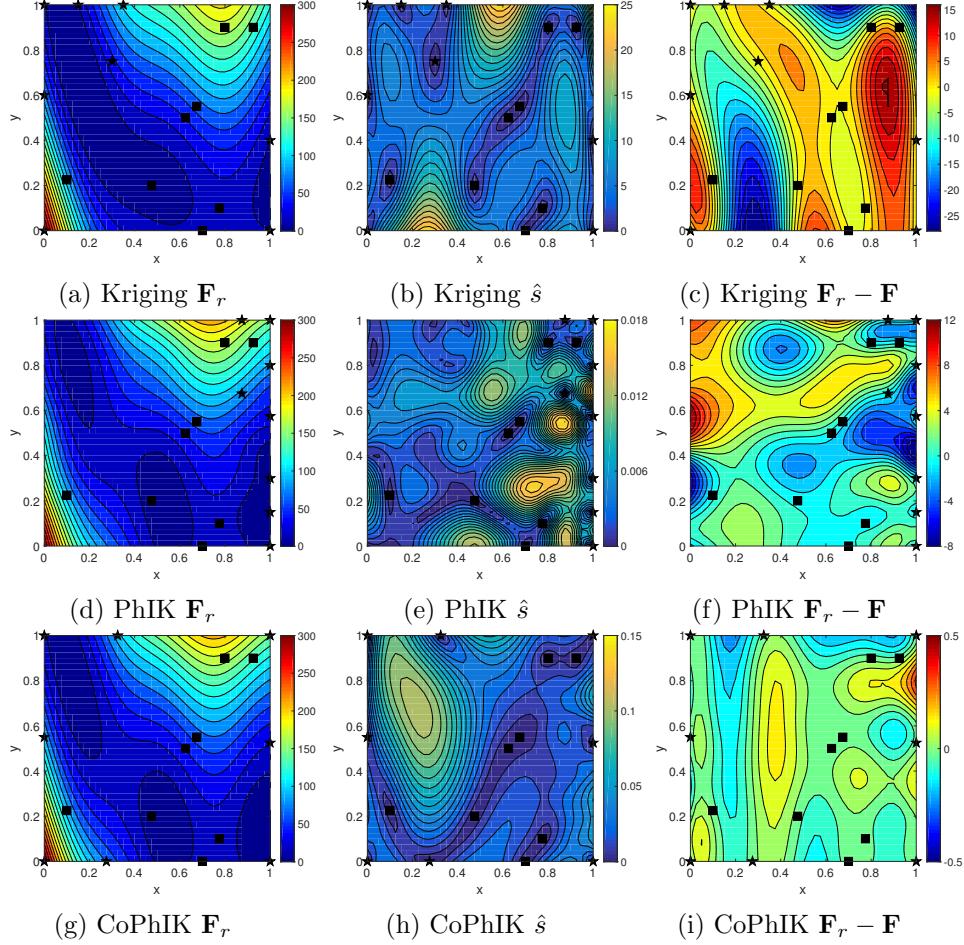


Figure 5: Reconstruction of the modified Branin function via active learning. Black squares are the locations of the original eight observations. Stars are newly added eight observations based on the actively learning algorithm.

subject to the boundary conditions:

$$\begin{cases} T = -30 \cos(2\pi x) + 40, & x \in \Gamma_1, \\ \frac{\partial T}{\partial \mathbf{n}} = -20, & x \in \Gamma_2, \\ T = 30 \cos(2\pi(x + 0.1)) + 40, & x \in \Gamma_3, \\ \frac{\partial T}{\partial \mathbf{n}} = 20, & x \in \Gamma_4, \\ \frac{\partial T}{\partial \mathbf{n}} = 0, & x \in \Gamma_5. \end{cases} \quad (3.4)$$

Here, $T(\mathbf{x}, t)$ is the temperature and $\kappa(T)$ is the temperature-dependent heat conductivity. The computational domain \mathbb{D} is the rectangle $[-0.5, 0.5] \times [-0.2, 0.2]$ with two circular cavities $R_1(O_1, r_1)$ and $R_2(O_2, r_2)$, with $O_1 = (-0.3, 0)$, $O_2 = (0.2, 0)$, $r_1 = 0.1$, $r_2 = 0.15$ (see Figure 7). The reference conductivity is set as

$$\kappa(T) = 1.0 + \exp(0.02T), \quad (3.5)$$

which results in the reference steady state temperature field shown in Figure 8. This solution was obtained by solving Eq. 3.3 and 3.4 using the finite element method with unstructured

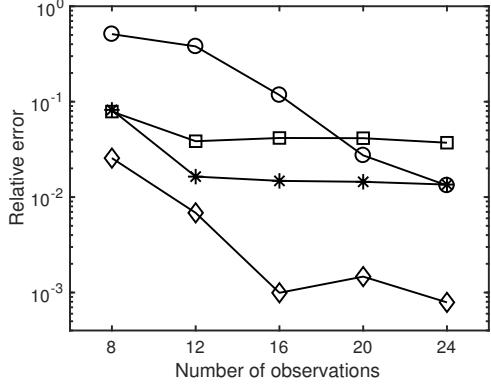


Figure 6: Relative error of reconstructed modified Branin function $\|\mathbf{F}_r - \mathbf{F}\|_F/\|\mathbf{F}\|_F$ using Kriging (“o”), PhIK (“□”), modified PhIK (“*”) and CoPhIK (“◊”) with different numbers of total observations via active learning.

triangular mesh implemented by the MATLAB PDE toolbox. The number of degrees of freedom is 1319, with a maximum grid size of 0.02. Observations of this exact profile are collected at six locations, marked by black squares in Figure 8.

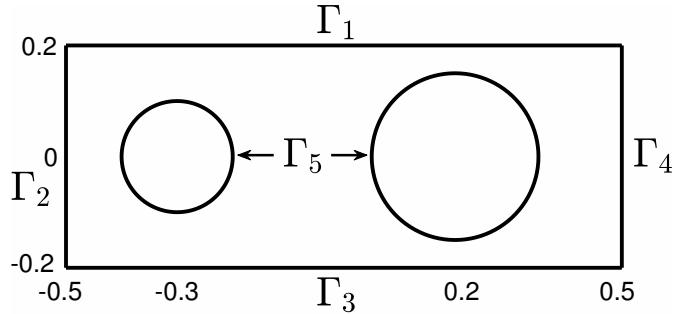


Figure 7: Heat transfer solution (steady state) computational domain.

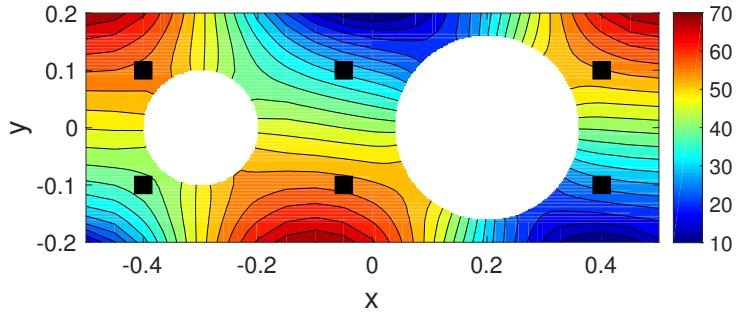


Figure 8: Contours of heat transfer solution (steady state) and locations of six observations (black squares).

Now we assume that the conductivity model (3.5) is unknown and an “expert knowledge” of κ is expressed as

$$\kappa(T; \omega) = 0.1 + \xi(\omega)T, \quad (3.6)$$

where $\xi(\omega)$ is a uniform random variable $\mathcal{U}[0.0012, 0.0108]$. Note that this example represents a biased expert knowledge that systematically underestimates the heat conductivity and assumes

an incorrect functional dependence of κ on T . We sample the stochastic model by generating $M = 400$ samples of $\xi(\omega)$ and then solving Eq. (3.3) for each realization. We denote the resulting ensemble of temperatures solutions by $\{\hat{\mathbf{F}}^m\}_{m=1}^M$.

The first row in Figure 9 presents the posterior mean, RMSE and pointwise reconstruction error of Kriging regression obtained with six measurements whose locations are also shown in this figure. The relative error is large (about 27%). The second row in Figure 9 shows the mean and standard deviation of the ensemble $\{\hat{\mathbf{F}}^m\}_{m=1}^M$ and the difference between the ensemble mean and the exact field. In this case, the relative error of the ensemble average is 8%, which may be acceptable in some application. For the selected number and locations of observations, Kriging performs worse than the unconditional stochastic model.

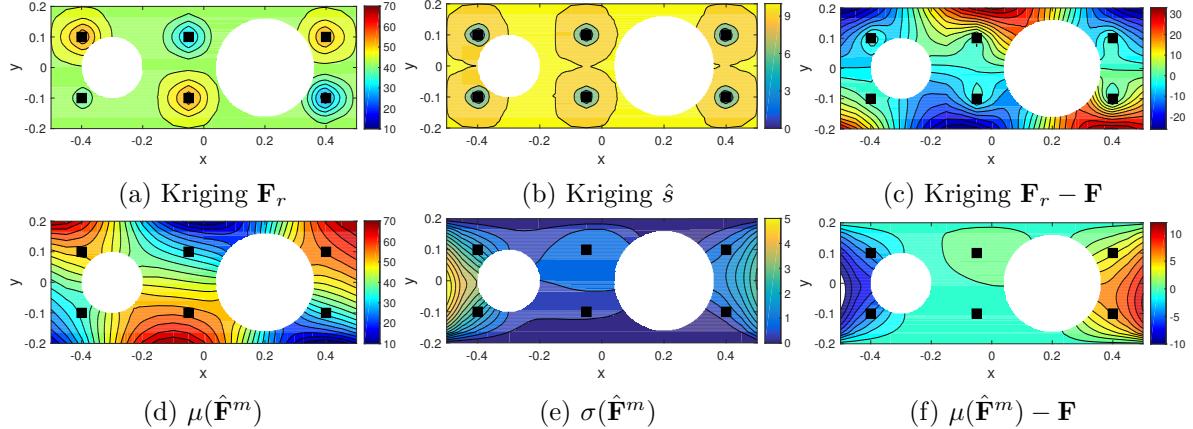


Figure 9: Reconstruction of the steady state solution for heat transfer problem by Kriging (first row) and statistics of the ensemble $\{\hat{\mathbf{F}}^m\}_{m=1}^M$ (second row).

Next, we use PhIK and CoPhIK to obtain prediction of T . Figure 10 shows the results for PhIK in the top row and for CoPhIK in the bottom row. CoPhIK outperforms PhIK as it results in smaller reconstruction errors. The relative errors of the reconstruction are 4.8% for CoPhIK and 7.8% for PhIK. As before, \hat{s} in PhIK is smaller than in CoPhIK. Both, PhIK and CoPhIK are more accurate and certain than Kriging.

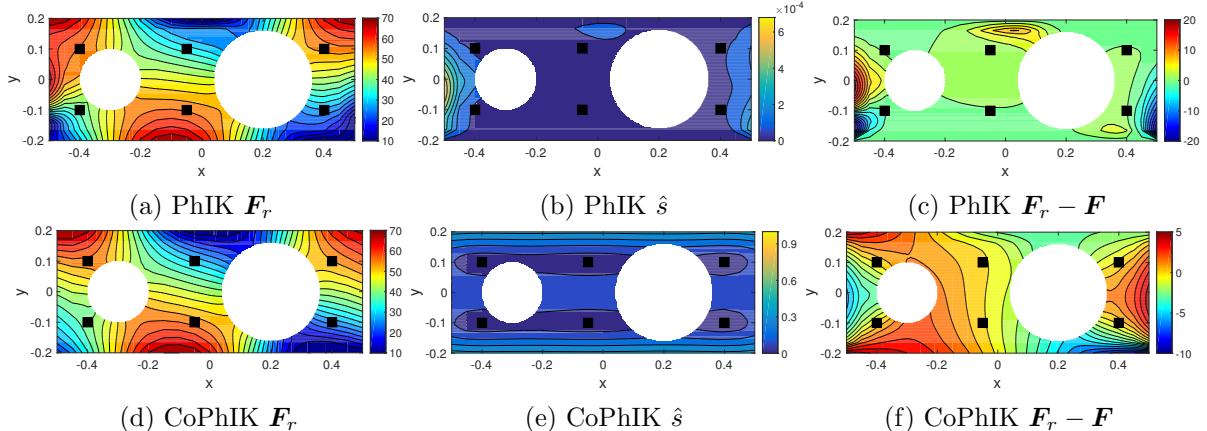


Figure 10: Reconstruction of the steady state solution for heat transfer problem by PhIK (first row) and CoPhIK (second row).

Finally, we employ active learning to identify additional observation locations. Figure 11 displays Kriging, PhIK, and CoPhIK predictions obtained with 14 observations. Eight new

observations are marked with stars and original six observations are denoted with squares. These three methods place additional observations at different locations. Kriging suggests new observation locations mostly along the external boundaries as there are no original observations on the boundaries and extrapolation in Kriging is the most uncertain in these subdomains. PhiIK identifies additional observation on the Neumann boundaries Γ_2 and Γ_4 . CoPhiIK identifies new observations locations on boundaries in a manner similar to Kriging, but also adds an observation location in the interior of \mathbb{D} . Figure 12 presents a quantitative study of the relative error as a function of the total number of observations for the three methods. It shows that CoPhiIK is more accurate than Kriging and PhiIK for a given number of observation points. As more observations are available, the errors in Kriging and CoPhiIK decrease while the error of PhiIK reaches a constant value after the first few observations. In this case, when 22 observations are used (six original ones plus 14 added ones through active learning), the relative errors are 3%, 4%, and less than 1% for Kriging, PhiIK and CoPhiIK, respectively. We also used the modified PhiIK method to model the data and found that the relative error in this method is slightly smaller than in PhiIK. However, the modified PhiIK reconstruction does not satisfy the Dirichlet boundary condition on Γ_1 and Γ_3 . Therefore, we do not report the modified PhiIK results.

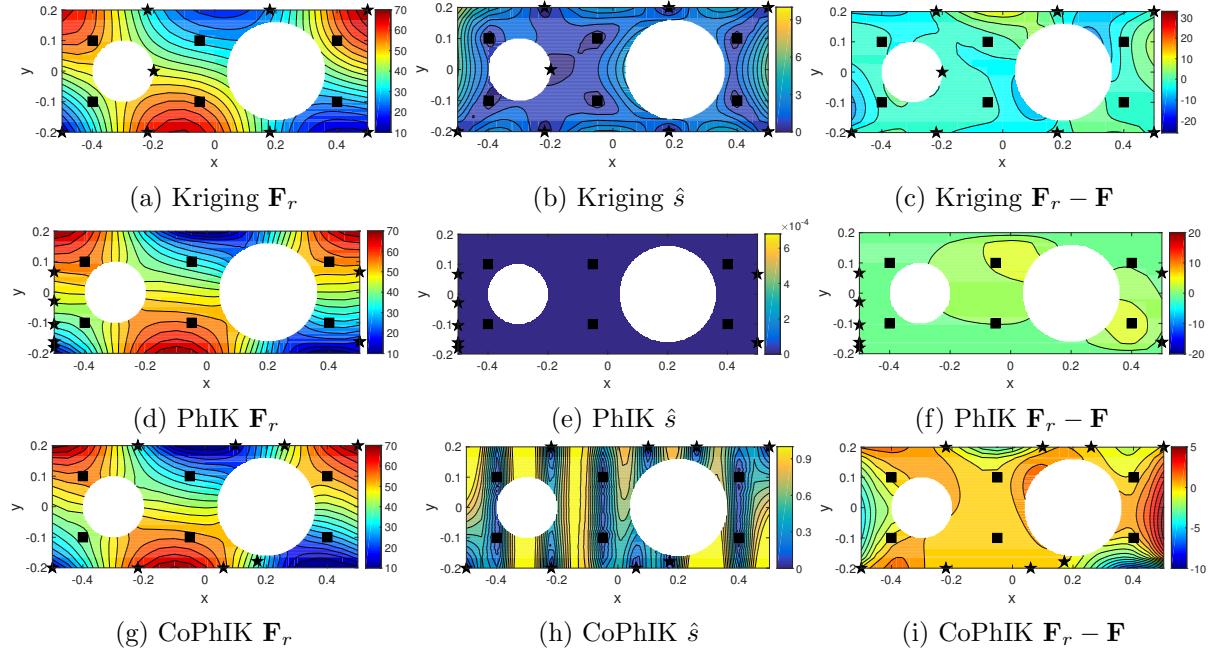


Figure 11: Reconstruction of the heat transfer via active learning. Black squares are the locations of the original six observation. Stars are newly added eight observations based on the actively learning algorithm.

3.3 Solute transport in heterogeneous porous media

In this example, we consider conservative transport in a steady-state velocity field in heterogeneous porous media. Let $C(\mathbf{x}, t)$ ($\mathbf{x} = (x, y)^\top$) denote the solute concentration. We assume that measurements of $C(\mathbf{x}, t)$ are available at several locations at different times. The flow and transport processes can be described by conservation laws. In particular, the flow is described

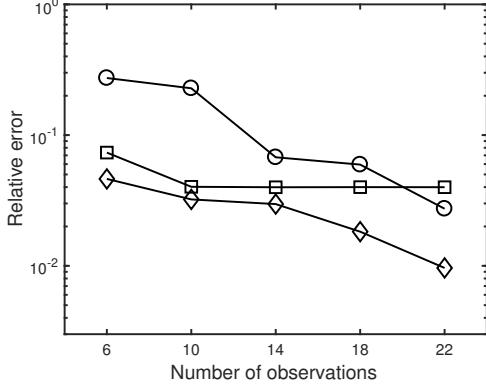


Figure 12: Relative error of reconstructed steady state solution for heat transfer problem $\|\mathbf{F}_r - \mathbf{F}\|_F/\|\mathbf{F}\|_F$ using Kriging (“ \circ ”), PhIK (“ \square ”) and CoPhIK (“ \diamond ”) with different numbers of total observations via active learning.

by the Darcy flow equation

$$\begin{cases} \nabla \cdot (K \nabla h) = 0, & \mathbf{x} \in \mathbb{D}, \\ \frac{\partial h}{\partial \mathbf{n}} = 0, & y = 0 \text{ or } y = L_2, \\ h = H_1 & x = 0, \\ h = H_2 & x = L_1, \end{cases} \quad (3.7)$$

where $h(\mathbf{x}; \omega)$ is the hydraulic head, $\mathbb{D} = [0, L_1] \times [0, L_2]$, $L_1 = 256$, $L_2 = 128$ is the simulation domain, H_1 and H_2 are known boundary head values, and $K(\mathbf{x})$ is the unknown hydraulic conductivity field. This field is modeled as a random log-normally distributed field $K(\mathbf{x}; \omega) = \exp(Z(\mathbf{x}; \omega))$, where $Z(\mathbf{x}; \omega)$ is a second-order stationary GP with known exponential covariance function $\text{Cov}\{Z(\mathbf{x}), Z(\mathbf{x}')\} = \sigma_Z^2 \exp(-|\mathbf{x} - \mathbf{x}'|/l_z)$, variance $\sigma_Z^2 = 2$, and correlation length $l_z = 5$. The solute transport is governed by the advection-dispersion equation [12, 26]:

$$\begin{cases} \frac{\partial C}{\partial t} + \nabla \cdot (\mathbf{v}C) = \nabla \cdot \left[\left(\frac{D_w}{\tau} + \boldsymbol{\alpha} \|\mathbf{v}\|_2 \right) \nabla C \right], & \mathbf{x} \text{ in } \mathbb{D}, \\ C = Q\delta(\mathbf{x} - \mathbf{x}^*), & t = 0, \\ \frac{\partial C}{\partial \mathbf{n}} = 0, & y = 0 \text{ or } y = L_2 \text{ or } x = L_1, \\ C = 0, & x = 0. \end{cases} \quad (3.8)$$

Here, $C(\mathbf{x}, t; \omega)$ is the solute concentration defined on $\mathbb{D} \times [0, T] \times \Omega$; \mathbf{v} is the fluid velocity given by $\mathbf{v}(\mathbf{x}; \omega) = -K(\mathbf{x}; \omega) \nabla h(\mathbf{x}; \omega)/\phi$, where ϕ is the porosity; D_w is the diffusion coefficient; τ is the tortuosity; and $\boldsymbol{\alpha}$ is the dispersivity tensor with the diagonal components α_L and α_T . In the present work, the transport parameters are set to $\phi = 0.317$, $\tau = \phi^{1/3}$, $D_w = 2.5 \times 10^{-5}$, $\alpha_L = 5$, and $\alpha_T = 0.5$. Finally, the solute is instantaneously injected at $\mathbf{x}^* = (50, 64)$ at $t = 0$ with the intensity $Q = 1$.

We are interested in reconstructing the concentration field at $T = 192$ (eight days) from sparse observations collected at $t = T$. We generate M realizations of $Z(\mathbf{x})$ using the SGSIM (sequential Gaussian simulation) code [11], and solve the governing equations for each realization of $K(\mathbf{x}) = \exp(Z(\mathbf{x}))$ using the finite volume code STOMP (subsurface transport over multiple phases) [47] with grid size 1×1 . The ground truth $C_e(\mathbf{x}, T)$ is randomly selected from the M realizations of $C(\mathbf{x}, T)$; this C_e is excluded from the ensembles used in PhIK or CoPhIK.

Figure 13 shows $C_e(\mathbf{x}, T)$ with sparse observation locations marked by black squares. We assume that six uniformly spaced observations are available near the boundary of the simulation domain, and nine randomly placed observations are available in the interior of the domain. As Kriging is known to be less accurate for extrapolation, it is a common practice to collect data near the boundary of the domain of interest (e.g., [10]).

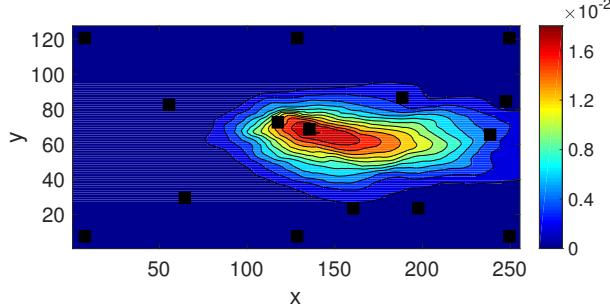


Figure 13: Ground truth of the solute concentration when $T = 192$ and observation locations (black squares).

The first row in Figure 14 shows the reconstruction results obtained using Kriging with Gaussian kernel. The relative error is nearly 50% because a stationary kernel is not capable of resolving the reference field accurately. The second row displays the ensemble mean, standard deviation and the difference of mean and the ground truth, estimated from the stochastic flow and advection-dispersion equation without conditioning on data. Solving these stochastic equations with standard MC requires a large number of simulations. To reduce computational cost of MC, we use MLMC (described in Appendix A) to compute mean and variance of C as in [52]. Later, we also use MLMC to compute the covariance of C . In MLMC, we use $M_H = 10$ high-resolution simulations (grid size 1×1) and $M_L = 150$ low-resolution simulations (grid size 4×4). The MLMC mean is almost symmetric which does not reflect the real pattern of the ground truth, which is not symmetric. The relative error of using ensemble mean to estimate the ground truth is 30%. This figure shows that Kriging prediction is less accurate and have larger predictive uncertainty except for the neighborhoods of the observations.

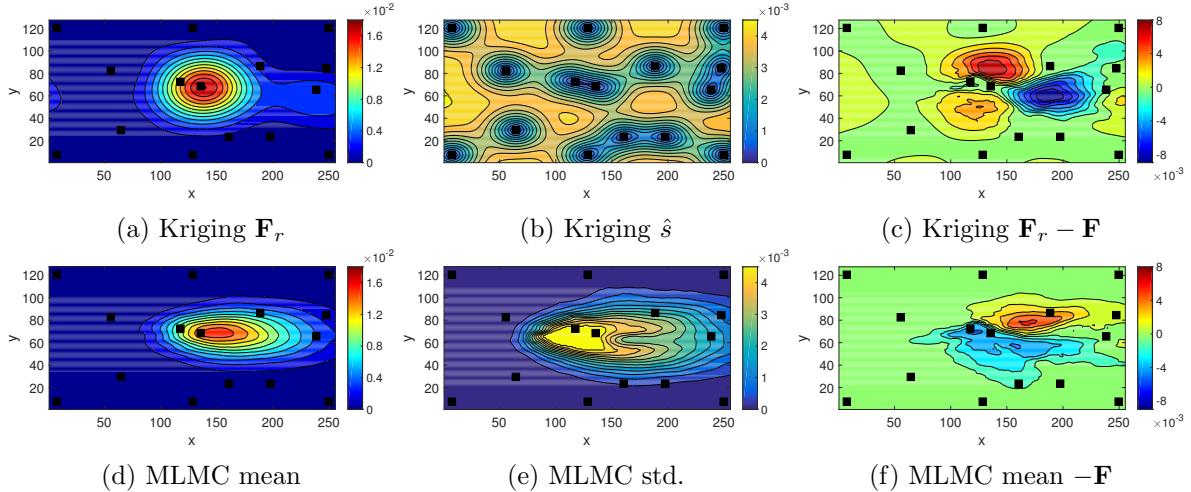


Figure 14: Reconstruction of the solute concentration by Kriging (first row) and statistics of MLMC combining the ensemble $\{\hat{\mathbf{F}}_L^m\}_{m=1}^{M_L}$ and $\{\hat{\mathbf{F}}_H^m\}_{m=1}^{M_H}$ (second row).

Figure 15 shows \mathbf{F}_r , \hat{s} , and $\mathbf{F}_r - \mathbf{F}$ obtained with PhIK and CoPhIK. In this case, PhIK is

more accurate than CoPhiK. The reconstructed field \mathbf{F}_r from both methods are closer to the ground truth than the Kriging results, as evident from smaller \hat{s} and $\mathbf{F}_r - \mathbf{F}$.

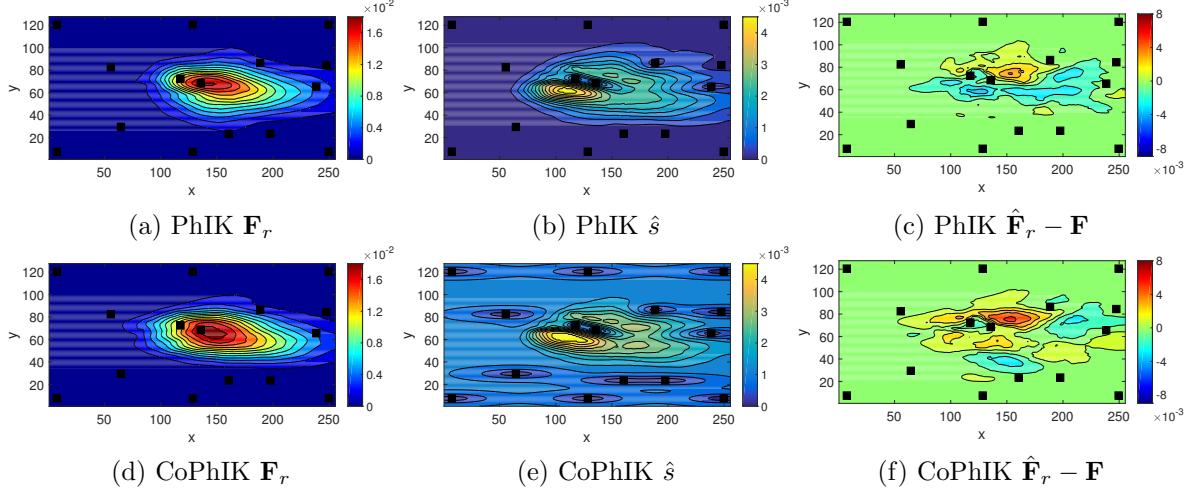


Figure 15: Reconstruction of the solute concentration field using PhIK (first row) and CoPhIK (second row).

Finally, we employ the active learning algorithm 2 to identify additional observation locations. Figure 16 presents 15 additional observation locations, indicated by black stars, identified using Kriging, PhIK and CoPhIK, and the resulting field reconstructions. Both PhIK and CoPhIK outperforms Kriging, which can be seen qualitatively in terms of the structure of the reconstructed plume and quantitative from the pointwise difference $\mathbf{F}_r - \mathbf{F}$. It can be seen that the additional observations identified by PhIK cluster around the plume, where the concentration is high, while Kriging distributes additional observations more uniformly throughout the entire simulation domain. The behavior of CoPhIK is between that of Kriging and PhIK, placing additional observations around the plume less tightly than PhIK but less spread out than Kriging.

Figure 17 presents the relative error as a function of the number of additional observation locations identified via active learning. It can be seen that PhIK is more accurate than CoPhIK, especially when number of observations is small. The difference between these two methods becomes smaller as more observations are introduced. The error of Kriging decreases in general with increasing number of observations, but is much larger than that of PhIK and CoPhIK. For modified PhIK, the magnitude of $\Delta\mu$ is $\mathcal{O}(10^{-6})$, so that its behavior is similar to that of PhIK.

This example is different from the previous two in that a stationary kernel is not suitable for reconstructing the reference solute concentration field. Similarly, a stationary kernel is not adequate for modeling the GP Y_d in CoPhIK. In addition, in this case PhIK outperforms CoPhIK because the ground truth is a realization of the stochastic model, i.e., the stochastic model in this case is accurate. This is different from the previous two examples where we use incorrect stochastic physical models. A carefully chosen non-stationary kernel function would be necessary to improve the accuracy of Kriging and CoPhIK.

4 Conclusion

In this work, we propose CoPhiK, a CoKriging-based multifidelity method that uses the PhIK method to combine numerical simulations of physical systems with accurate observations. The

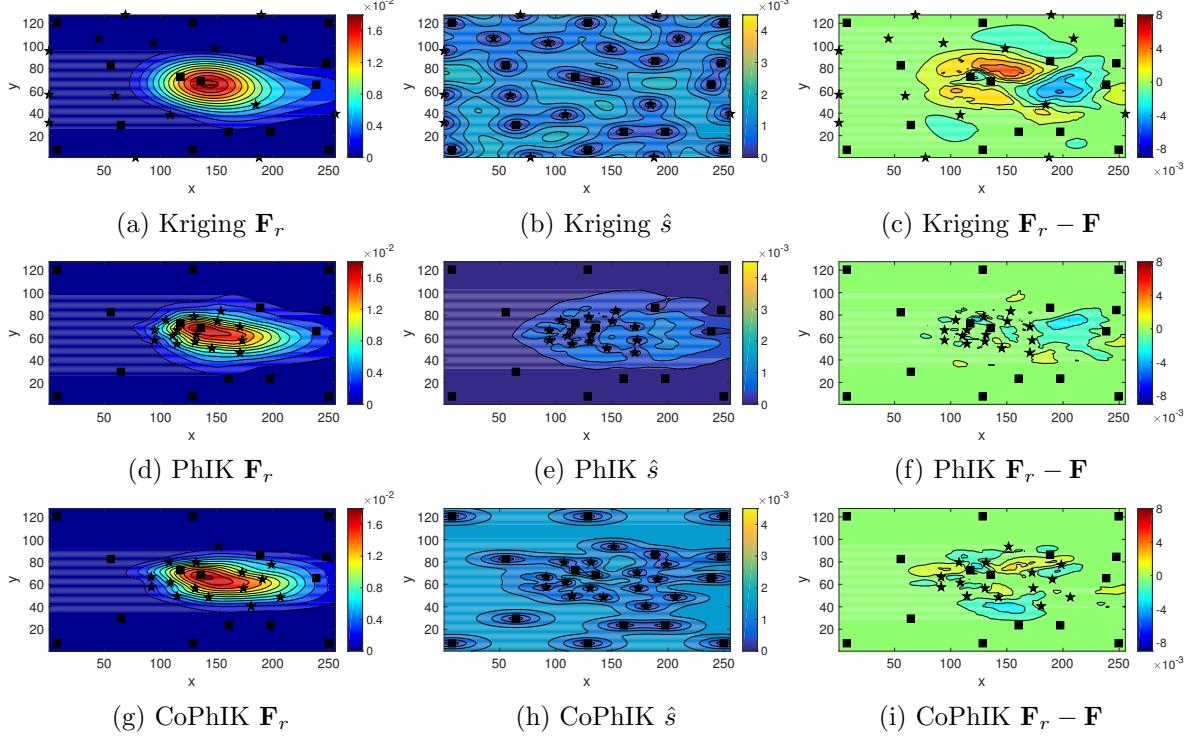


Figure 16: Reconstruction of the solute concentration field via active learning using Kriging (first row), PhIK (second row) and CoPhIK (third row). Black squares are the locations of the original 15 observation. Stars are 15 newly added observations.

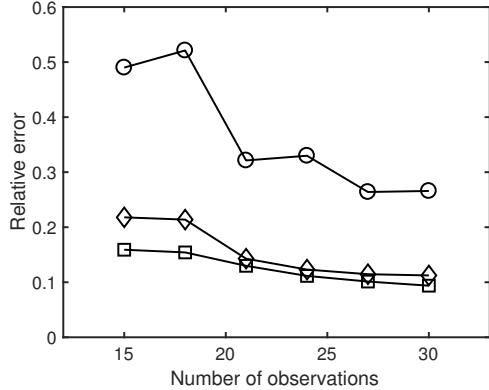


Figure 17: Relative error of reconstructed solute concentration $\|\mathbf{F}_r - \mathbf{F}\|_F / \|\mathbf{F}\|_F$ of Kriging (“o”), PhIK (“□”) and CoPhIK (“◊”) using different numbers of total observations via active learning.

CoPhIK method first constructs a “low-fidelity” GP Y_L via PhIK by estimating its mean and covariance function from the output of a stochastic physical model reflecting partial knowledge of the system; then, it models the discrepancy between high-fidelity data (e.g., observations of the system) and the low-fidelity GP using auxiliary GP Y_d . Whereas in PhIK the (prior) mean and covariance function are entirely defined by the stochastic model outputs, CoPhIK incorporates high-fidelity data in constructing the prior mean and covariance. In addition, we propose a modified version of PhIK by introducing a correction term to the prior mean. We also provide upper bounds for the error in enforcing physical constraints using CoPhIK and

modified PhIK. Finally, we demonstrate that an active learning algorithm in combination with Kriging, PhIK and CoPhIK suggests very different locations for new observations, and the two physics-informed methods result in significantly more accurate predictions with reduced uncertainty.

The CoPhIK method presented in this work consists of a non-stationary part Y_L , and a stationary part Y_d , in contrast with the “data-driven” Kriging method, for which the prior mean and covariance are estimated from data only, usually requiring an assumption of stationarity. The accuracy of CoPhIK predictions and of enforcing physical constraints depends both on the accuracy of the physical model and the selection of the kernel for Y_d . One can further improve CoPhIK by employing a non-stationary model for Y_d , thus rendering the GP Y_H fully non-stationary. The choice of non-stationary kernel for Y_d is problem dependent, and it may be achieved by exploiting additional physical information whenever available.

The presented physics-informed methods are nonintrusive, and can utilize existing domain codes to compute the necessary ensembles. Therefore, these methods are suitable for large-scale complex applications for which physical models and codes are available.

Acknowledgments

We thank Dr. Nathan Baker for fruitful discussion. This work was supported by the U.S. Department of Energy (DOE), Office of Science, Office of Advanced Scientific Computing Research (ASCR) as part of the Uncertainty Quantification in Advection-Diffusion-Reaction Systems. A portion of the research described in this paper was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle for the DOE under Contract DE-AC05-76RL01830.

Appendices

A. Constructing GP $Y(\mathbf{x})$ in PhIK using MLMC

For simplicity, we demonstrate the idea via two-level MLMC [52]. We use $u_L^m(\mathbf{x})$ ($m = 1, \dots, M_L$) and $u_H^m(\mathbf{x})$ ($m = 1, \dots, M_H$) to denote M_L low-accuracy and M_H high-accuracy realizations of the stochastic model $u(\mathbf{x}; \omega)$ for the system. In this work, u_L^m are simulation results on coarse grids \mathbb{D}_L , and u_H^m are simulations results on fine grids \mathbb{D}_H . We denote $\bar{u}(\mathbf{x}) = u_H(\mathbf{x}) - u_L(\mathbf{x})$. Here, when computing \bar{u} , we interpolate u_L from \mathbb{D}_L to \mathbb{D}_H . The mean of $Y(\mathbf{x})$ is estimated as

$$\mathbb{E}\{Y(\mathbf{x})\} = \mu(\mathbf{x}) \approx \mu_{\text{MLMC}}(\mathbf{x}) = \frac{1}{M_L} \sum_{m=1}^{M_L} u_L^m(\mathbf{x}) + \frac{1}{M_H} \sum_{m=1}^{M_H} \bar{u}^m(\mathbf{x}). \quad (\text{A.1})$$

which is the standard MLMC estimate of the mean [17]. The covariance function of $Y(\mathbf{x})$ is estimated as:

$$\begin{aligned} \text{Cov}\{Y(\mathbf{x}), Y(\mathbf{x}')\} &\approx k_{\text{MLMC}}(\mathbf{x}, \mathbf{x}') \\ &= \frac{1}{M_L - 1} \sum_{m=1}^{M_L} \left(u_L^m(\mathbf{x}) - \frac{1}{M_L} \sum_{m=1}^{M_L} u_L^m(\mathbf{x}) \right) \left(u_L^m(\mathbf{x}') - \frac{1}{M_L} \sum_{m=1}^{M_L} u_L^m(\mathbf{x}') \right) \\ &\quad + \frac{1}{M_H - 1} \sum_{m=1}^{M_H} \left(\bar{u}^m(\mathbf{x}) - \frac{1}{M_H} \sum_{m=1}^{M_H} \bar{u}^m(\mathbf{x}) \right) \left(\bar{u}^m(\mathbf{x}') - \frac{1}{M_H} \sum_{m=1}^{M_H} \bar{u}^m(\mathbf{x}') \right). \end{aligned} \quad (\text{A.2})$$

Finally, the MLMC-based PhIK model takes the form

$$\hat{y}(\mathbf{x}^*) = \mu_{\text{MLMC}}(\mathbf{x}^*) + \mathbf{c}_{\text{MLMC}}^\top \mathbf{C}_{\text{MLMC}}^{-1} (\mathbf{y} - \boldsymbol{\mu}_{\text{MLMC}}), \quad (\text{A.3})$$

where $\boldsymbol{\mu}_{\text{MLMC}} = (\mu_{\text{MLMC}}(\mathbf{x}^{(1)}), \dots, \mu_{\text{MLMC}}(\mathbf{x}^{(N)}))^\top$. The matrix \mathbf{C}_{MLMC} and vector \mathbf{c}_{MLMC} are approximations of \mathbf{C} in Eq. (2.4) and \mathbf{c} in Eq. (2.8) using k_{MLMC} in Eq. (A.2). The MSE of this prediction is

$$\hat{s}^2(\mathbf{x}^*) = \sigma_{\text{MLMC}}^2(\mathbf{x}^*) - \mathbf{c}_{\text{MLMC}}^\top \mathbf{C}_{\text{MLMC}}^{-1} \mathbf{c}_{\text{MLMC}}, \quad (\text{A.4})$$

where $\sigma_{\text{MLMC}}^2(\mathbf{x}^*) = k_{\text{MLMC}}(\mathbf{x}^*, \mathbf{x}^*)$. If i.i.d. Gaussian noise is assumed in the observation, replace \mathbf{C}_{MLMC} with $\mathbf{C}_{\text{MLMC}} + \delta^2 \mathbf{I}$, where δ^2 is the variance of the noise.

B. Active learning

In this work, *active learning* is a process of identifying locations for additional observations that minimize the prediction error and reduce MSE or uncertainty, e.g., [8, 19, 46, 9]. We use a greedy algorithm to add additional observations, i.e., to add new observations at the maxima of $s(\mathbf{x})$, e.g., [14, 38]. Then, we can make a new prediction $\hat{y}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{D}$ and compute a new $\hat{s}^2(\mathbf{x})$ to select the next location for additional observation (see Algorithm 2). This selection

Algorithm 2 Active learning based on GPR

- 1: Specify the locations \mathbf{X} , corresponding observations \mathbf{y} , and the maximum number of observations N_{max} affordable. The number of available observations is denoted as N .
 - 2: **while** $N_{\text{max}} > N$ **do**
 - 3: Compute the MSE $\hat{s}^2(\mathbf{x})$ of MLE prediction $\hat{y}(\mathbf{x})$ for $\mathbf{x} \in \mathbb{D}$.
 - 4: Locate the location \mathbf{x}_m for the maximum of $\hat{s}^2(\mathbf{x})$ for $\mathbf{x} \in \mathbb{D}$.
 - 5: Obtain observation y_m at \mathbf{x}_m and set $\mathbf{X} = \{\mathbf{X}, \mathbf{x}_m\}$, $\mathbf{y} = (\mathbf{y}^\top, y_m)^\top$, $N = N + 1$.
 - 6: **end while**
 - 7: Construct the MLE prediction of $\hat{y}(\mathbf{x})$ on \mathbb{D} using \mathbf{X} and \mathbf{y} .
-

criterion is based on the statistical interpretation of the interpolation. More sophisticated sensor placement algorithms can be found in literature, e.g., [19, 23, 16], and PhIK or CoPhIK are complementary to these methods.

References

- [1] Margaret Armstrong. Problems with universal kriging. *Journal of the International Association for Mathematical Geology*, 16(1):101–108, 1984.
- [2] David A Barajas-Solano and Alexandre M Tartakovsky. Multivariate gaussian process regression for multiscale data assimilation and uncertainty reduction. *arXiv preprint arXiv:1804.06490*, 2018.
- [3] David A. Barajas-Solano and Alexandre M. Tartakovsky. Probability and cumulative density function methods for the stochastic advection-reaction equation. *SIAM/ASA Journal on Uncertainty Quantification*, 6(1):180–212, 2018.
- [4] Sofiane Brahim-Belhouari and Amine Bermak. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, 47(4):705–712, 2004.
- [5] Christopher James Brooks, AJ Forrester, AJ Keane, and S Shahpar. Multi-fidelity design optimisation of a transonic compressor rotor. 2011.
- [6] Oksana A Chkrebtii, David A Campbell, Ben Calderhead, Mark A Girolami, et al. Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis*, 11(4):1239–1267, 2016.

- [7] Jon Cockayne, Chris Oates, Tim Sullivan, and Mark Girolami. Bayesian probabilistic numerical methods. *arXiv preprint arXiv:1702.03673*, 2017.
- [8] David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [9] Timothé Collet and Olivier Pietquin. Optimism in active learning with Gaussian processes. In *International Conference on Neural Information Processing*, pages 152–160. Springer, 2015.
- [10] Heng Dai, Xingyuan Chen, Ming Ye, Xuehang Song, and John M Zachara. A geostatistics-informed hierarchical sensitivity analysis method for complex groundwater flow and transport modeling. *Water Resources Research*, 53(5):4327–4343, 2017.
- [11] Clayton V Deutsch and André G Journel. *GSLIB: Geostatistical Software Library and User’s Guide*. Oxford University Press, 1992.
- [12] Simon Emmanuel and Brian Berkowitz. Mixing-induced precipitation and porosity evolution in porous media. *Advances in Water Resources*, 28(4):337–344, 2005.
- [13] Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367, 2003.
- [14] Alexander Forrester, Andy Keane, et al. *Engineering Design via Surrogate Modelling: A Practical Guide*. John Wiley & Sons, 2008.
- [15] Alexander IJ Forrester, András Sóbester, and Andy J Keane. Multi-fidelity optimization via surrogate modelling. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 463, pages 3251–3269. The Royal Society, 2007.
- [16] Roman Garnett, Michael A Osborne, and Stephen J Roberts. Bayesian optimization for sensor set selection. In *Proceedings of the 9th ACM/IEEE international conference on Information Processing in Sensor Networks*, pages 209–219. ACM, 2010.
- [17] Michael B Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.
- [18] Philipp Hennig, Michael A Osborne, and Mark Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings of the Royal Society London A*, 471(2179):20150142, 2015.
- [19] Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.
- [20] Marc C Kennedy and Anthony O’Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.
- [21] Peter K Kitanidis. *Introduction to Geostatistics: Applications in Hydrogeology*. Cambridge University Press, 1997.
- [22] M Knotters, DJ Brus, and JH Oude Voshaar. A comparison of kriging, co-kriging and kriging combined with regression for spatial interpolation of horizon depth with censored observations. *Geoderma*, 67(3-4):227–246, 1995.

- [23] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(Feb):235–284, 2008.
- [24] Julien Laurenceau and P Sagaut. Building efficient response surfaces of aerodynamic functions with kriging and cokriging. *AIAA journal*, 46(2):498–507, 2008.
- [25] Loic Le Gratiet and Josselin Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.
- [26] Guang Lin and Alexandre M Tartakovsky. An efficient, high-order probabilistic collocation method on sparse grids for three-dimensional flow and solute transport in randomly heterogeneous porous media. *Advances in Water Resources*, 32(5):712–722, 2009.
- [27] David JC MacKay. Introduction to gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, 168:133–166, 1998.
- [28] Karla Monterrubio-Gómez, Lassi Roininen, Sara Wade, Theo Damoulas, and Mark Girolami. Posterior inference for sparse hierarchical non-stationary models. *arXiv preprint arXiv:1804.01431*, 2018.
- [29] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.
- [30] Harald Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 63. SIAM, 1992.
- [31] Anthony O’Hagan. A Markov property for covariance structures, 1998.
- [32] Christopher J Paciorek and Mark J Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems*, pages 273–280, 2004.
- [33] Wenxiao Pan, Xiu Yang, Jie Bao, and Michelle Wang. Optimizing discharge capacity of li-o₂ batteries by design of air-electrode porous structure: Multifidelity modeling and optimization. *Journal of The Electrochemical Society*, 164(11):E3499–E3511, 2017.
- [34] Guofei Pang, Liu Yang, and George Em Karniadakis. Neural-net-induced Gaussian process regression for function approximation and PDE solution. *arXiv preprint arXiv:1806.11187*, 2018.
- [35] P Perdikaris, D Venturi, JO Royset, and GE Karniadakis. Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields. *Proc. R. Soc. A*, 471(2179):20150018, 2015.
- [36] Paris Perdikaris, Maziar Raissi, Andreas Damianou, ND Lawrence, and George Em Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society London A*, 473(2198):20160751, 2017.
- [37] Christian Plagemann, Kristian Kersting, and Wolfram Burgard. Nonstationary gaussian process regression using point estimates of local smoothness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 204–219. Springer, 2008.

- [38] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017.
- [39] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Numerical Gaussian processes for time-dependent and nonlinear partial differential equations. *SIAM Journal on Scientific Computing*, 40(1):A172–A198, 2018.
- [40] Jerome Sacks, William J Welch, Toby J Mitchell, and Henry P Wynn. Design and analysis of computer experiments. *Statistical Science*, pages 409–423, 1989.
- [41] Paul D Sampson and Peter Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.
- [42] Michael Schober, David K Duvenaud, and Philipp Hennig. Probabilistic ode solvers with Runge-Kutta means. In *Advances in Neural Information Processing Systems*, pages 739–747, 2014.
- [43] A Stein and LCA Corsten. Universal kriging and cokriging as a regression procedure. *Biometrics*, pages 575–587, 1991.
- [44] Michael L Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 2012.
- [45] A. M. Tartakovsky, M. Panzeri, G. D. Tartakovsky, and A. Guadagnini. Uncertainty quantification in scale-dependent models of flow in porous media. *Water Resources Research*, 53:9392–9401, 2017.
- [46] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2(Nov):45–66, 2001.
- [47] Mark D White and Martinus Oostrom. STOMP subsurface transport over multiple phases, version 4.0, users guide. Technical report, PNNL-15782, Richland, WA, 2006.
- [48] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for machine learning. *The MIT Press*, 2(3):4, 2006.
- [49] Dongbin Xiu and Jan S. Hesthaven. High-order collocation methods for differential equations with random inputs. *SIAM Journal on Scientific Computing*, 27(3):1118–1139, 2005.
- [50] Xiu Yang, Minseok Choi, Guang Lin, and George Em Karniadakis. Adaptive ANOVA decomposition of stochastic incompressible and compressible flows. *Journal of Computational Physics*, 231(4):1587–1614, 2012.
- [51] Xiu Yang and George Em Karniadakis. Reweighted ℓ_1 minimization method for stochastic elliptic differential equations. *Journal of Computational Physics*, 248(1):87–108, 2013.
- [52] Xiu Yang, Guzel Tartakovsky, and Alexandre Tartakovsky. PhIK: A physics informed Gaussian process regression method for data-model convergence. *arXiv preprint arXiv:1809.03461*, 2018.
- [53] Xueyu Zhu, Erin M Linebarger, and Dongbin Xiu. Multi-fidelity stochastic collocation method for computation of statistical moments. *Journal of Computational Physics*, 341:386–396, 2017.