Bayesian analysis of multifidelity computer models with local features and non-nested experimental designs: Application to the WRF model

Bledar A. Konomi [*]
Department of Mathematical Sciences, University of Cincinnati, USA
and
Georgios Karagiannis [*]
Department of Mathematical Sciences, Durham University, UK

October 2, 2020

Abstract

Motivated by a multi-fidelity Weather Research and Forecasting (WRF) climate model application where the available simulations are not generated based on hierarchically nested experimental design, we develop a new ko-criging procedure called Augmented Bayesian Treed Co-Kriging. The proposed procedure extends the scope of co-kriging in two major ways. We introduce a binary treed partition latent process in the multifidelity setting to account for non-stationary and potential discontinuities in the model outputs at different fidelity levels. Moreover, we introduce an efficient imputation mechanism which allows the practical implementation of co-kriging when the experimental design is non-hierarchically nested by enabling the specification of semi-conjugate priors. Our imputation strategy allows the design of an efficient RJ-MCMC implementation that involves collapsed blocks and direct simulation from conditional distributions. We develop the Monte Carlo recursive emulator which provides a Monte Carlo proxy for the full predictive distribution of the model output at each fidelity level, in a computationally feasible manner. The performance of our method is demonstrated on benchmark examples and used for the analysis of a large-scale climate modeling application which involves the WRF model. *Supplementary materials are available online.*

Keywords: Augmented hierarchically nested design, Binary treed partition, Gaussian process, Collapsed MCMC

[*] The two authors contributed equally to this work. Corresponding authors: Bledar A. Konomi (alex.konomi@uc.edu) and Georgios Karagiannis (georgios.karagiannis@durham.ac.uk).

1 Introduction

Understanding the behavior as well as the underlying mechanisms of real systems such as physical procedures is central to many applications such as weather forecasting. Direct investigation of the real system is often impossible due to limited resources, and hence it is simulated by computer models aiming at reproducing the real system's behavior with high accuracy. Our case study involves an expensive computer model which requires a significant amount of resources to perform a single run; and hence, only a limited number of simulations can be performed. Gaussian process (GP) regression models (Sacks et al., 1989) are statistical models that allow the emulation of the computer model output by using only a few runs of the computer model.

Computer models are often able to run at different levels of fidelity, sophistication, or resolution. As high fidelity runs are usually more expensive, collecting data by simulating the model at different fidelity levels is preferred for a given budget of resources. Statistical inference is preferable to be made against the whole simulated data set, and thus account for across fidelity level dependence, rather than against simulated data sets associated with individual fidelity levels (Kennedy and O'Hagan, 2000). Assume there are available S deterministic computer models $\{\mathfrak{C}_t\}_{t=1}^{S}$ aiming at simulating the same real system. The models are ordered by ascending fidelity level t. Let $y_t(\boldsymbol{x}) : \mathcal{X} \rightarrow \mathbb{R}$ denote the output function of the computer model $\mathscr{C}_t$ with respect to a m-dimensional input $x \in \mathcal{X}$. Autoregressive co-kriging assumes an autoregressive model

$$y_t(\boldsymbol{x}) = \xi_{t-1}(\boldsymbol{x}) y_{t-1}(\boldsymbol{x}) + \delta_t(\boldsymbol{x}) \qquad \text{for } x \in \mathcal{X},\ t = 2, ..., S \tag{1}$$

where $y_{t-1}(\boldsymbol{x})$, $\delta_t(\boldsymbol{x})$, $\xi_{t-1}(\boldsymbol{x})$ are independent unknown functions a priori modeled as Gaussian processes. Here, $\delta_t(\cdot)$ is the location discrepancy function (representing a local adjustment from $\mathfrak{C}_{t-1}$ to $\mathfrak{C}_t$), and $\xi_t(\cdot)$ is the scale discrepancy (representing a scale change from

$\mathfrak{C}_{t-1}$ to $\mathfrak{C}_t$ for $t = 1, ..., S$). Discrepancy terms, $\{\delta_t(\cdot)\}$ and $\{\xi_t(\cdot)\}$, can be thought of as accounting for 'missing' or 'misrepresented' physical properties in the lower fidelity computer model $\mathfrak{C}_{t-1}$ with respect to the higher one $\mathfrak{C}_t$. Model (1) is induced by the Markovian condition $\text{cov}(y_t(\boldsymbol{x}), y_{t-1}(\boldsymbol{x}')|y_{t-1}(\boldsymbol{x})) = 0$ of the heuristic 'there is nothing more to learn about $y_t(\boldsymbol{x})$ from $y_{t-1}(\boldsymbol{x}')$ for any $\boldsymbol{x}' \neq \boldsymbol{x}$ given $y_{t-1}(\boldsymbol{x})$ is known ' which is broadly accepted in computer experiments (OH́agan, 1998). Originally, Kennedy and O'Hagan (2000) considered a constant $\xi_{t-1} = \xi_{t-1}(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$ in (1) to impose stationarity throughout t.

A number of important variations of the autoregressive co-kriging have been proposed. Qian et al. (2005) used $\xi_{t-1}(x)$ as a polynomial expansion to model more general autoregressive dependencies. Based on this, Qian and Wu (2008) considered the scale discrepancy as a function of the input space by casting it as a GP. In practice, their approach is applicable to problems with only two fidelity levels, as the computational overhead caused by using more fidelity levels is increased dramatically. Le Gratiet (2013); Le Gratiet and Garnier (2014) modeled the scale discrepancy as an expansion of bases defined on the inputs, and presented conditional conjugate priors which lead to standard conditional posterior distributions for the unknown coefficients of the expansion. However, casting the scale discrepancy as a basis expansion may require an undesirably large number of bases in order to explain small scale discrepancies; while it cannot represent discontinuities and sudden changes. Furthermore, this may aggravate non-identifiability between the scale and additive discrepancies. Perdikaris et al. (2015) proposed a machine learning framework, which uses sparse precision matrices of Gaussian-Markov random fields introduced by Lindgren et al. (2011). This facilitates computations that leverage on the sparsity of the resulting discrete operators. Perdikaris et al. (2017) relaxed the auto-regressive structure by using deep learning ideas, however the computational demands to train the model are significantly increased. The aforementioned developments require hierarchically nested experimental designs for computational reasons, otherwise the computational demands become impractical. This constraint prevents their practical implementation on a number of important real world problems where the available data set is not based on such nested designs.

We are motivated by a real world application that involves the Weather Research and Forecasting (WRF) regional climate model (Skamarock et al., 2008). WRF is an expensive computer model that allows the use of different resolutions leading to different fidelity levels. We consider the WRF with the Rapid Radiative Transfer Model for General Circulation Model (Pincus et al., 2003), with the Kain-Fritsch convective parametrisation scheme (KF CPS) (Kain, 2004), and with five input parameters, while we are interested in the average precipitation as an output. The available simulations were generated by running WRF at two resolution levels, 12.5km and 25km grid spacing. The fidelity of the simulations increases when the grid spacing gets finer. The available simulations have not been generated based on a hierarchically nested design, while it is not possible to re-run the expensive computer model in our facilities and generate simulations based on such a design due to the high computational cost required. The aforesaid co-kriging methods cannot be implemented directly due to the lack of nested design, and hence new developments are required. We are interested in designing an accurate emulator that aggregates all the available simulations as well as represents features of the WRF. Previous research in (Yan et al., 2014; Yang et al., 2012) suggested that discrepancies between the two levels may depend on the five inputs of the KF CPS. Interest also lies in better understanding how different grid spacing affects the discrepancies in WRF with respect to the input parameters. Existing co-kriging methods do not model/account for such behaviors, thus suitable extensions must be introduced.

We propose the Augmented Bayesian Treed co-kriging (ABTCK); a fully Bayesian method for building multifidelity emulators of computer models that extends the scope of co-kriging mainly in two ways. The proposed method is able to address applications where the available training data set has not necessarily been generated according to a hierarchically nested experimental design. To achieve this, we introduce a suitable imputation mechanism that augments the original data set with uncertain quantities which can be thought of as missing data from a complete data set generated based on an hierarchically nested design. The proposed imputation allows the specification of conditional conjugate priors, and analytic integration of a large number of dimensions from the posterior. A different remedy for non-hierarchically nested designs is studied in the Thesis of Zertuche (2015), however, unlike our approach, that approach leads to an approximation of the posterior distribution while it is concentrated to only two fidelity levels. Moreover, our method is able to account for non-stationary, and possible discontinuities. This is achieved by suitably specifying the statistical model as a combination of computationally convenient and GP regression models by using a binary treed partition which a priori follows a process similar to (Chipman et al., 1998; Gramacy and Lee, 2008). The additional flexibility of the proposed model aims at producing more accurate predictions as well as providing an insight of the model discrepancies. To facilitate inference, we propose a reversible jump Markov chain Monte Carlo (RJ-MCMC) implementation, tailored to the proposed model, that involves an efficient MCMC sampler which operates on the joint space of the missing data and the parameters, and consists of collapsed blocks. Due to the augmentation, the MCMC loop consists of local RJ updates operating on a lower dimensional state space and producing more acceptable proposals, and a block simulating the missing data directly from the conditionals. Finally, we propose the Monte Carlo recursive emulator, as an alternative to those in (Kennedy and O'Hagan, 2000; Le Gratiet and Garnier, 2014; Le Gratiet, 2013), which is able to provide fully Bayesian posterior predictive inference even with non-nested designs while keeping the computational cost lower than the others.

The rest of the paper is organized as follows. In Section 2, we present the proposed procedure; in Section 3, we provide numerical comparisons with other methods; and in Section 4, we implement our procedure for the analysis of the WRF model. Conclusions are summarised in Section 5.

2 The Augmented Bayesian Treed co-Kriging

We describe the development of our Augmented Bayesian treed co-kriging model (ABTCK) which extends the scope of co-kriging to applications with non-nested designs and/or non-stationary model outputs. A schematic is available in Supplementary Section S.1.

2.1 Treed auto-regressive co-kriging model

To account for non-stationarity we consider an unknown partition $\{\mathcal{X}_k\}_{k=1}^{K}$ of the input space $\mathcal{X}$, whose sub-regions are assumed to be homogeneous in the sense that a co-kriging model (1) can be defined independently at each sub-region, i.e.

$$y_{k,t}(\boldsymbol{x}) = \xi_{k,t-1}(\boldsymbol{x}) y_{k,t-1}(\boldsymbol{x}) + \delta_{k,t}(\boldsymbol{x}) \qquad \text{for } \boldsymbol{x} \in \mathcal{X}_k,\ t = 2, ..., S\,; \tag{2}$$

such that input dependencies are represented accurately enough by parameterizing the unknown scale discrepancies $\{\xi_{k,t}(\boldsymbol{x})\}$, location discrepancies $\{\delta_{k,t}(\boldsymbol{x})\}$, and output functions $\{y_{k,1}(\boldsymbol{x})\}$ with computationally convenient forms.

We cast $\{\mathcal{X}_k\}_{k=1}^{K}$ as a binary tree partition with rectangular sub-regions $\mathcal{X}_k := \mathcal{X}_k(\mathcal{T})$, for $k = 1, ..., K(\mathcal{T})$, determined by a binary tree $\mathcal{T}$. This specification adds structure to the model for the sake of computational convenience, however it can still provide a reasonable approximation to the reality. Binary treed partitioning has been successfully used in other problems (Denison et al., 1998; Chipman et al., 1998; Gramacy and Lee, 2008; Pratola et al., 2017; Konomi et al., 2017; Karagiannis et al., 2017). To account for the uncertainty about $\mathcal{T}$, we use the binary tree process prior of Chipman et al. (1998) specified as

$$\pi(\mathcal{T}) = P_{\text{rule}}(\rho|v, \mathcal{T}) \prod_{v_i \in \mathcal{I}} P_{\text{split}}(v_i, \mathcal{T}) \prod_{v_j \in \mathcal{E}} (1 - P_{\text{split}}(v_j, \mathcal{T})), \tag{3}$$

where $\mathcal{E}$ denotes the set of external nodes corresponding to sub-regions of the partition

$\{\mathcal{X}_k(\mathcal{T})\}$ and $\mathcal{I}$ denotes the internal nodes. It describes a process where input space is recursively sub-divided by spiting a sub-region, one at a time, into two regions according to a probability low. In particular here, prior tree process $\mathcal{T}$ has origin denoting the whole input space $\mathcal{X}$, while each node $v \in \mathcal{T}$ represents a sub-region of the input space. Each node splits with probability $P_{\text{split}}(v, \mathcal{T}) = \zeta(1 + u_v)^{-d}$ where u_v is the depth of $v \in \mathcal{T}$, ζ controls the balance of the shape of the tree, and d controls the size of the tree. The splits are performed based on a random splitting rule ρ which follows a distribution $P_{\text{rule}}(\rho|v, \mathcal{T}) \propto 1$ specifying that the dimension and the location of the split are drawn independently and randomly.

We specify mutually independent Gaussian processes (GP) priors for $y_{k,1}(\cdot)$, and $\delta_{k,t}(\cdot)$

$$y_{k,1}(\cdot)|\mathcal{T} \sim \text{GP}\left(\mu_1(\cdot|\boldsymbol{\beta}_{k,1}), \sigma^2_{k,1} R_1(\cdot, \cdot|\boldsymbol{\phi}_{k,1})\right); \tag{4}$$

$$\delta_{k,t}(\cdot)|\mathcal{T} \sim \text{GP}\left(\mu_t(\cdot|\boldsymbol{\beta}_{k,t}), \sigma^2_{k,t} R_t(\cdot, \cdot|\boldsymbol{\phi}_{k,t})\right), \text{ for } t = 2, \ldots, S, \tag{5}$$

for $k = 1, ..., K$, to account for their uncertainty. Given a suitable partition $\{\mathcal{X}_k\}_{k=1}^K$ for the model (2), we can use simple and computationally convenient functions to model $\mu_t(\cdot|\boldsymbol{\beta}_{k,t})$, $R_t(\cdot, \cdot|\boldsymbol{\phi}_{k,t})$, and $\xi_{k,t}(\boldsymbol{x})$. The mean functions are parametrized as basis expansions $\mu_t(\cdot|\boldsymbol{\beta}_{k,t}) = \boldsymbol{h}_t(\cdot)^T\boldsymbol{\beta}_{k,t}$, where $\boldsymbol{h}_t(\cdot)$ is a vector of basis functions and $\beta_{k,t}$ are vectors of coefficients, at fidelity level t, and sub-region $\mathcal{X}_k$. The GP mean parameterized as an expansion of basis functions allows modeling long scale variations along x which facilitates the use of computationally convenient correlation functions modeling low scale variations in a similar manner to the standard GP regression as in (Ba and Joseph, 2012; Sang and Huang, 2012). We consider a the family of square exponential correlation functions in separable form $R_t(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\phi}_{k,t}) = \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^\top \text{diag}(\boldsymbol{\phi}_{k,t})(\boldsymbol{x} - \boldsymbol{x}'))$; more sophisticated ones can be used (Williams and Rasmussen, 2006). The unknown functions $\{\xi_{k,t}(\boldsymbol{x})\}$ are modeled as low degree basis expansions $\xi_{k,t}(\boldsymbol{x}|\boldsymbol{\gamma}_{k,t}) = \boldsymbol{w}_t(\boldsymbol{x})^T\boldsymbol{\gamma}_{k,t}$ where $\{\boldsymbol{w}_t(\boldsymbol{x})\}$ are polynomial bases and $\{\boldsymbol{\gamma}_{k,t}\}$ are uncertain coefficients. Modeling $\boldsymbol{\mu}_t(\cdot|\boldsymbol{\beta}_{k,t})$, and $\xi_{k,t}(\boldsymbol{x})$ as basis expansions facilitates the specification of conjugate priors and leads to computational savings given a suitable treatment in the likelihood to be introduced in Section 2.2.

2.2 Conditional conjugacy via augmentation

We do not require the available experimental design to be hierarchically nested, unlike existing co-kriging methods (Kennedy and O'Hagan, 2000; Le Gratiet, 2013). Namely, if $\{\boldsymbol{y}_t, \mathscr{D}_t\}$ denotes the available training data set with output values $y_t \in \mathbb{R}^{n_t}$ at the experimental design $\mathscr{D}_t$ of size n_t at fidelity level $t = 1, .., S$, it may be $\mathscr{D}_{t+1} \not\subseteq \mathscr{D}_t$ for some t. This realistic generalization prevents the direct specification of priors conjugate to the Gaussian likelihood $f(\boldsymbol{y}_{1:S}|\mathcal{T}, \boldsymbol{\sigma}^2_{1:S}, \boldsymbol{\phi}_{1:S}, \boldsymbol{\beta}_{1:S}, \boldsymbol{\gamma}_{1:S-1})$, and hence makes the Bayesian computations prohibitively expensive. In such cases, direct implementation of existing co-kriging methods would require the inversion of large covariance matrices with size $\sum_t n_t \times \sum_t n_t$ for the computation of the likelihood, and possibly the use of Metropolis-Hastings operations in high-dimensional state spaces which would lead to practically infeasible computations. The introduction of the binary partition exacerbates this issue as it increases the dimensionality of the posterior by introducing additional unknown parameters $\boldsymbol{\beta}_{k,t}, \boldsymbol{\gamma}_{k,t}, \sigma^2_{k,t}, \boldsymbol{\phi}_{k,t}$ associated to each sub-region; this necessitates the specification of conjugate priors.

We address this issue by properly imputing the observed data with uncertain quantities, that can be thought of as missing data of a hierarchically nested experimental design, able to induce a conditional independence that enables the specification of conjugate priors, facilitates tractability of posterior marginals and conditionals, and allows the design of efficient MCMC implementations, while it leads to the same Bayesian inference as if we had considered the original data set only.

Augmentation Let $\{\boldsymbol{y}_{k,t}, \mathscr{D}_{k,t}\}$ be the observed data set with output values $\boldsymbol{y}_{k,t} = y_t(\mathscr{D}_{k,t})$ and design $\mathscr{D}_{k,t}$ at sub-region $\mathcal{X}_k$ and fidelity level t. Assume sets of points $\tilde{\mathscr{D}}_{k,t}$ and $\mathring{\mathscr{D}}_{k,t}$ such that $\tilde{\mathscr{D}}_{k,S} = \mathscr{D}_{k,S}$ with $\mathring{\mathscr{D}}_{k,S} = \emptyset$, and $\tilde{\mathscr{D}}_{k,t} = \mathscr{D}_{k,t} \cup \mathring{\mathscr{D}}_{k,t}$ where $\mathring{\mathscr{D}}_{k,t} = \tilde{\mathscr{D}}_{k,t+1} - \mathscr{D}_{k,t}$ is defined as the relative complement of $\mathscr{D}_{k,t}$ in $\tilde{\mathscr{D}}_{k,t+1}$ for $t = S-1, ..., 1$. It is easy to check that $\tilde{\mathscr{D}}_{k,t} = \cup_{j=t}^{S} \mathscr{D}_{k,j}$, and that $\{\tilde{\mathscr{D}}_{k,t}\}_{t=1}^{S}$ is hierarchically nested; i.e. $\tilde{\mathscr{D}}_{k,t} \subseteq \tilde{\mathscr{D}}_{k,t-1}$. By construction, $\{\mathring{\mathscr{D}}_{k,t}\}$ is the smallest collection of sets of input points required to be added to the original design $\{\mathscr{D}_{k,t}\}$ in order to obtain a hierarchically nested experimental design $\{\tilde{\mathscr{D}}_{k,t}\}$. Let $\mathring{\boldsymbol{y}}_{k,t} = y_t(\mathring{\mathscr{D}}_{k,t})$ be the missing output values of the computer model at the corresponding input points in $\mathring{\mathscr{D}}_{k,t}$. We refer to $\{\mathring{\boldsymbol{y}}_{k,t}, \mathring{\mathscr{D}}_{k,t}\}$ as missing data set, and $\{\tilde{\boldsymbol{y}}_{k,t}, \tilde{\mathscr{D}}_{k,t}\}$ as complete data set, where $\tilde{\mathscr{D}}_{k,t}$ is the complete experimental design, and $\tilde{\boldsymbol{y}}_{k,t} = y_t(\tilde{\mathscr{D}}_{k,t})$ are the output model values at input points in $\tilde{\mathscr{D}}_{k,t}$. The number of input points at sub-region $\mathcal{X}_k$ and fidelity level t after augmentation is denoted as $\tilde{n}_{k,t} = |\tilde{\mathscr{D}}_{k,t}|$.

The joint distribution of $\tilde{\boldsymbol{y}} = (\tilde{y}_{k,t})$ given the parameters $(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi})$ is

$$f(\tilde{y}|\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}) = \prod_{k=1}^{K} f_k(\tilde{\boldsymbol{y}}_{k,1}|\boldsymbol{\beta}_{k,1}, \sigma^2_{k,1}, \boldsymbol{\phi}_{k,1}) \prod_{t=2}^{S} f_k(\tilde{\boldsymbol{y}}_{k,t}|\tilde{\boldsymbol{y}}_{k,t-1}, \boldsymbol{\beta}_{k,t}, \gamma_{k,t-1}, \sigma^2_{k,t}, \boldsymbol{\phi}_{k,t}) \quad (6)$$

where each conditional $f_k(\tilde{\boldsymbol{y}}_{k,t}|...)$ is a Gaussian distribution with mean $\xi_{t-1}(\tilde{\mathscr{D}}_{k,t}|\gamma_{k,t-1}) \circ y_{k,t-1}(\tilde{\mathscr{D}}_{k,t}) + \mu_t(\tilde{\mathscr{D}}_{k,t}|\boldsymbol{\beta}_{k,t})$, and covariance $\sigma^2_{k,t} R_t(\tilde{\mathscr{D}}_{k,t}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})$. Here, $\circ$ denotes the Hadamard product. The joint distribution of $\tilde{\boldsymbol{y}}$ can be factorized as in (6) because the proposed augmentation artificially creates a hierarchically nested design which due to the Markovian condition of (2) induces the required conditional independence. The computation of the augmented likelihood (6) is broken down into that of S Gaussian densities requiring the inversion of $\tilde{n}_{k,t} \times \tilde{n}_{k,t}$ covariance matrices. Otherwise, we would be unable to factorize (6) and we would be required to invert a larger covariance matrices with sizes $\sum_t n_t \times \sum_t n_t$.

Priors To account for the uncertainty about unknowns $\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}$, we specify a prior factorized as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}|\mathcal{T}) = \prod_{k=1}^{K} \pi(\boldsymbol{\beta}_{k,1}, \sigma^2_{k,1}|\mathcal{T}) \pi(\boldsymbol{\phi}_{k,1}|\mathcal{T}) \prod_{t=2}^{S} \pi(\boldsymbol{\beta}_{k,t}, \gamma_{k,t-1}, \sigma^2_{k,t}|\mathcal{T}) \pi(\boldsymbol{\phi}_{k,t}|\mathcal{T}). \quad (7)$$

We assign Normal-inverse-Gamma prior distributions on $(\beta, \gamma, \sigma^2)$ such as

$$\begin{aligned} \boldsymbol{\beta}_{k,1}|\mathcal{T}, \sigma^2_{k,1} &\sim \mathrm{N}_{p_1}\left(\boldsymbol{b}_1, \sigma^2_{k,1}\boldsymbol{B}_1\right) ; & \sigma^2_{k,1}|\mathcal{T} &\sim \mathrm{IG}(\lambda_1, \chi_1) ; \\ \boldsymbol{\beta}_{k,t}, \boldsymbol{\gamma}_{k,t-1}|\mathcal{T}, \sigma^2_{k,t} &\sim \mathrm{N}_{p_t+q_{t-1}}\left(\left[\boldsymbol{b}_t, \boldsymbol{g}_{t-1}\right]^\top, \sigma^2_{k,t}\mathrm{diag}\left(\boldsymbol{B}_t, \boldsymbol{G}_{t-1}\right)^\top\right) ; & \sigma^2_{k,t}|\mathcal{T} &\sim \mathrm{IG}(\lambda_t, \chi_t) ; \end{aligned}$$

which are conjugate to the conditionals $f_k(\tilde{\boldsymbol{y}}_{k,t}|...)$ in augmented likelihood (6). This allows the analytic marginalization of the posterior and leads to important computational benefits discussed in Section 2.3. Without augmentation, we would be unable to specify conjugate priors for the actual likelihood, and computations for learning $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2})$ would be impractical. Elicitation of the priors is performed according to (Oakley, 2002; Brynjarsdóttir and O'Hagan, 2014). Weakly informative priors are obtained by adjusting $\boldsymbol{b}_t$, $\boldsymbol{g}_{t-1}$, $\boldsymbol{B}_t^{-1}$ and $\boldsymbol{G}_t^{-1}$ to be close to zero as they place equal amount of prior mass above and below zero in $\mu_t(\cdot)$ and $\xi_t(\cdot)$, and $\lambda_t \to 1 + (p_t + q_{t-1})/2$ for $t = 2, ..., S$, and $\lambda_1 \to 1 + p_1/2$. Here, $\{\pi(\boldsymbol{\phi}_{k,t}|\mathcal{T})\}$ are proper priors chosen by the researcher.

The posterior distribution of ABTCK model is

$$\pi(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \mathring{\boldsymbol{y}}|\boldsymbol{y}) \propto f(\mathring{\boldsymbol{y}}|\boldsymbol{y}, \mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}) f(\boldsymbol{y}|\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}) \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}|\mathcal{T}) \pi(\mathcal{T}) \tag{8}$$

admits the posterior of interest $\pi(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}|\boldsymbol{y})$ as marginal by construction, and hence leads to the same Bayesian analysis.

2.3 Bayesian inference and computations

We design a RJMCMC sampler, targeting the augmented posterior (8), that involves a random permutation scan of blocks updating $[\mathring{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\sigma^2}, \boldsymbol{\gamma}, \mathcal{T}]$, $[\boldsymbol{\phi}, \mathcal{T}|\tilde{\boldsymbol{y}}]$, and $[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}, \mathcal{T}]$. The blocks are collapsed to avoid undesired high Monte Carlo (MC) standard errors due to the originally high-dimensional sampling space (Liu, 1994). The sampler is computationally efficient as it breaks down the inversion of covariance matrices and involves parallel sampling at different sub-regions k and fidelity levels t. Details regarding the MCMC blocks are explained below.

Update $[\mathring{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \mathcal{T}]$ The full conditional posterior of $\mathring{\boldsymbol{y}}_{k,t}$, after integrating out $\boldsymbol{\beta}$'s from the joint posterior (8), is a Normal distribution with mean and covariance matrix

$$\begin{aligned}
\mathring{\boldsymbol{\mu}}_{k,t} =& \mathring{\boldsymbol{\Sigma}}_{k,t} \; \Bigg[\frac{\hat{R}_t^{-1}(\phi_{k,t}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}_{k,t})}{\sigma^2_{k,t}} \; \hat{\mu}_{(t-1)\to t}(\boldsymbol{\phi}_{k,t}, \boldsymbol{\gamma}_{k,t-1}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}_{k,t}) \\
& \qquad + \Xi_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) \; \frac{\hat{R}_{t+1}^{-1}(\boldsymbol{\phi}_{k,t+1}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}^{\otimes}_{k,t+1})}{\sigma^2_{k,t+1}} \; \hat{\mu}_{(t+1)\to t}(\boldsymbol{\phi}_{k,t+1}, \boldsymbol{\gamma}_{k,t}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}^{\otimes}_{k,t+1}) \Bigg] \qquad (9) \\
\mathring{\boldsymbol{\Sigma}}_{k,t} =& \left[\frac{\hat{R}_t^{-1}(\boldsymbol{\phi}_{k,t}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}_{k,t})}{\sigma^2_{k,t}} + \boldsymbol{\Xi}_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) \; \frac{\hat{R}_{t+1}^{-1}(\boldsymbol{\phi}_{k,t+1}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}^{\otimes}_{k,t+1})}{\sigma^2_{k,t+1}} \; \boldsymbol{\Xi}_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) \right]^{-1}
\end{aligned}$$

where $\mathscr{D}^{\otimes}_{k,t+1} := \tilde{\mathscr{D}}_{k,t+1} - \mathring{\mathscr{D}}_{k,t}$ is the relative complement of $\mathring{\mathscr{D}}_{k,t}$ in $\tilde{\mathscr{D}}_{k,t+1}$, $\boldsymbol{\Xi}_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) = \text{diag}(\xi_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}))$, for $k = 1, ..., K$ and $t = 1, ..., S-1$. The functions $\hat{R}_t$, $\hat{\mu}_{(t-1)\to t}$, and $\hat{\mu}_{(t+1)\to t}$ are given in Appendix A. We observe that updating missing data $\mathring{y}_{k,t}$ takes into account information from the lower level $t-1$, same level t, and higher level $t+1$ by interpolating the associated moments. For instance, $\hat{\mu}_{(t-1)\to t}$ (and $\hat{\mu}_{(t+1)\to t}$) provide information about the location of $\mathring{\boldsymbol{y}}_{k,t}$ from levels $t-1$, t (and levels $t+1$, t). It is worth mentioning that (9) can be re-written as a matrix-weighted average of $\hat{\mu}_{(t-1)\to t}$, and scaled $\hat{\mu}_{(t+1)\to t}$; i.e.

$$\begin{aligned}
\mathring{\boldsymbol{\mu}}_{k,t} &= \mathring{\Omega}_{k,t}\left(\phi_{k,t}, \boldsymbol{\gamma}_{k,t}, \sigma^2_{k,t}, \sigma^2_{k,t+1}\right) \; \hat{\mu}_{(t-1)\to t}(\boldsymbol{\phi}_{k,t}, \boldsymbol{\gamma}_{k,t-1}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}_{k,t}) \\
& \qquad + \Big(I - \mathring{\Omega}_{k,t}\left(\phi_{k,t}, \boldsymbol{\gamma}_{k,t}, \sigma^2_{k,t}, \sigma^2_{k,t+1}\right)\Big) \; \boldsymbol{\Xi}_t^{-1}(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) \, \hat{\mu}_{(t+1)\to t}(\boldsymbol{\phi}_{k,t+1}, \boldsymbol{\gamma}_{k,t}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}^{\otimes}_{k,t+1}). \\
\mathring{\Omega}_{k,t}\left(\phi_{k,t}, \boldsymbol{\gamma}_{k,t}, \sigma^2_{k,t}, \sigma^2_{k,t+1}\right) \; &= \; \mathring{\boldsymbol{\Sigma}}_{k,t} \; \frac{\hat{R}_t^{-1}(\boldsymbol{\phi}_{k,t}|\mathring{\mathscr{D}}_{k,t};\mathscr{D}_{k,t})}{\sigma^2_{k,t}}.
\end{aligned}$$

Hence, each update interpolates both across the input space at an individual fidelity level and across the fidelity levels. Simulating $[\mathring{\boldsymbol{y}}_{k,t}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \mathcal{T}]$ can be performed in parallel for k which is a computational benefit, and it can be suppressed if $\mathring{\mathscr{D}}_{k,t} = \emptyset$.

Elaborating further into specific cases of the above imputation, if levels t and $t+1$ do not share any design points at all, at sub-region $\mathcal{X}_k$, i.e., $\mathscr{D}^{\otimes}_{k,t+1} = \emptyset$, then $\hat{R}^{-1}_{t+1}(\boldsymbol{\phi}_{k,t+1}|\mathring{\mathscr{D}}_{k,t};\emptyset) = R^{-1}_{t+1}(\mathring{\mathscr{D}}_{k,t}, \mathring{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t+1})$, and $\hat{\mu}_{(t+1)\to t}(\boldsymbol{\phi}_{k,t+1}, \boldsymbol{\gamma}_{k,t}|\mathring{\mathscr{D}}_{k,t};\emptyset) = y_{k,t+1}(\mathring{\mathscr{D}}_{k,t}) - \boldsymbol{H}_{t+1}(\mathring{\mathscr{D}}_{k,t})\boldsymbol{b}_{t+1}$. This implies that, given weak priors on $\delta_{k,t+1}(\cdot)$ are specified, i.e. $\boldsymbol{b}_{t+1} \to 0$, the update of missing $\mathring{\boldsymbol{y}}_{k,t}$ obtains information from the upper level $t+1$ which entirely relies on the observed output $\boldsymbol{y}_{k,t+1}$ and not from the discrepancy terms $\delta_{k,t+1}(\cdot)$ and $\xi_{k,t}(\cdot)$ of the two levels. If levels t and $t+1$ share design points, $\mathscr{D}^{\otimes}_{k,t+1} \neq \emptyset$, the extra structure of the equations of $\hat{\mu}_{(t+1)\to t}$ and $\hat{R}^{-1}_{t+1}$ in (27) and (26) (see Appendix A) can be interpreted as the factor quantifying the discrepancy between levels t and $t+1$. Finally, we can see that when the correlation between the two levels t and $t+1$, at sub-region $\mathcal{X}_k$, is weak, e.g. $\Xi_t(\mathring{\mathscr{D}}_{k,t}|\boldsymbol{\gamma}_{k,t}) \to 0$, the missing data update resembles the prediction relying only on the information from the current level t. Based on these observations, it may be preferable to consider designs with some overlap at adjacent levels not only for computational convenience but also for modeling reasons. However, a theoretical proof of this statement is out of scope.

Update $[\mathcal{T}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}]$ To update $[\mathcal{T}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}]$, we propose a mixture of the Markov transitions targeting the augmented marginal posterior $\pi(\mathcal{T}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}})$ whose density is proportional to

$$\pi(\tilde{\boldsymbol{y}}, \mathcal{T}, \boldsymbol{\phi}) = \pi(\mathcal{T}) \prod_{k=1}^{K} \pi(\tilde{\boldsymbol{y}}_{k,1}, \boldsymbol{\phi}_{k,1}|\mathcal{T}) \prod_{t=2}^{S} \pi(\tilde{\boldsymbol{y}}_{k,t}, \boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}}_{k,t-1}, \mathcal{T}), \tag{10}$$

$$\pi(\tilde{\boldsymbol{y}}_{k,t}, \boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}}_{k,t-1}, \mathcal{T}) = \pi(\boldsymbol{\phi}_{k,t}) \frac{|\hat{\boldsymbol{A}}_{k,t}(\phi_{k,t})|^{\frac{1}{2}}}{|\boldsymbol{B}_t|^{\frac{1}{2}} |\boldsymbol{G}_t|^{\frac{1}{2}}} \frac{\chi_t^{\lambda_t}}{\pi^{\frac{\tilde{n}_{k,t}}{2}}} \frac{\Gamma(\lambda_t + \frac{\tilde{n}_{k,t}}{2})}{\Gamma(\lambda_t)} \left(\mathrm{SSE}_{k,t}(\boldsymbol{\phi}_{k,t})\right)^{-\lambda_t - \frac{\tilde{n}_{k,t}}{2}} \tag{11}$$

where $\mathrm{SSE}_{k,t}(\boldsymbol{\phi}_{k,t}) = (\tilde{n}_{k,t} + 2\lambda_t - 2)\hat{\sigma}^2_{k,t}(\boldsymbol{\phi}_{k,t})$. Functions $\hat{\sigma}^2_{k,t}$ and $\hat{\boldsymbol{A}}_{k,t}$ are given in (23) and (24) in Appendix A. The Markov transitions are based on the operations change, swap, rotate, and grow & prune, introduced by (Chipman et al., 1998; Gramacy and Lee, 2008). The first three operations are Metropolis-Hastings algorithms (Hastings, 1970) whose implementation is straightforward. The grow & prune operations are local reversible jump

(RJ) transitions and further specification is required.

The grow operation performing a transition from state $(\mathcal{T}, \boldsymbol{\phi})$ to $(\mathcal{T}^*, \boldsymbol{\phi}^*)$ works as follows. We randomly select an external node ω_{j_0} and assume it corresponds to a sub-region $\mathcal{X}_{j_0}$, data set $\{\tilde{\mathscr{D}}_{j_0}, \tilde{\boldsymbol{y}}_{j_0}\}$, and parameters $\boldsymbol{\phi}_{j_0,t}$ though the augmented statistical model. We propose node ω_{j_0} to split into two new child nodes ω_{j_1} and ω_{j_2} according to the splitting rule P_{rule} in prior (7), and we denote the proposed tree as $\mathcal{T}^*$. Nodes ω_{j_1} and ω_{j_2} correspond to disjoint sub-regions $\mathcal{X}_{j_1}$ and $\mathcal{X}_{j_2}$ (with $\mathcal{X}_{j_0} = \mathcal{X}_{j_0} \cup \mathcal{X}_{j_1}$), data sets $\{\tilde{\mathscr{D}}_{j_1,t}, \tilde{\boldsymbol{y}}_{j_1,t}\}$ and $\{\tilde{\mathscr{D}}_{j_2,t}, \tilde{\boldsymbol{y}}_{j_2,t}\}$, and parameters $\boldsymbol{\phi}^*_{j_1,t}$ and $\boldsymbol{\phi}^*_{j_2,t}$, respectively. Randomly, one of the parameters $\boldsymbol{\phi}^*_{j_1,t}$ or $\boldsymbol{\phi}^*_{j_2,t}$ inherits the values from the parent ones; e.g., $\boldsymbol{\phi}^*_{j_1,t} = \boldsymbol{\phi}_{j_0,t}$. The values of the other parameter are proposed by simulating from a probability distribution; e.g., $\boldsymbol{\phi}^*_{j_2,t} \sim Q_t(\cdot)$, such as the corresponding priors. The rest elements of $\boldsymbol{\phi}^*_t$ inherit their values from $\boldsymbol{\phi}_t$. The proposed transition is accepted with probability $\min(1, \Delta)$ where

$$\begin{aligned}\Delta =& \frac{\zeta(1+u_{\omega_{j_0}})^{-d}(1-\zeta(2+u_{\omega_{j_0}})^{-d})^2}{1-\zeta(1+u_{\omega_{j_0}})^{-d}} \frac{|\mathcal{G}|}{|\mathcal{P}^*|} \prod_{t=2}^{S} \frac{\pi(\tilde{\boldsymbol{y}}_{j_1,t}, \boldsymbol{\phi}^*_{j_1} | \tilde{\boldsymbol{y}}_{j_1,t-1}, \mathcal{T}^*)\pi(\tilde{\boldsymbol{y}}_{j_2,t}, \boldsymbol{\phi}^*_{j_2,t} | \tilde{\boldsymbol{y}}_{j_2,t-1}, \mathcal{T}^*)}{\pi(\tilde{\boldsymbol{y}}_{j_0,t}, \boldsymbol{\phi}^*_{j_1,t} | \tilde{\boldsymbol{y}}_{j_0,t-1}, \mathcal{T})Q_t(\boldsymbol{\phi}^*_{j_2,t})} \\ &\times \frac{\pi(\tilde{\boldsymbol{y}}_{j_1,1}, \boldsymbol{\phi}^*_{j_1,1} | \mathcal{T}^*)\pi(\tilde{\boldsymbol{y}}_{j_2,1}, \boldsymbol{\phi}^*_{j_2,1} | \mathcal{T}^*)}{\pi(\tilde{\boldsymbol{y}}_{j_0,1}, \boldsymbol{\phi}^*_{j_1,1} | \mathcal{T})Q_t(\boldsymbol{\phi}^*_{j_2,1})},\end{aligned} \tag{12}$$

$\mathcal{G}$ is the set of growable nodes in tree $\mathcal{T}$, and $\mathcal{P}^*$ is the set of prounable nodes in tree $\mathcal{T}^*$. The prune operation, performing a transition from state $(\mathcal{T}^*, \boldsymbol{\phi}^*)$ to $(\mathcal{T}, \boldsymbol{\phi})$, is fully defined as the reverse operation of the grow operation, and is accepted with probability $\min(1, 1/\Delta)$.

Due to the proposed augmentation in Section 2.2, we are able to analytically integrate out a potentially high-dimensional parameter vector $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2})$ from the joint density (8), and hence design local RJ moves targeting the marginal $\pi(\mathcal{T}, \boldsymbol{\phi} | \tilde{\boldsymbol{y}})$. The benefit from this collapsed update is that the proposed RJ algorithm operates on a lower dimensional state space, which allows for shorter and more acceptable jumps in practice. If necessary, grow and prune operations can be further improved by using the annealing mechanism of Karagiannis and Andrieu (2013).

Update $[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}, \mathcal{T}]$ The conditional posterior $\pi(\beta, \gamma, \sigma^2, \phi|\tilde{y}, \mathcal{T})$ has the form

$$\boldsymbol{\beta}_{k,t}|\tilde{\boldsymbol{y}}_{k,t}, \tilde{\boldsymbol{y}}_{k,t-1}, \boldsymbol{\gamma}_{k,t-1}, \sigma^2_{k,t}, \boldsymbol{\phi}_{k,t} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}_{k,t}(\boldsymbol{\phi}_{k,t}), \hat{\boldsymbol{B}}_{k,t}(\boldsymbol{\phi}_{k,t})\sigma^2_{k,t}), \text{ for } t = 2, ...S \tag{13}$$

$$\boldsymbol{\beta}_{k,1}|\tilde{\boldsymbol{y}}_{k,1}, \sigma^2_{k,1}, \boldsymbol{\phi}_{k,1} \sim \mathrm{N}(\hat{\boldsymbol{\beta}}_{k,1}(\boldsymbol{\phi}_{k,1}), \hat{\boldsymbol{B}}_{k,1}(\boldsymbol{\phi}_{k,1})\sigma^2_{k,1}), \tag{14}$$

$$\boldsymbol{\gamma}_{k,t-1}|\tilde{\boldsymbol{y}}_{k,t}, \tilde{\boldsymbol{y}}_{k,t-1}, \sigma^2_{k,t}, \boldsymbol{\phi}_{k,t} \sim \mathrm{N}(\hat{\boldsymbol{\gamma}}_{k,t-1}(\boldsymbol{\phi}_{k,t}), \hat{\boldsymbol{G}}_{k,t-1}(\boldsymbol{\phi}_{k,t})\sigma^2_{k,t}), \text{ for } t = 2, ...S$$

$$\sigma^2_{k,t}|\tilde{\boldsymbol{y}}_{k,t}, \tilde{\boldsymbol{y}}_{k,t-1}, \boldsymbol{\phi}_{k,t} \sim \mathrm{IG}(\hat{\lambda}_{k,t}, \hat{\chi}_{k,t}(\boldsymbol{\phi}_{k,t})), \text{ for } t = 2, ...S \tag{15}$$

$$\sigma^2_{k,1}|\tilde{\boldsymbol{y}}_{k,1}, \boldsymbol{\phi}_{k,1} \sim \mathrm{IG}(\hat{\lambda}_{k,1}, \hat{\chi}_{k,1}(\boldsymbol{\phi}_{k,1})), \tag{16}$$

$$\boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}}, \mathcal{T} \sim \pi(\boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}}, \mathcal{T}), \tag{17}$$

where the hatted quantities are given in (21)-(23) of Appendix A.

Conditional distributions (13)-(16) can be sampled directly, and in parallel for different (k, t). Sampling from the full conditional of $\boldsymbol{\beta}$'s (13) and (14) is not necessary and can be ignored from the MCMC sweep if prediction is the only concern of the analysis. This is because $\boldsymbol{\beta}$'s can be analytically integrated out from the proposed emulator in Section 2.4. Alternatively, $\boldsymbol{\beta}$'s can be sampled outside the MCMC sweep (13) and (14) by conditioning.

Updating ϕ by simulating from $\pi(\boldsymbol{\phi}|\tilde{\boldsymbol{y}}, \mathcal{T})$ is not necessary in theory, as it is updated in block $[\mathcal{T}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}]$, however it improves mixing in practice. The marginal posterior (17) cannot be sampled directly. Conditional independence in (10) implies that $\{\boldsymbol{\phi}_{k,t}\}$ can be simulated by running in parallel $K \times S$ Metropolis-Hastings algorithms each of them targeting distributions with densities proportional to (11).

2.4 Posterior analysis and emulation

Assume there is available a MCMC sample $\mathcal{S}^N = (\mathring{\boldsymbol{y}}^{(j)}, \mathcal{T}^{(j)}, \boldsymbol{\gamma}^{(j)}, \boldsymbol{\sigma}^{\boldsymbol{2},(j)}, \boldsymbol{\phi}^{(j)})_{j=1}^N$ generated from the RJMCMC sampler in Section 2.3, and let $\{\mathcal{X}_k^{(j)}\}_{k=1}^{K(j)}$ denote the partition corresponding to tree $\mathcal{T}^{(j)}$. Central Limit Theorem can be applied to facilitate inference as the proposed sampler is aperiodic, irreducible, and reversible (Roberts et al., 2004).

The proposed procedure ABTCK allows inference to be performed for the missing output values $\mathring{\boldsymbol{y}}_t = \boldsymbol{y}_t(\mathring{\mathscr{D}}_t)$ at input points in $\mathring{\mathscr{D}}_t = \bigcup_{\forall k} \mathring{\mathscr{D}}_{k,t}$. Inference on $\mathring{\boldsymbol{y}}_t$ can be particularly useful when the computer model has been unable to generate simulations at these input points due to numerical crash or limitations. The marginal posterior distribution of $\mathring{\boldsymbol{y}}_t$, along with its expectations, can be approximated via standard Monte Carlo (MC) using the generated samples $\{\mathring{\boldsymbol{y}}_t^{(j)}\}$ at each level t. Alternatively, point estimates of $\mathring{\boldsymbol{y}}_{k,t}$ at $\mathring{\mathscr{D}}_{k,t}$ can be approximated by the more accurate Rao-Blackwell MC estimator $\mathrm{E}(\mathring{\boldsymbol{y}}_{k,t}|\boldsymbol{y}_{1:S}) \approx \frac{1}{N}\sum_{j=1}^{N} \mathring{\boldsymbol{\mu}}_{k,t}^{(j)}$, where $\{\mathring{\boldsymbol{\mu}}_{k,t}^{(j)}\}$ is the j-th MCMC realization of (9).

A Monte Carlo recursive emulator able to facilitate fully Bayesian predictive inference on the output $y_t(\mathscr{D}^*)$ at untried input points $\mathscr{D}^*$ at every fidelity level $t = 1, ..., S$ can be derived. The conditional distribution $[\boldsymbol{y}_{1:S}(\cdot)|\boldsymbol{y}_{1:S}, \mathring{\boldsymbol{y}}_{1:S}, \boldsymbol{\beta}_{1:S}, \boldsymbol{\gamma}_{1:S}, \boldsymbol{\sigma^2}_{1:S}, \boldsymbol{\phi}_{1:S}]$ inherits a conditional independence similar to (6) due to the augmentation of the data with $\mathring{\boldsymbol{y}}_{1:S}$ that allows it to be analytically integrated out with respect to (13)-(16). Hence the distribution of $[\boldsymbol{y}_{1:S}(\cdot)|\boldsymbol{y}_{k,1:S}, \mathring{\boldsymbol{y}}_{k,1:S}, \boldsymbol{\phi}_{k,1:S}, \mathcal{T}]$, at sub-region $\mathcal{X}_k$, is calculated as

$$y_1(\cdot)|\mathring{\boldsymbol{y}}_1, \boldsymbol{\phi}_1, \mathcal{T} \sim \mathrm{STP}\left(\boldsymbol{\mu}^*_{k,1}(\cdot|\mathring{\boldsymbol{y}}_{k,1}, \boldsymbol{\phi}_{k,1}),\ \hat{\sigma}^2_{k,1}\ R^*_{k,1}(\cdot,\cdot|\mathring{\boldsymbol{y}}_{k,1}, \boldsymbol{\phi}_{k,1}), 2\lambda_1 + \tilde{n}_{k,1}\right); \tag{18}$$

$$y_t(\cdot)|y_{t-1}(\cdot), \mathring{\boldsymbol{y}}_{t:t-1}, \boldsymbol{\phi}_{k,t}, \mathcal{T} \sim \mathrm{STP}\left(\mu^*_{k,t}(\cdot|\mathring{\boldsymbol{y}}_{k,t}, \boldsymbol{\phi}_{k,t}),\ \hat{\sigma}^2_{k,t} R^*_{k,t}(\cdot,\cdot|\mathring{\boldsymbol{y}}_{k,t}, \boldsymbol{\phi}_{k,t}), 2\lambda_t + \tilde{n}_{k,t}\right), \tag{19}$$

where the conditionals are Student-T processes (STP) with

$$\begin{aligned}
\mu^*_t(x|\mathring{\boldsymbol{y}}_{k,t}, \boldsymbol{\phi}_{k,t}) =& \boldsymbol{L}_t(\boldsymbol{x};\boldsymbol{y}_t)\hat{\boldsymbol{a}}_t + \boldsymbol{R}_t(x, \tilde{\mathscr{D}}_t|\boldsymbol{\phi}_{k,t})\boldsymbol{R}_t^{-1}(\tilde{\mathscr{D}}_{k,t}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t}) \left[\boldsymbol{L}_t(\tilde{\mathscr{D}}_{k,t};\boldsymbol{y}_t)\hat{\boldsymbol{a}}_t - \boldsymbol{y}_t(\tilde{\mathscr{D}}_{k,t})\right] \\
R^*_t(\boldsymbol{x}, \boldsymbol{x}'|\boldsymbol{\phi}_{k,t}) =& R_t(x, x'|\boldsymbol{\phi}_{k,t}) - \boldsymbol{R}_t(\boldsymbol{x}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})\boldsymbol{R}_t^{-1}(\tilde{\mathscr{D}}_{k,t}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})R_t^{\top}(\boldsymbol{x}', \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t}) \\
&+ \left[\boldsymbol{L}_t(\boldsymbol{x};\boldsymbol{y}_t) - \boldsymbol{R}_t(\boldsymbol{x}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})\boldsymbol{R}_t^{-1}(\tilde{\mathscr{D}}_{k,t}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})\boldsymbol{L}_t(\tilde{\mathscr{D}}_{k,t};\boldsymbol{y}_t)\right]\hat{\boldsymbol{A}}_t \\
&\times \left[\boldsymbol{L}_t(\boldsymbol{x}';\boldsymbol{y}_t) - \boldsymbol{R}_t(\boldsymbol{x}', \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})\boldsymbol{R}_t^{-1}(\tilde{\mathscr{D}}_{k,t}, \tilde{\mathscr{D}}_{k,t}|\boldsymbol{\phi}_{k,t})\boldsymbol{L}_t(\tilde{\mathscr{D}}_{k,t};\boldsymbol{y}_t)\right]^{\top}
\end{aligned}$$

for $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}_k$, and $\boldsymbol{L}_t(\mathfrak{z};\boldsymbol{y}_{t-1}) = \left[\boldsymbol{H}_t(\mathfrak{z}), \mathrm{diag}(\boldsymbol{y}_{t-1}(\mathfrak{z})\boldsymbol{W}_{t-1}(\mathfrak{z}))\right]$ for $t = 2, ..., S$ and

$\boldsymbol{L}_1(\mathfrak{Z};\cdot) = \boldsymbol{H}_1(\mathfrak{Z})$ for a set $\mathfrak{Z}$. An MCMC sample from the predictive distribution of $[\boldsymbol{y}_{1:S}(\cdot)|\boldsymbol{y}_{1:S}]$, at $\boldsymbol{x} \in \mathscr{D}^*$, can be obtained by simulating (18)-(19) given the sample values $\mathcal{S}^N = \{\mathring{\boldsymbol{y}}^{(j)}, \boldsymbol{\phi}^{(j)}, \mathcal{T}^{(j)}\}$. This allows the computation of a Monte Carlo approximation of the emulator of $[y_t(\cdot)|\boldsymbol{y}_{1:S}]$, and its moments, at any fidelity level t. The conditional independence in the predictive distribution (18) and (19) results because of our imputation strategy.

The proposed emulator accounts for non-stationariy and discontinuity by aggregating simpler GP emulators in a Bayesian model averaging manner, while it integrates uncertainty regarding the unknown 'missing data' $\mathring{\boldsymbol{y}}$ and parameters. It is computationally preferable compared to existing co-kriging one (Kennedy and O'Hagan, 2000; Le Gratiet, 2013) because it allows the parallel inversion of smaller covariance matrices with sizes $\tilde{n}_{t,k} \times \tilde{n}_{t,k}$ while the others require the inversion of a large co-variance matrix of size $\sum_{t=1}^{S} \tilde{n}_t \times \sum_{t=1}^{S} \tilde{n}_t$. Moreover, it is able to recover the whole predictive distribution and its moments, unlike the derivation in Le Gratiet and Garnier (2014) where only the predictive mean and variance are derived recursively. More importantly, it is able to be applied in problems where the training data set is not hierarchically nested, while its competitors cannot.

2.5 Further details

Two novel co-kriging procedures can be distinguished as special cases of the proposed ABTCK. In applications where the design is non hierarchically nested, but the computer model outputs can be assumed as stationary, one can consider to drop the partitioning by setting $K = 1$ and suppressing the MCMC update $[\mathcal{T}, \boldsymbol{\phi}|\tilde{\boldsymbol{y}}]$. We will refer to this reduced version of ABTCK as *Augmented Bayesian co-kriging (ABCK).* Unlike standard co-kriging, our ABCK can be applied with non-nested designs as it makes the computations for training the Bayesian model or computing the emulator practically feasible. In fact, the proposed augmentation strategy separates the posterior into conditionally independent quantities and allows closed form inference for the majority of the hyper-parameters. Another special case is where the design is hierarchically nested but the model outputs present non-stationarity, the imputation mechanism can be dropped by setting $\{\mathring{\mathscr{D}}_{k,t} \equiv \emptyset\}$ and suppressing the update $[\mathring{\boldsymbol{y}}|\boldsymbol{y}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \mathcal{T}]$. We will refer to this reduced version of ABTCK as *Bayesian treed co-kriging (BTCK).* In such a case, BTCK can be preferable to the standard co-kriging as it can model the aforesaid stationarity by properly combining simple stationary GPs.

The computational complexity of the proposed ABTCK compared to existing co-kriging methods is reduced in two ways: a) by breaking the emulation into K parts via the partitioning, and b) by breaking the emulation into S parts via the recursively prediction procedure. In ABTCK the computational complexity of evaluating the augmented likelihood or the predictive distribution is $\mathcal{O}(\sum_{t=1}^{S}\sum_{k=1}^{K}\tilde{n}_{k,t}^{3})$ in sequential computing environments, while it can be further reduced to $\mathcal{O}(\sum_{t=1}^{S}\max_{k=1,...,K}(\tilde{n}_{k,t})^{3})$ in parallel computing environments since operations at each k can be performed in parallel. Under non-hierarchical designs, our ABCK (assuming the partitioning is dropped) requires $\mathcal{O}(\sum_{t=1}^{S}\tilde{n}_{t}^{3})$ for the evaluation of the augmented likelihood or the Monte Carlo emulator which is smaller than $\mathcal{O}((\sum_{t=1}^{S}n_{t})^{3})$ required by (Kennedy and O'Hagan, 2000; Le Gratiet et al., 2014) for the evaluation of the associated likelihoods since $\tilde{n}_t \leq n_t$.

3 Case study

We examine the predictive ability of the proposed method, refereed to as augmented Bayesian treed co-kriging (ABTCK), its special cases ABCK, and BTCK, as well as we provide comparisons with existing approaches of (Le Gratiet, 2013; Kennedy and O'Hagan, 2000) showing the good performance of the proposed method. Details about the performance measures used can be found in Supplementary Section S.3. For the simulations, we used in MATLAB R2017b on a computer with specifications (IntelCore™i7-7700K CPU @ 4.20GHz × 8, and 62.8 GiBRAM).

3.1 Numerical example

Consider functions

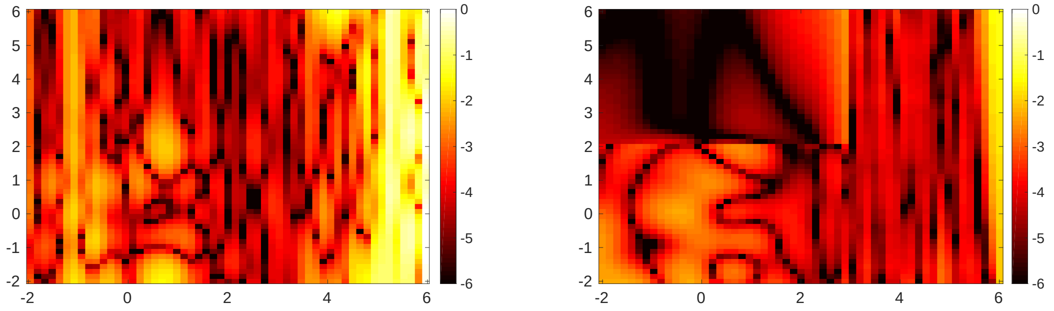
$$\begin{aligned} y_1(\boldsymbol{x}) &= 2x_1\exp(-x_1^2 - x_2^2) + 0.5\exp\{\sin((0.9(\tfrac{x_1+2}{8} + 0.48)^{10}))\} + 1.2,\ \boldsymbol{x} \in [-2,6]^2; \\ y_2(\boldsymbol{x}) &= 4x_1\exp(-x_1^2 - x_2^2) + 0.2\exp\{\sin((0.9(\tfrac{x_1+2}{8} + 0.48)^{10}))\} + 0.5,\ \boldsymbol{x} \in [-2,6]^2, \end{aligned} \tag{20}$$

which are assumed to be output functions of computer models $\mathfrak{C}_1$ and $\mathfrak{C}_2$, with $\mathfrak{C}_2$ being more accurate but slower to run than $\mathfrak{C}_1$. By expressing (20) as (1), it can be seen that the discrepancy functions $\delta_1(\cdot)$ and $\xi_1(\boldsymbol{x})$ change over $\mathcal{X}$. We pretend that equations in (20) are unknown, and we are interested in learning the high fidelity $y_2(\cdot)$.

We consider a non-hierarchically nested design $\mathscr{D} = \{\mathscr{D}_1, \mathscr{D}_2\}$, generated as follows. For level $t = 1$, the observed data are generated by employing a Latin Hypercube Sampling (LHS) (McKay et al., 1979) to generate a point set $\mathscr{D}_1$ of size $n_1 = 120$, and computing the corresponding observations $\boldsymbol{y}_1$ from (20). For level $t = 2$, the observations are generated likewise by generating a point set $\mathscr{D}_2$ of size $n_2 = 30$ via LHS such that $\mathscr{D}_2 \not\subseteq \mathscr{D}_1$. For our comparisons against, we consider a second hierarchically nested experimental design $\mathscr{D}' = \{\mathscr{D}_1, \mathscr{D}_2'\}$, where the high fidelity point set $\mathscr{D}_2'$ is randomly generated via the condition Latin Hypercube Sampling (cLHS) design (Minasny and McBratney, 2006).

To implement our ABTCK, we consider weakly informative priors with hyper-parameters $\boldsymbol{b}_t = \boldsymbol{g}_t = 0$, $\boldsymbol{B}_t = 10$, $\lambda_t = 2$, $\chi_t = 2$ and a mixture prior of Gamma distributions $\boldsymbol{\phi}_t|\mathcal{T} \sim 0.5\text{G}(1, 20) + 0.5\text{G}(10, 10)$ for $\boldsymbol{\phi}_t$ distributing the prior mass on areas of smaller and larger values (Gramacy and Lee, 2008). The scale discrepancy is parametrised as a zero-degree basis expansion $\xi_{k,t}(\boldsymbol{x}|\boldsymbol{\gamma}_{k,t}) = \boldsymbol{\gamma}_{k,t}$. The tree process prior has hyper-parameters $\zeta = 0.5$ and $d = 2$. The statistical model was trained by MCMC running for 25000 iterations where the first 5000 iterations where discarded as as burn-in. Also, we consider ABCK which is a special case of ABTCK where binary partitioning is suspended. Furthermore, for comparisons involve the existing approaches: the standard GP considering only the high fidelity data on $\mathscr{D}_2$ (HFGP), Zertuche's co-kriging approach (ZBCK) in Zertuche (2015) against non-nested design $\mathscr{D}$, Bayesian co-kriging on nested design $\mathscr{D}'$ (NBCK), and the Kenedy & O'Hagan's CK (K&O) approach on non-nested design $\mathscr{D}$.

Figure 1 present the relative absolute error (RAE), on a 100 × 100 grid of $\mathcal{X}$, produced by ABTCK and ABCK (where partitioning is suppressed). We observe that ABTCK has produced a significantly smaller RAE than ABCK suggesting the benefit in predictions from including the binary partitioning mechanism in our proposed procedure.



(a) ABCK

(b) ABTCK

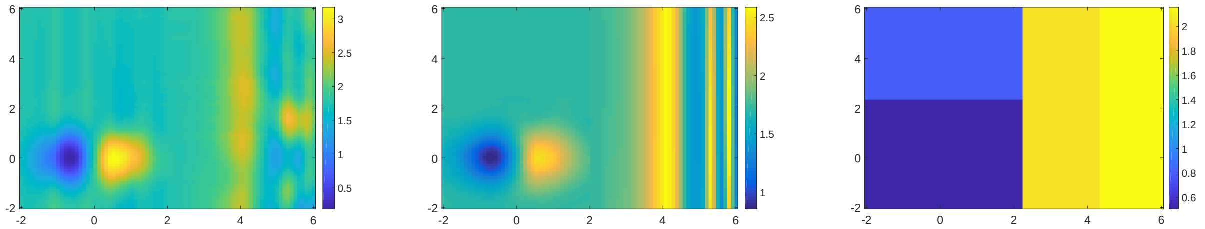
Figure 1: Absolute relative error (in log scale) between the real response $y_2(\cdot)$ and the predictive mean of the high-level computer model using the augmented Bayesian co-kriging (ABCK) and augmented Bayesian treed co-kriging (ABTCK).

Table 1 presents the mean squared predictive error (MSPE) on a 100 × 100 grid of $\mathcal{X}$ produced by the procedures under comparison, as well as the model fitting times. We observe that MSPE produced by ABTCK is around 40 times smaller than that by ABCK suggesting that the partitioning mechanism as implemented in our ABTCK has been able to successfully capture and model the non-stationarity, and hence produce more accurate predictions, in the multifidelity setting. Our ABTCK and ABCK produced smaller MSPE than ZCK which is reasonable as the latter introduced a bias while dealing with non-nested designs. Moreover ABTCK and ABCK on the non-nested design $\mathscr{D}$ produced smaller MSPEs than NBCK and K&O on nested designs $\mathscr{D}'$. This suggests the possibility that non-nested designs can lead to more accurate emulations (at the expense of execution time) and hence the limitation of approaches requiring only nested designs. We believe that this may happen because non-nested designs allow the procedure learn from different input locations at different fidelity levels, however a more detailed examination of this phenomenon is out of the scope of this study. Finally, we observe that ABTCK presented smaller model fitting time than the other co-kriging approaches. This is the result of the partitioning in ABTCK that requires inversion of smaller covariance matrices than other methods. In fact, the computational overhead introduced by the RJ operation is dominated by the computational gain due to the partition and subsequent inversion of smaller matrices.

	HFGP	NBCK	K&O	ZBCK	ABCK	ABTCK
MSPE	0.107	0.0602	0.0550	0.0339	0.0221	0.0006
Time (sec)	71.12	314.11	738.74	484.86	500.75	220.84

Table 1: The computed MSPE and the corresponding model fitting time in seconds using: the High Fidelity Gaussian process (HFGP), nested Bayesian Co-kriging with nested design (NBCK), Zertuche's Bayesian Co-kriging (ZBCK), Augmented Bayesian Co-kriging (ABCK), Augmented Bayesian Treed Co-kriging (ABTCK).

In Figures 2a and 2b, we can see the that predictive mean of high fidelity response $y_2(\boldsymbol{x})$ by ABTCK h predictive mean of high fidelity response $y_2(\boldsymbol{x})$ for ABCK and ABTCKas been able to emulate the sinusoidal dependence of $y_2(\boldsymbol{x})$ on the left hand side more accurately than ABCK. The Monte Carlo approximation of the posterior mean of the scalar discrepancy $\hat{\xi}(\boldsymbol{x}) \approx \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{w}_t(\boldsymbol{x})^T \left(\sum_{k=1}^{K^{(i)}} 1(\boldsymbol{x})\hat{\gamma}_{k,t}(\boldsymbol{\phi}_{k,t}^{(j)})\right)$ produced by the ABTCK is presented in Figure 2c. We observe that ABTCK has recovered a representation of the scalar discrepancy which suggests that $\xi(\boldsymbol{x})$ changes value. In contrast, ABCK produces a posterior scalar discrepancy which is equal to 0.525 and constant throughout the input space due to the lack of partitioning.



(a) $\mathrm{E}(y_2(x)|\mathscr{D}_1, \mathscr{D}_2)$ by ABCK

(b) $\mathrm{E}(y_2(x)|\mathscr{D}_1, \mathscr{D}_2)$ by ABTCK

(c) $\mathrm{E}(\xi_1(x)|\mathscr{D}_1, \mathscr{D}_2)$ by ; ABTCK

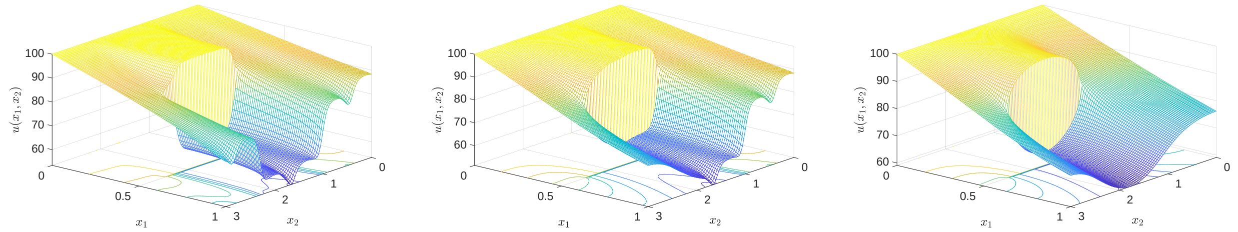
Figure 2: The predictive mean of $y_2(\boldsymbol{x})$ produced from ABCK in (a), predictive mean of $y_2(\boldsymbol{x})$ produced from ABTCK in (b), and posterior mean of the scalar discrepancy $\xi_1(\boldsymbol{x})$ between low and high fidelity computer models from ABTCK in (c). ABCK produced a posterior mean for $\xi_1(\boldsymbol{x})$ around 0.525.

3.2 Heat transfer benchmark example

We consider the benchmark problem of a heated metal block of size $\mathcal{X} = [0, 1] \times [0, 3]$ with a rectangular cavity of size $[0.5, 0.015] \times [1, 2.5]$ where the temperature $u(\boldsymbol{x})$ is modeled as an elliptic partial differential equation. Let us consider 2D elliptic PDEs $-\nabla \cdot c^{(j)}(\boldsymbol{x}) \nabla u^{(j)}(\boldsymbol{x}) = f(\boldsymbol{x})$ with $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{x} \in \mathcal{X} - \partial\mathcal{X}$, for $j = 1, 2, 3$. The left side of the block is heated to 100 degrees and hence we consider Dirichlet condition $u = 100$. At the right side of the metal block, heat is flowing from the block to the surrounding air at a constant rate and we assume Neumann condition $u'(\boldsymbol{x}) = -20$. The rest boundary conditions are Neumann condition $u'(\boldsymbol{x}) = 0$. The internal heat source is $f(\boldsymbol{x}) = 1$.

Assume that there are three computer models aiming at describing the steady state of the temperature, and they are arranged in ascending order of fidelity as $\{\mathfrak{C}^{(t)}\}_{t=1}^{3}$. The spatial dependent thermal connectivity is denoted as $c^{(j)}(\boldsymbol{x})$; it is $c^{(1)}(\boldsymbol{x}) = 1$ for the least accurate computer model, $c^{(2)}(\boldsymbol{x}) = \exp(1.5 \sin(3.33\pi x_2))1(x_2 < 1.8)$ for more accurate computer model, and $c^{(3)}(\boldsymbol{x}) = \exp(1.5 \sin(3.33\pi x_2))$ for most accurate computer model. The PDE is solved via a FEM solver with the domain $\mathcal{X}$ discretized in 24119 nodes. The temperature produced by the three computer models is presented in Figures 3a, 3b, and 3c.

There is an obvious discontinuity at $x_1 = 0.5$. The accurate model $\mathfrak{C}^{(3)}$ has high frequen-



(a) Fine grid

(b) Intermidiate grid

(c) Corse grid

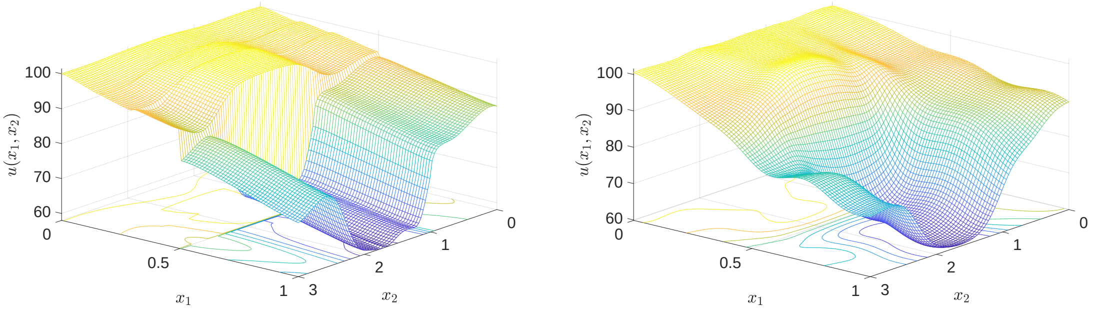
Figure 3: Response surface for the temperature, steady state solution at three levels of fidelity (fine, intermediate, and coarse).

cies not captured by the lower fidelity models $\mathfrak{C}^{(1)}$ and $\mathfrak{C}^{(2)}$. The discrepancy function δ_2 varies throughout the input space, and presents discontinuity.

We compare our proposed approach against the the existing co-kriging approach of Le Gratiet (2013) referred to as GLCK. As the latter approach requires hierarchically nested designs, we generated a nested experimental design for models $\{\mathfrak{C}^{(t)}\}_{t=1}^{3}$ according to the condition Latin Hypercube Sampling (cLHS) design (Minasny and McBratney, 2006) with sample sizes $n^{(1)} = 150, n^{(2)} = 100$ and $n^{(3)} = 50$. For the same reason, we implement our BTCK; the special case of ABTCK where augmentation is suppressed as not needed in nested designs. We assume a prior $\boldsymbol{\phi}_t|\mathcal{T} \sim 0.5\text{G}(1, 20) + 0.5\text{G}(10, 10)$ on $\boldsymbol{\phi}_t$. The model was trained by running the suggested MCMC sampler for 25000 iterations and discarding the first 5000 values as burn in. For or GLCK we used similar particularization.

In Figures 4a and 4b, we present the prediction of the high fidelity model output for the proposed BTCK and the competitor in a 100 × 100 grid.

We observe that BTCK managed to adequately capture the discontinuity and the smaller scale variations in the output while the competitor failed. We speculate that the behavior of the surface produced by the competitor in Figure 4b is because the basis expansion is unable to represent efficiently sudden changes. The proposed BTCK produced a significantly smaller MSPE equal to 1.4613 compared to the competitor whose MSPE was 14.1599; hence BTCK has being able to produce more accurate predictions. Figures 5a and 5b demonstrate the

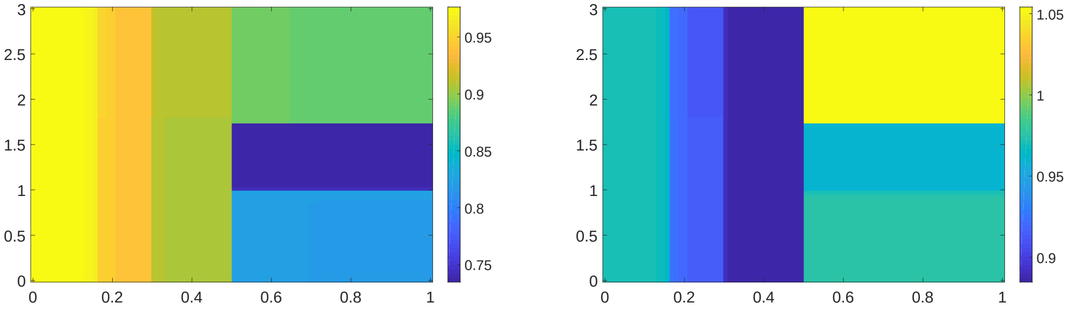


(a) Bayesian treed co-kriging

(b) GP co-kriging

Figure 4: Prediction mean of the temperature steady state solution for the fine computer model using two different methods (a) co-kriging GP and (b) proposed Bayesian treed co-kriging.

estimation of the scale discrepancy function between computer models $\mathfrak{C}^{(1)}$ vs. $\mathfrak{C}^{(2)}$ and $\mathfrak{C}^{(2)}$ vs. $\mathfrak{C}^{(3)}$ respectively, as produced by the proposed ABTCK.



(a) Scalar discrepancy $\xi_1(x)$

(b) Scalar discrepancy $\xi_2(x)$

Figure 5: Posterior mean of the scale discrepancies between $\mathfrak{C}^{(1)}$ vs. $\mathfrak{C}^{(2)}$ in (a) and $\mathfrak{C}^{(2)}$ vs. $\mathfrak{C}^{(3)}$ in (b) using the proposed Bayesian treed co-kriging.

4 Application to large-scale climate modeling

We consider the Advanced Research Weather Research and Forecasting Version 3.2.1 (WRF Version 3.2.1) climate model (Skamarock et al., 2008) constrained in the geographical domain 25°–44°N and 112°–90°W over the Southern Great Plains (SGP) region, and we concentrate on the average precipitation response over the area.

WRF is employed with the Morrison 2-moment cloud microphysics scheme (Morrison et al., 2005) and the Kain-Fritsch convective parametrisation scheme (KF CPS) (Kain, 2004) as in (Yang et al., 2012). The most critical parameters (Yang et al., 2012; Yan et al., 2014) of the KF scheme are: the coefficient related to downdraft mass flux rate P_{d} that takes values in range $[-1, 1]$; the coefficient related to entrainment mass flux rate P_{e} that takes values in range $[-1, 1]$; the maximum turbulent kinetic energy in sub-cloud layer $(m^2 s^{-2})$ P_{t} that takes values in range $[3, 12]$; the starting height of downdraft above updraft source layer (hPa) P_{h} that takes values in range $[50, 350]$; and the average consumption time of convective available potential energy P_{c} that takes values in range $[900, 7200]$. The ranges of the KF CPS parameters are quite wide and cause higher uncertainties in climate simulations due to the non linear interactions and compensating errors of the parameters (Gilmore et al., 2004; Murphy et al., 2007; Yang et al., 2012). We consider the Rapid Radiative Transfer Model (RRTMG) for General Circulation Models (Mlawer et al., 1997) as a more accurate radiation scheme for the geological domain of interest. We are interested in modeling the average precipitation with respect to the five KF CPS parameters.

The available simulations have been generated by running WRF model 240 times at two resolution levels; 90 model runs for 12.5km grid spacing and 150 model runs 25km grid spacing. The fidelity of the simulations increases when the grid spacing gets finer. The available data set has been generated via a simulated stochastic approximation annealing (SSAA) calibration algorithm in (Yan et al., 2014). As the SSAA procedure progresses, the sampling range of each parameter gradually narrows. Different resolutions correspond to different narrowing range on the input space. This has produced simulations on a non hierarchically nested design at the five input parameters (see Supplemental Section S.4). Due to the high cost, it is not possible to re-run the expensive WRF model in order to generate simulations based on a hierarchically nested design as existing co-kriging methods require. As discussed in (Yang et al., 2012; Yan et al., 2014) the discrepancies between the two fidelity levels may depend on the five inputs, however no formal statistical analysis have been performed. The atmospheric humidity at all levels is lower in the fine resolution than coarse resolution, and the drier atmosphere may result from more condensation (so more precipitation generated) which consumes more moisture at the finer resolution. The explicit precipitation increases with spatial resolution because more clouds are resolved at finer resolution. Moreover, interest lies in better understanding how different grid spacing affects the discrepancies in WRF with respect to the input parameters.

We implement our proposed ABTCK approach to analyze the data set. For the GP priors, we use separable square exponential covariance functions. On the correlation parameters we assign Gamma mixture priors $\boldsymbol{\phi}_{k,t} \sim 0.5\text{G}(1,10)+0.5\text{G}(5,2)$ distributing the mass on areas of smaller and larger values; for the binary treed partition priors, we consider hyper-parameters $a = 0.8$ and $b = 5$; and for the rest parameters we consider weak informative priors as $\boldsymbol{b}_t = 0$, $\boldsymbol{B}_t = 100$, $\lambda_t = 0.2$, and $\chi_t = 0.2$. Regarding the grow & prune update, we use the prior distributions as the dimensional matching proposals $\boldsymbol{\phi}_{k,t} \sim 0.5\text{G}(1,10)+0.5\text{G}(5,2)$. We have re-scaled the input space for the five parameters to be between $[-1, 1]$ in order to be able to use the same proposal distribution for all $\boldsymbol{\phi}_{k,t}$'s. To train the model, we run the MCMC sampler for $30,000$ iterations from which we discard 5000 as burn in. To demonstrate the necessity of the binary partitioning mechanism, we also consider our ABCK which is the special case of ABTCK where partitioning is suppressed. Moreover, we consider the standard Gaussian process emulator (HFGP) trained against the high fidelity data set only to demonstrate the importance of using co-kriging in multi-fidelity problems even under non-hierarchically nested designs. Existing co-kriging techniques cannot be implemented of this application because the available experimental design is not hierarchically nested.

We study the performance of the proposed approach by performing two separate k-fold cross validation (CV) studies; the 20%-CV leaves the 20% of the available data as a test set and keeps the rest as a training test, while 50%-CV leaves the 50% of the data as a test set and keeps the rest as a training test. For the comparisons, as performance measures, we used the MSPE, the coverage probability of the 95% equal-tail credible interval 95%-CVG, the Nash-Sutcliffe model efficiency coefficient (NSME), and the computational time in sec; for details see Supplementary Section S.3. Their values are the averages of 60 realizations performed by re-running the procedures for variance reduction purposes. The results are presented in Table 2.

We observe that both ABCK and ABTCK outperform the HFGP, by far, in terms of accuracy and constructing more accurate credible intervals. This is reasonable as, unlike HFGP, our ABCK and ABTCK are co-kriging approaches taking into account lower fidelity data able to provide important information from locations at which we have no high fidelity observations. This demonstrates the importance of enabling co-kriging approaches to use non-hierarchinal nested designs via the proposed augmentation mechanism. We observe that that ABTCK which involves the partitioning mechanism, taking into account non-stationarity, outperformed ABTCK with respect to all measures. The partitioning mechanism in ABTCK caused an improvement about 20% compared to ABTCK with respect to MSPE, indicating higher accuracy. NSME produced by ABTCK is closer to one than that of ABTCK and HFGP, suggesting that ABTCK has been more accurate than the other two. Figure 6 presents the variation of MSPE from the 60 realizations of the approaches under comparison. It shows that GP is clearly outperformed from our ABCK, and ABTCK, while MSPE of ABTCK presents a smaller variation than that of ABCK as the former can captures non-stationarities. Based on 95%-CVG, ABTCK produced the best representation of the uncertainty than ABCK and HFGP, as well as mode accurate credible sets. Not only the ABTCK produced more accurate predictions but also it gave a better picture of the uncertainty associated with these predictions. By running ABTCK, we observed that the average number of the generated sub-regions varies from 2 to 5; this evidence supports the use of

ABTCK instead of ABCK in this application by the inclusion of the binary partition hence the use of a non-stationary process via partitioning. Finally, it is important to notice that the computational time in ABTCK is approximately two-third of the computational time in ABCK. This means that the improvements on the prediction and uncertainty described above come at a lower computational cost. It is worth noticing that we can further reduce the computational cost of ABTCK if we run it in a parallel computing environment.

	20%-leave out CV				50%-leave out CV			
	MSPE	95%-CVG	NSME	Time	MSPE	95%-CVG	NSME	Time
HFGP	0.1501	0.691	0.61	418	0.2118	0.613	0.31	368
ABCK	0.1002	0.862	0.84	2404	0.1205	0.840	0.79	1804
ABTCK	0.0701	0.954	0.91	1602	0.0974	0.945	0.87	1240

Table 2: Performance measures of High Fidelity Gaussian process, Augmented Bayesian Co-kriging, Augmented Bayesian Treed Co-kriging. The values are the resulted averages after re-running each approach 60 times. Time is measured in sec.

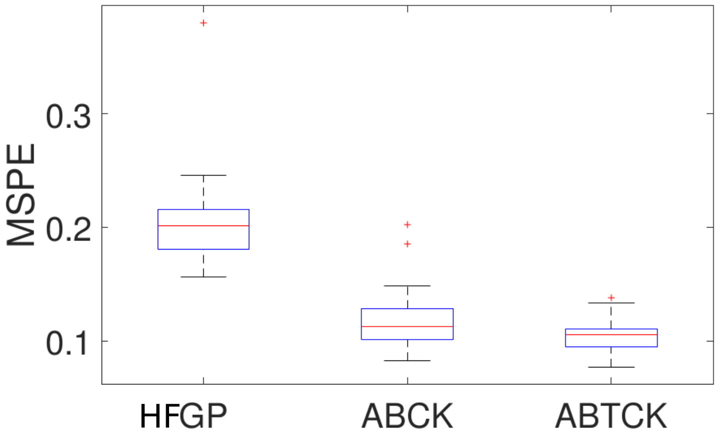
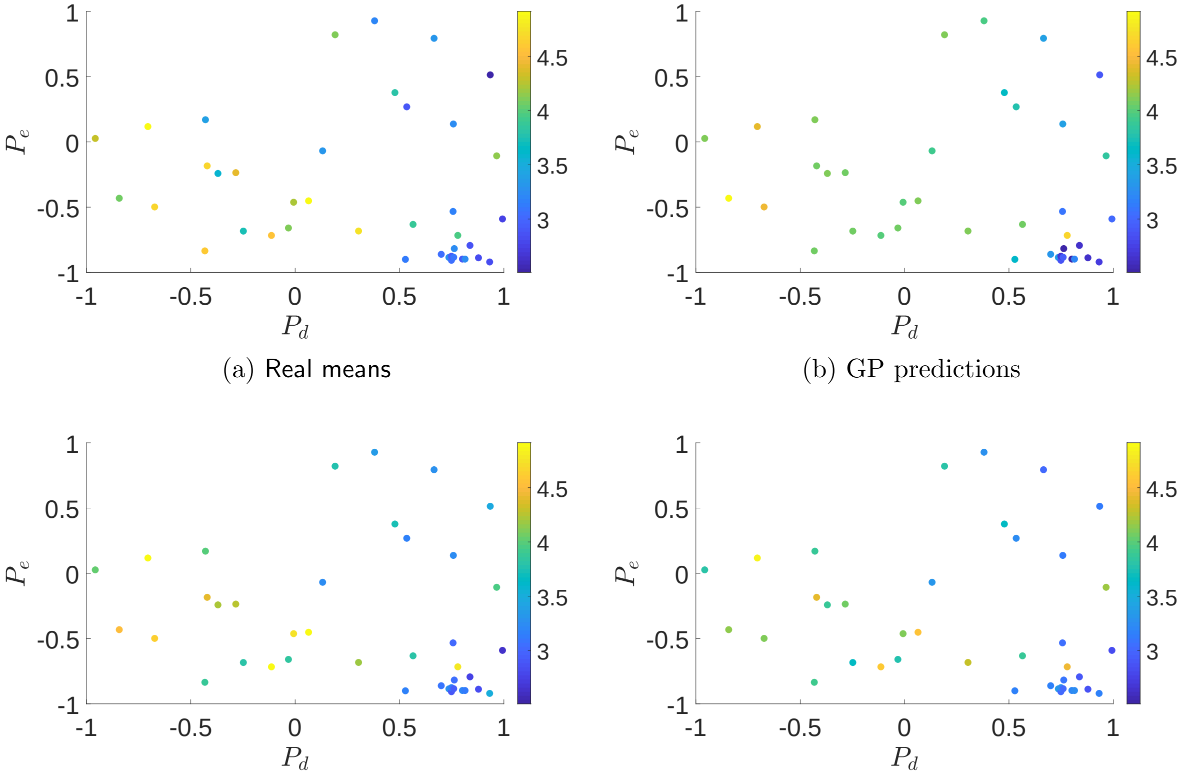


Figure 6: Boxplot of the MSPE for three different procedures

Figure 7 presents the simulated average precipitation from WRF at high fidelity, as well as the predicted average precipitation produced from ABTCK, ABCK, and HFGP against the downdraft mass flux rate P_{d} and entrainment mass flux rate coefficient P_{e}. This is the case corresponding to realizations with the highest MSPE differences between ABTCK and HFGP. We can observe that the HFGP have not been able to capture the variation in the



(a) Real means

(b) GP predictions

(c) ABCK predictions

(d) ABTCK predictions

Figure 7: Real output mean precipitation values and their corresponding predictions for different procedures.

central part of the input domain where high fidelity observations are sparse. Both ABCK and ABTCK were able to capture that variation by taking into account low fidelity data. We observe that ABCK produced a smoother representation of the precipitation, however ABTCK was able to represent local features more accurately. Finally, prediction is much improved over the whole test data-set even in the small clustered range.

5 Conclusions and further work

We proposed a new method, called Augmented Bayesian Treed Autoregressive Co-Kriging, which extends the scope of the co-kriging methods. First our procedure can be implemented in problems where the experimental design is not necessarily hierarchically nested in a principled manner while keeping the computational demands low. This overcomes the difficulty of existing co-kriging methods which require hierarchically nested designs in order to keep the computations practically feasible. Secondly, our method can account for non-stationarity, and potential discontinuity, in the output of the computer models without the need to specify complicated or problem specific GP priors, in the multifidelity setting. Moreover, we proposed a Monte Carlo recursive emulator which can recover the predictive distribution of the computer model output at every level, and can be used with non-hierarchically nested designs as well, while keeping the computational cost lower than the existing emulators as it requires operations with smaller matrices. Finally, for the computations, we designed an efficient trans-dimensional MCMC method tailored to the proposed model.

We analyzed the Weather Research and Forecasting (WRF) simulator using the Kain-Fritsch convective parametrisation scheme by using our novel procedure. This is a large-scale climate modeling application where the available simulations are performed at different fidelity levels at non hierarchically nested designs, and hence we used our method to build a Bayesian multifidelity emulator on WRF. Our method discovered non-stationarity in the WRF output precipitation with respect to the KFC input parameters. We observed that the use of Bayesian treed partition in the co-kriging framework as utilized in our method is able to provide more accurate predictions than ignoring it. In the WRF application we observed the use of the partition was able to reduce the MSPE around 21% on average when we compared the ABTCK with the ABCK where the partitioning was dropped out. In our simulation example considering non-nested designs, we observed that the augmentation mechanism was able to recover the model output accurately.

The procedure can be modified to involve a basis selection mechanism for $\boldsymbol{h}_t(\cdot)$ of $\delta_t(\boldsymbol{x})$ and $\boldsymbol{w}_t(\cdot)$ of $\xi_t(\cdot)$ at different input sub-regions $\mathcal{X}_{k,t}$, by properly specifying spike-and-slab priors on $\boldsymbol{\beta}_{k,t}$ and $\boldsymbol{\gamma}_{k,t}$ and calculating Gibbs updates. One can use the fixed hyper-parameters of the latent treed process $\pi(\mathcal{T})$ to control or mitigate possible non-identifiability between the discrepancy functions, while setting input invariant scale discrepancy $\xi_{k,t}(\boldsymbol{x}) = \gamma_{k,t}$ and meaningful priors on the location discrepancy $\delta_{k,t}(\cdot)$ in the sense of (Brynjarsdóttir and O'Hagan, 2014). The rational is that the treed prior can act as a penalty favoring simpler partitions, which can mitigate the competition between the two discrepancies. An extension of ABTCK would be to specify different partitions for $\xi_t(\boldsymbol{x})$, $\delta_t(\boldsymbol{x})$, $y_1(\boldsymbol{x})$, which may lead to a more flexible model, however, research has to be done on whether the conditional posteriors can be analytically marginalized to keep the computational demands feasible. The authors are currently working on a sequential design procedure with multifidelity simulations that take into account non-hierarchically nested designs.

Supplementary Materials

Supplementary material: It contains Schematics & algorithms of the proposed approach in Section S.1; a list of the performance metrics used in the numerical examples in Section S.2, an additional toy example in Section S.3, additional plots referenced in the application section of the main paper in Section S.4, and information of the codes and data sets used in Section S.5.

Code & data: It contains the data-sets and code used to produce the numerical results in the paper (Link: https://github.com/georgios-stats/ABTCK)

References

Ba, S. and Joseph, V. R. (2012), "Composite Gaussian process models for emulating expensive functions," *The Annals of Applied Statistics*, 6, 1838–1860.

Brynjarsdóttir, J. and O'Hagan, A. (2014), "Learning about physical parameters: The importance of model discrepancy," *Inverse problems*, 30, 114007.

Chipman, H., George, E., and McCulloch, R. (1998), "Bayesian CART Model Search," *Journal of the American Statistical Association*, 93, 935–960.

Denison, D., Mallick, B., and Smith, A. (1998), "A Bayesian CART Algorithm," *Biometrika*, 85, 363–377.

Gilmore, M. S., Straka, J. M., and Rasmussen, E. N. (2004), "Precipitation uncertainty due to variations in precipitation particle parameters within a simple microphysics scheme," *Monthly weather review*, 132, 2610–2627.

Gramacy, R. B. and Lee, H. K. H. (2008), "Bayesian treed Gaussian process Models with an application to computer modeling," *Journal of the American Statistical Association*, 103, 1119–1130.

Hastings, W. K. (1970), "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57, 97–109.

Kain, J. S. (2004), "The Kain-Fritsch convective parameterization: an update," *Journal of Applied Meteorology*, 43, 170–181.

Karagiannis, G. and Andrieu, C. (2013), "Annealed importance sampling reversible jump MCMC algorithms," *Journal of Computational and Graphical Statistics*, 22, 623–648.

Karagiannis, G., Konomi, B. A., and Lin, G. (2017), "On the Bayesian calibration of expensive computer models with input dependent parameters," *Spatial Statistics*.

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

Konomi, B. A., Karagiannis, G., Lai, K., and Lin, G. (2017), "Bayesian Treed Calibration: An Application to Carbon Capture With AX Sorbent," *Journal of the American Statistical Association*, 112, 37–53.

Le Gratiet, L. (2013), "Bayesian analysis of hierarchical multifidelity codes," *SIAM/ASA Journal on Uncertainty Quantification*, 1, 244–269.

Le Gratiet, L., Cannamela, C., and Iooss, B. (2014), "A Bayesian approach for global sensitivity analysis of (multifidelity) computer codes," *SIAM/ASA Journal on Uncertainty Quantification*, 2, 336–363.

Le Gratiet, L. and Garnier, J. (2014), "Recursive co-kriging model for design of computer experiments with multiple levels of fidelity," *International Journal for Uncertainty Quantification*, 4.

Lindgren, F., Rue, H., and Lindström, J. (2011), "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach," *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 74, 423–498.

Liu, J. S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *Journal of the American Statistical Association*, 89, 958–966.

McKay, M., Beckman, R., and Conover, W. (1979), "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, 21, 239–245.

Minasny, B. and McBratney, A. B. (2006), "A conditioned Latin hypercube method for sampling in the presence of ancillary information," *Computers & geosciences*, 32, 1378–1388.

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., and Clough, S. A. (1997), "Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave," *Journal of Geophysical Research: Atmospheres (1984–2012)*, 102, 16663–16682.

Morrison, H., Curry, J., and Khvorostyanov, V. (2005), "A new double-moment microphysics parameterization for application in cloud and climate models. Part I: Description," *Journal of the Atmospheric Sciences*, 62, 1665–1677.

Murphy, J. M., Booth, B. B., Collins, M., Harris, G. R., Sexton, D. M., and Webb, M. J. (2007), "A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 365, 1993–2028.

Oakley, J. (2002), "Eliciting Gaussian process priors for complex computer codes," *Journal of the Royal Statistical Society: Series D (The Statistician)*, 51, 81–97.

OH́agan, A. (1998), "A Markov property for covariance structures," *Statistics Research Report*, 98, 13.

Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N., and Karniadakis, G. E. (2017), "Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473, 20160751.

Perdikaris, P., Venturi, D., Royset, J., and Karniadakis, G. (2015), "Multi-fidelity modelling via recursive co-kriging and Gaussian–Markov random fields," *Proc. R. Soc. A*, 471, 20150018.

Pincus, R., Barker, H. W., and Morcrette, J.-J. (2003), "A fast, flexible, approximate technique for computing radiative transfer in inhomogeneous cloud fields," *Journal of Geophysical Research: Atmospheres (1984–2012)*, 108.

Pratola, M., Chipman, H., George, E., and McCulloch, R. (2017), "Heteroscedastic BART Using Multiplicative Regression Trees," *arXiv preprint arXiv:1709.07542*.

Qian, P. Z. and Wu, C. J. (2008), "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments," *Technometrics*, 50, 192–204.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Jeff Wu, C. F. (2005), "Building Surrogate Models Based on Detailed and Approximate Simulations," *Journal of Mechanical Design*, 128, 668–677.

Roberts, G. O., Rosenthal, J. S., et al. (2004), "General state space Markov chains and MCMC algorithms," *Probability surveys*, 1, 20–71.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), "Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction," *Statistical Science*, 4, 409–435.

Sang, H. and Huang, J. Z. (2012), "A full-scale approximation of covariance functions for large spatial data sets," *Journal of the Royal Statistical Society, Series B, In press*, 74, 19–741.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, M., Duda, K. G., Huang, X. Y., Wang, W., and Powers, J. G. (2008), "A description of the Advanced Research WRF Version 3," Tech. rep., National Center for Atmospheric Research.

Williams, C. K. and Rasmussen, C. E. (2006), "Gaussian processes for machine learning," *the MIT Press*, 2, 4.

Yan, H., Qian, Y., Lin, G., Leung, L., Yang, B., and Fu, Q. (2014), "Parametric sensitivity and calibration for Kain–Fritsch convective parameterization scheme in the WRF model," *Clim Res*, 59, 135–147.

Yang, B., Qian, Y., Lin, G., Leung, R., and Zhang, Y. (2012), "Some issues in uncertainty quantification and parameter tuning: a case study of convective parameterization scheme in the WRF regional climate model," *Atmospheric Chemistry and Physics*, 12, 2409.

Zertuche, F. (2015), "Assessment of uncertainty in computer experiments when working with multifidelity simulators." Theses, Université Grenoble Alpes.

A Appendix

Let $\mathfrak{Z}$, $\mathfrak{J}$ denote any sub-sets of the design $\tilde{\mathscr{D}}_t$ for $t = 1, ..., S$. Let $|\mathfrak{Z}|$ denote the size of $\mathfrak{Z}$, and let $y_0(\cdot) = 0$ and $\xi_0(\cdot) = 0$. The parameters of the conditional distributions in (13)-(16) are

$$
\begin{aligned}
\hat{\boldsymbol{B}}_t(\boldsymbol{\phi}|\mathfrak{Z}) =& [\boldsymbol{H}_t^\top(\mathfrak{Z})\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi})\boldsymbol{H}_t(\mathfrak{Z}) + \boldsymbol{B}_t^{-1}]^{-1},\ t = 1:S & (21)\\
\hat{\boldsymbol{\beta}}_t(\boldsymbol{\phi}|\mathfrak{Z}) =& \hat{\boldsymbol{B}}_t(\boldsymbol{\phi}|\mathfrak{Z})[\boldsymbol{H}_t^\top(\mathfrak{Z})\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi})[y_t(\mathfrak{Z}) - \xi_{t-1}(\mathfrak{Z}|\boldsymbol{\gamma}_{t-1}) \circ y_{t-1}(\mathfrak{Z})] + \boldsymbol{B}_t^{-1}\boldsymbol{b}_t] & (22)\\
\hat{\boldsymbol{G}}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z}) =& [\boldsymbol{W}_{t-1}(\mathfrak{Z};\boldsymbol{y}_{t-1})\boldsymbol{C}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z})\boldsymbol{W}_{t-1}^\top(\mathfrak{Z};\boldsymbol{y}_{t-1}) + \boldsymbol{G}_{t-1}^{-1}]^{-1},\ t = 2:S\\
\hat{\gamma}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z}) =& \hat{\boldsymbol{G}}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z})[\boldsymbol{G}_{t-1}^{-1}\boldsymbol{g}_{t-1} + \boldsymbol{W}_{t-1}^\top(\mathfrak{Z};\boldsymbol{y}_{t-1})\hat{\boldsymbol{C}}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z})[y_t(\mathfrak{Z}) - \boldsymbol{H}_t(\mathfrak{Z})\boldsymbol{b}_t]],\ t = 2, ..., S\\
\hat{\boldsymbol{C}}_{t-1}(\boldsymbol{\phi}|\mathfrak{Z}) =& \boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi}) + \boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi})\boldsymbol{H}_t(\mathfrak{Z})\\
&\times [\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi}) + \boldsymbol{H}_t(\mathfrak{Z})\boldsymbol{B}_t^{-1}\boldsymbol{H}_t^\top(\mathfrak{Z})]\boldsymbol{H}_t^T(\mathfrak{Z})\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi}),\ t = 2, ..., S\\
\hat{\lambda}_t(\mathfrak{Z}) =& \lambda_t + \frac{|\mathfrak{Z}|}{2},\ t = 1, ..., S\\
\hat{\chi}_t(\boldsymbol{\phi}|\mathfrak{Z}) =& (|\mathfrak{Z}| + 2\lambda_t - 2)\hat{\sigma}_t^2(\boldsymbol{\phi}|\mathfrak{Z}),\ t = 1, ..., S\\
\hat{\sigma}_t^2(\boldsymbol{\phi}|\mathfrak{Z}) =& \frac{1}{2\lambda_t + |\mathfrak{Z}| - 2}\Big(2\chi_t + \boldsymbol{y}_t^\top(\mathfrak{Z})\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi})y_t(\mathfrak{Z}) + \boldsymbol{b}_t^\top\boldsymbol{B}_t^{-1}\boldsymbol{b}_t\\
&+\boldsymbol{g}_{t-1}^\top\boldsymbol{G}_{t-1}^{-1}\boldsymbol{g}_{t-1} - \hat{\boldsymbol{\alpha}}_t^\top(\boldsymbol{\phi},\mathfrak{Z})\hat{\boldsymbol{A}}_t^{-1}(\boldsymbol{\phi}|\mathfrak{Z})\hat{\boldsymbol{\alpha}}_t(\boldsymbol{\phi}|\mathfrak{Z})\Big),\ t = 1, ..., S & (23)\\
\hat{\boldsymbol{A}}_t(\boldsymbol{\phi}|\mathfrak{Z}) =& \Big[\boldsymbol{L}_t(\mathfrak{Z};\boldsymbol{y}_{t-1})^\top R_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi})L_t(\mathfrak{Z};\boldsymbol{y}_{t-1}) + \operatorname{diag}(\boldsymbol{B}_t^{-1},\boldsymbol{G}_{t-1}^{-1})\Big]^{-1}; & (24)\\
\hat{\boldsymbol{\alpha}}_t(\boldsymbol{\phi}|\mathfrak{Z}) =& \hat{\boldsymbol{A}}_t(\boldsymbol{\phi}|\mathfrak{Z})\left(\boldsymbol{L}_t(\mathfrak{Z};\boldsymbol{y}_{t-1})^\top\boldsymbol{R}_t^{-1}(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi}) + \Big[\boldsymbol{b}_t^\top\boldsymbol{B}_t^{-1},\boldsymbol{g}_{t-1}^\top\boldsymbol{G}_{t-1}^{-1}\Big]^\top\right). & (25)
\end{aligned}
$$

where: $\boldsymbol{W}_{t-1}(\mathfrak{Z};\boldsymbol{y}_{t-1}) = \operatorname{diag}(y_{t-1}(\mathfrak{Z}))\boldsymbol{w}_{t-1}(\mathfrak{Z})$ for $t = 2, ..., S$ and $\boldsymbol{W}_0(\mathfrak{Z};\cdot) = 0$; $\boldsymbol{L}_t(\mathfrak{Z};\boldsymbol{y}_{t-1}) = \big[\boldsymbol{H}_t(\mathfrak{Z}), \operatorname{diag}(\boldsymbol{y}_{t-1}(\mathfrak{Z})\boldsymbol{W}_{t-1}(\mathfrak{Z}))\big]$ for $t = 2, ..., S$ and $\boldsymbol{L}_1(\mathfrak{Z};\cdot) = \boldsymbol{H}_1(\mathfrak{Z})$. In the manuscript, when $\mathfrak{Z} = \mathscr{D}_{k,t}$, we use the notation $\hat{\boldsymbol{B}}_{k,t} := \hat{\boldsymbol{B}}_t(\boldsymbol{\phi}|\tilde{\mathscr{D}}_{k,t})$, $\hat{\boldsymbol{\beta}}_t(\boldsymbol{\phi}) := \hat{\boldsymbol{\beta}}_t(\boldsymbol{\phi}|\tilde{\mathscr{D}}_{k,t})$, etc... to easy the notation.

The equations of the functions $\hat{\boldsymbol{R}}_{k,t}$, $\hat{\boldsymbol{\mu}}_{(t-1)\to t}$, and $\hat{\boldsymbol{\mu}}_{(t+1)\to t}$ in (9)

$$
\begin{aligned}
\hat{\boldsymbol{R}}_t(\boldsymbol{\phi}|\mathfrak{Z};\mathfrak{J}) =& \boldsymbol{R}_t(\mathfrak{Z},\mathfrak{Z}|\boldsymbol{\phi}) - \boldsymbol{R}_t(\mathfrak{Z},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{R}_t^{-1}(\mathfrak{J},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{R}_t^\top(\mathfrak{Z},\mathfrak{J}|\boldsymbol{\phi})\\
&+ \big[\boldsymbol{H}_t(\mathfrak{Z}) + \boldsymbol{R}_t(\mathfrak{Z},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{R}_t^{-1}(\mathfrak{J},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{H}_t(\mathfrak{J})\big]\,\hat{\boldsymbol{B}}_t(\boldsymbol{\phi}|\mathfrak{J})\\
&\times \big[\boldsymbol{H}_t(\mathfrak{Z}) + \boldsymbol{R}_t(\mathfrak{Z},\mathfrak{J}|\phi)\boldsymbol{R}_t^{-1}(\mathfrak{J},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{H}_t(\mathfrak{J})\big]^\top & (26)\\
\hat{\boldsymbol{\mu}}_{(t-1)\to t}(\boldsymbol{\phi},\boldsymbol{\gamma}|\mathfrak{Z};\mathfrak{J}) =& \xi_{t-1}(\mathfrak{Z}|\boldsymbol{\gamma}) \circ y_{t-1}(\mathfrak{Z}) + \boldsymbol{H}_t(\mathfrak{Z})\hat{\boldsymbol{\beta}}_t(\boldsymbol{\phi}|\mathfrak{J})\\
&+ \boldsymbol{R}_t(\mathfrak{Z},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{R}_t^{-1}(\mathfrak{J},\mathfrak{J}|\boldsymbol{\phi})\\
&\times \Big[y_t(\mathfrak{J}) - \xi_{t-1}(\mathfrak{J}|\gamma) \circ y_{t-1}(\mathfrak{J}) - \boldsymbol{H}_t(\mathfrak{J})\hat{\boldsymbol{\beta}}_t(\boldsymbol{\phi}|\mathfrak{J})\Big],\ t = 1:S\\
\hat{\boldsymbol{\mu}}_{(t+1)\to t}(\boldsymbol{\phi},\boldsymbol{\gamma}|\mathfrak{Z};\mathfrak{J}) =& y_{t+1}(\mathfrak{Z}) - \boldsymbol{H}_{t+1}(\mathfrak{Z})\hat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\phi}|\mathfrak{J})\\
&- \boldsymbol{R}_{t+1}(\mathfrak{Z},\mathfrak{J}|\boldsymbol{\phi})\boldsymbol{R}_{t+1}^{-1}(\mathfrak{J},\mathfrak{J}|\boldsymbol{\phi}) & (27)\\
&\times \Big[y_{t+1}(\mathfrak{J}) - \xi_t(\mathfrak{J}|\gamma) \circ y_t(\mathfrak{J}) - \boldsymbol{H}_{t+1}(\mathfrak{J})\hat{\boldsymbol{\beta}}_{t+1}(\boldsymbol{\phi}|\mathfrak{J})\Big],\ t = 1:S-1
\end{aligned}
$$

Supplementary material for the paper: Bayesian analysis of multifidelity computer models with local features and non-nested experimental designs: Application to the WRF model

Bledar A. Konomi [*]
Department of Mathematical Sciences, University of Cincinnati, USA
and
Georgios Karagiannis [*]
Department of Mathematical Sciences, Durham University, UK

October 2, 2020

Contents

[*] The two authors contributed equally to this work. Corresponding authors: Bledar A. Konomi (alex.konomi@uc.edu) and Georgios Karagiannis (georgios.karagiannis@durham.ac.uk).

S.1 Schematics & algorithms of the proposed approach

A schematic of the proposed approach is presented in Algorithm 1.

Algorithm 1 Schematic of the proposed approach.

Returns: A MCMC sample $\mathcal{S}^N = (\mathring{\boldsymbol{y}}^{(j)}, \mathcal{T}^{(j)}, \boldsymbol{\gamma}^{(j)}, \boldsymbol{\sigma^2}^{,(j)}, \boldsymbol{\phi}^{(j)})_{j=1}^N$ from $\pi(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \mathring{\boldsymbol{y}}|\boldsymbol{y})$ in (8). Distribution of a function $g(\cdot)$ of $y(\cdot)$ at untried locations $x^* \in \mathscr{D}^*$.

Inquires: Seeds $\mathcal{T}^{(0)}, \mathring{\boldsymbol{y}}^{(0)}, \boldsymbol{\gamma}^{(0)}, \boldsymbol{\sigma^2}^{,(0)}, \boldsymbol{\phi}^{(0)}$

Set of untried input points $\mathscr{D}^*$

1. Posterior simulation:

   For $j = 1, ..., N$: sample

   $$\mathcal{T}^{(j)}, \boldsymbol{\beta}^{(j)}, \boldsymbol{\gamma}^{(j)}, \boldsymbol{\sigma^2}^{,(j)}, \boldsymbol{\phi}^{(j)}, \mathring{\boldsymbol{y}}^{(j)} \sim P(\cdot|\mathcal{T}^{(j-1)}, \boldsymbol{\beta}^{(j-1)}, \boldsymbol{\gamma}^{(j-1)}, \boldsymbol{\sigma^2}^{,(j-1)}, \boldsymbol{\phi}^{(j-1)}, \mathring{\boldsymbol{y}}^{(j-1)})$$

   where $P(\cdot|\cdot)$ is a Markov transition probability with stationary distribution $\pi(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \mathring{\boldsymbol{y}}|\boldsymbol{y})$ that can be simulated by Algorithm 2.

2. Parametric inference:

   Estimate $\mathring{\boldsymbol{y}} = (\mathring{\boldsymbol{y}}_{k,t}; k = 1, ..., K, t = 1, ..., S)$ either by a Blackwellized Monte Carlo estimator or vanilla Monte Carlo estimator ; Section 2.4.

3. Emulation:

   For $j = 1, ..., N$ (parallel); for $t = 1, ..., S$ (parallel) simulate (18) and (19) via composition, and compute the Monte Carlo predictions.

One random scan of the RJMCMC sampler (Section 2.3) targeting the posterior distribution $\pi(\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \mathring{\boldsymbol{y}}|\boldsymbol{y})$ is presented in Algorithm 2. The Markov transition probability of this process is denoted as $P(\cdot|\mathcal{T}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \mathring{\boldsymbol{y}})$.

Algorithm 2 The RJMCMC transition probability $P(\cdot|\mathcal{T},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma^2},\boldsymbol{\phi},\mathring{\boldsymbol{y}})$ targeting $\pi(\mathcal{T},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma^2},\boldsymbol{\phi},\mathring{\boldsymbol{y}}|\boldsymbol{y})$.

[BL.1] Update $\mathring{\boldsymbol{y}}|\boldsymbol{y},\boldsymbol{\phi},\boldsymbol{\sigma^2},\boldsymbol{\gamma},\mathcal{T}$

```
    For $t = 1, ..., S-1$ (sequential);
        for $k = 1, ..., K$ (parallel);
            draw $\mathring{\boldsymbol{y}}_{k,t}$ by simulating from
```

$$\mathring{\boldsymbol{y}}_{k,t}|\boldsymbol{\phi}_{k,t},\boldsymbol{\gamma}_{k,t-1},\boldsymbol{\gamma}_{k,t},\boldsymbol{\sigma^2_{k,t}},\boldsymbol{\sigma^2_{k,t+1}},\mathcal{T}\sim \mathrm{N}(\mathring{\boldsymbol{\mu}}_{k,t},\mathring{\boldsymbol{\Sigma}}_{k,t})$$

where $\mathring{\boldsymbol{\mu}}_{k,t},\mathring{\boldsymbol{\Sigma}}_{k,t}$ are given on page 11 in the main manuscript.

[BL.2] Update $\mathcal{T},\boldsymbol{\phi}|\tilde{y}$ via reversible jump algorithm:

Simulate $\boldsymbol{\phi},\mathcal{T}$ from a reversible jump transition prob. targeting $\pi(\mathcal{T},\boldsymbol{\phi}|\tilde{\boldsymbol{y}})$. Details can be found on page 12 in the main manuscript.

[BL.3] Update $\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma^2},\boldsymbol{\phi}|\tilde{\boldsymbol{y}},\mathcal{T}$

For $t = 1, ..., S$ and $k = 1, ..., K$ (parallel) simulate from

$$\begin{aligned}
\boldsymbol{\beta}_{k,t}|\tilde{\boldsymbol{y}}_{k,t},\tilde{\boldsymbol{y}}_{k,t-1},\boldsymbol{\gamma}_{k,t-1},\sigma^2_{k,t},\boldsymbol{\phi}_{k,t} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}_{k,t}(\boldsymbol{\phi}_{k,t}),\hat{\boldsymbol{B}}_{k,t}(\boldsymbol{\phi}_{k,t})\sigma^2_{k,t}), \text{ for } t=2,...S \qquad (1)\\
\boldsymbol{\beta}_{k,1}|\tilde{\boldsymbol{y}}_{k,1},\sigma^2_{k,1},\boldsymbol{\phi}_{k,1} &\sim \mathrm{N}(\hat{\boldsymbol{\beta}}_{k,1}(\boldsymbol{\phi}_{k,1}),\hat{\boldsymbol{B}}_{k,1}(\boldsymbol{\phi}_{k,1})\sigma^2_{k,1}), \qquad (2)\\
\boldsymbol{\gamma}_{k,t-1}|\tilde{\boldsymbol{y}}_{k,t},\tilde{\boldsymbol{y}}_{k,t-1},\sigma^2_{k,t},\boldsymbol{\phi}_{k,t} &\sim \mathrm{N}(\hat{\boldsymbol{\gamma}}_{k,t-1}(\boldsymbol{\phi}_{k,t}),\hat{\boldsymbol{G}}_{k,t-1}(\boldsymbol{\phi}_{k,t})\sigma^2_{k,t}), \text{ for } t=2,...S\\
\sigma^2_{k,t}|\tilde{\boldsymbol{y}}_{k,t},\tilde{\boldsymbol{y}}_{k,t-1},\boldsymbol{\phi}_{k,t} &\sim \mathrm{IG}(\hat{\lambda}_{k,t},\hat{\chi}_{k,t}(\boldsymbol{\phi}_{k,t})), \text{ for } t=2,...S \qquad (3)\\
\sigma^2_{k,1}|\tilde{\boldsymbol{y}}_{k,1},\boldsymbol{\phi}_{k,1} &\sim \mathrm{IG}(\hat{\lambda}_{k,1},\hat{\chi}_{k,1}(\boldsymbol{\phi}_{k,1})), \qquad (4)\\
\boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}},\mathcal{T} &\sim \pi(\boldsymbol{\phi}_{k,t}|\tilde{\boldsymbol{y}},\mathcal{T}), \qquad (5)
\end{aligned}$$

where the hatted quantities are given in (21)-(23) of Appendix A.

As discussed in the paper, simulating (1)-(2) can be suppressed because the generated values are not necessarily required to perform inference in Algorithm 1. In that case, Algorithm 2 samples from the marginal $P(\cdot|\mathcal{T},\boldsymbol{\gamma},\boldsymbol{\sigma^2},\boldsymbol{\phi},\mathring{\boldsymbol{y}})$ targeting $\pi(\mathcal{T},\boldsymbol{\gamma},\boldsymbol{\sigma^2},\boldsymbol{\phi},\mathring{\boldsymbol{y}}|\boldsymbol{y})$.

S.2 Performance metrics used in the numerical examples

In the examples, we used the following performance metrics:

1. Absolute relative error (RAE)

$$\mathrm{RAE}(x) = \left|\frac{\hat{y}(x) - y(x)}{y(x)}\right|$$

showing performance of a predictive model.

2. Root mean square prediction error (RMSPE) is defined as

$$\mathrm{RMSPE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{\mathrm{pred}} - y_i^{\mathrm{obs}})^2}$$

where y^{obs} is the observed value in test data-set and y_i^{pred} is the predicted value from the model. It measures the accuracy of the prediction from model. Smaller values of RMSPE indicate more a accurate model.

3. Nash-Sutcliffe model efficiency coefficient (NSME) is defined as:

$$\mathrm{NSME} = 1 - \frac{\sum_{i=1}^{n}(y_i^{\mathrm{pred}} - y_i^{\mathrm{obs}})^2}{\sum_{i=1}^{n}(y_i^{\mathrm{obs}} - \overline{y^{\mathrm{obs}}})^2}$$

where y^{obs} is the observed value in test data-set and y_i^{pred} is the predicted value from the model. NSME gives the relative magnitude of the residual variance from data and the model variance. NSME values closer to 1 indicate that the model has a better predictive performance.

4. 95% CVG is the coverage probability of 95% equal tail prediction interval. 95% CVG values closer to 0.95 indicate better prediction performance for the model.

5. 95% ALCI is average length of 95% equal tail prediction interval. Smaller 95% ALCI values indicate better prediction performance for the model.

S.3 1-D nonlinear toy example

We use the pedagogical example in Perdikaris et al. [2017] to show that our proposed Bayesian treed co-kriging approach can also be used to explore smooth nonlinear multi-fidelity modeling. Let us consider a low-fidelity model to be a sinusoidal wave with four periods $f_1(x) = \sin(8\pi x)$. The high-fidelity model is obtained through a transformation of the low-fidelity expression involving a non-uniform scaling and a quadratic non-linearity as $f_2(x) = (x - \sqrt{2}) f_1(x)^2)$. For the co-kriging statistical model in (2), we cast the scaling discrepancy $\xi_1(x)$ as a polynomial of degree 3, and the additive discrepancy function as a Gaussian process with constant mean. We generate by Latin Hypercube Sampling three training data-sets with different sizes , $(n_1 = 50, n_2 = 15)$, $(n_1 = 50, n_2 = 20)$, and $(n_1 = 50, n_2 = 25)$. Against each of these data sets, we train the associate emulators from our approach Bayesian treed co-kriging (BTCK), and Bayesian co-kriging (BCK). BCK differs from BTCK in that it does not involve the binary treed partitioning mechanism.

In Figure S1 , we plot the produced predictive means from the approaches under comparison trained against data set $(n_1 = 50, n_2 = 20)$, as well as the real function $f_2(x)$. We observe that the predictive mean produced from BTCK is much closer to the real $f_2(x)$ than that of BCK. In this particular example, we can hardly eyeball any discontinuity in the predictive mean of the BTCK; this is possibly due to the Bayesian model averaging effect which mitigates artificially imposed discontinuities due to partitioning. When the sample size is smaller than 20, we observe small discontinuities. In these cases, we can further improve the smoothness of the prediction results with at least three different ways: (a) use better basis functions such as b-spline, (b) more informative priors (c) tree with smoother transitions as proposed in Konomi et al. [2013]. However, these extensions are out of the scope of the paper.

In Table S.1 , we present the MSPE produced from the BCK, and BTCK approaches where their predictive means were used as surrogates. We observe that the MSPEs produced from BTCK are substantially smaller that those produced from BCK for all the training data set sizes. This indicates that the consideration of the binary treed partitioning mechanism in our approach is able to result in substantial improvements in prediction.

n_2	15	20	25
BCK	0.0782	0.0420	0.0121
BTCK	0.0401	0.0093	0.0022

Table S.1 : The MSPE produced from the Bayesian Co-kriging (BCK), and the Bayesian Treed Co-kriging (BTCK) approaches.

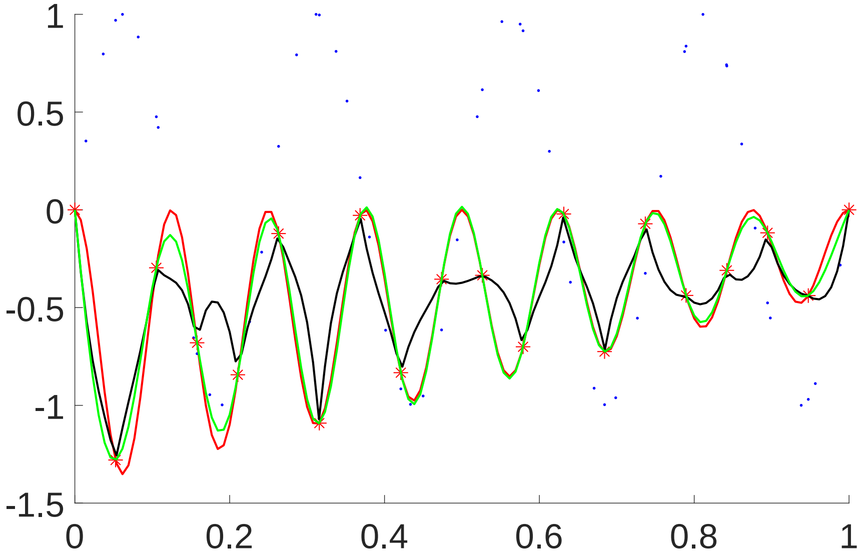


Figure S1 : 1-D Toy example: blue dots represent the low-fidelity simulations, red stars represents the high-fidelity simulation, red line represents the exact high-fidelity function, the black line represents the predictive mean of the multi-fidelity Bayesian co-kriging model, and the green line represents the predictive mean of the multi-fidelity Bayesian treed co-kriging model.

S.4 Graphical representation of input dimensions in Section 4.

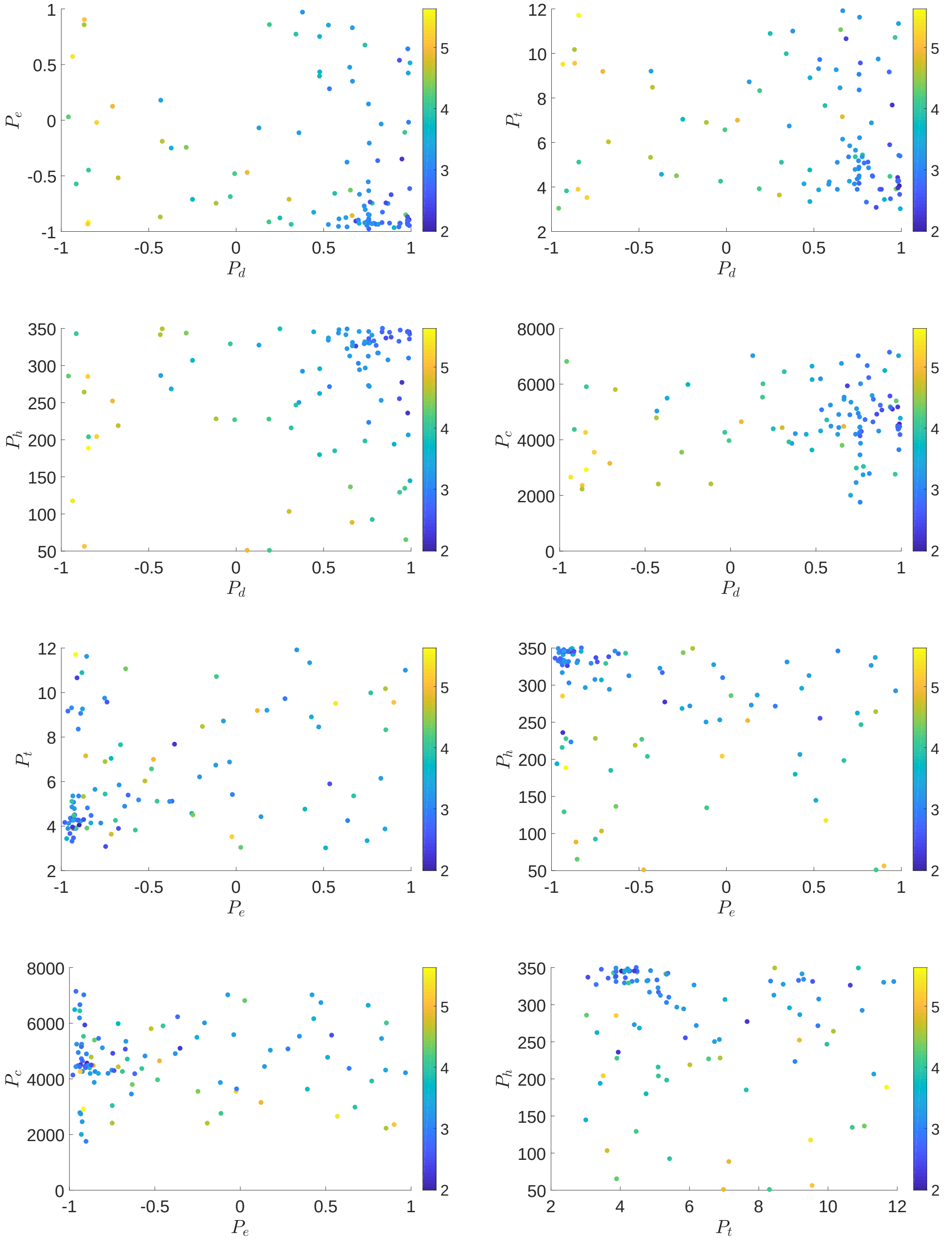


Figure S2 : Design plot for 25 × 25km

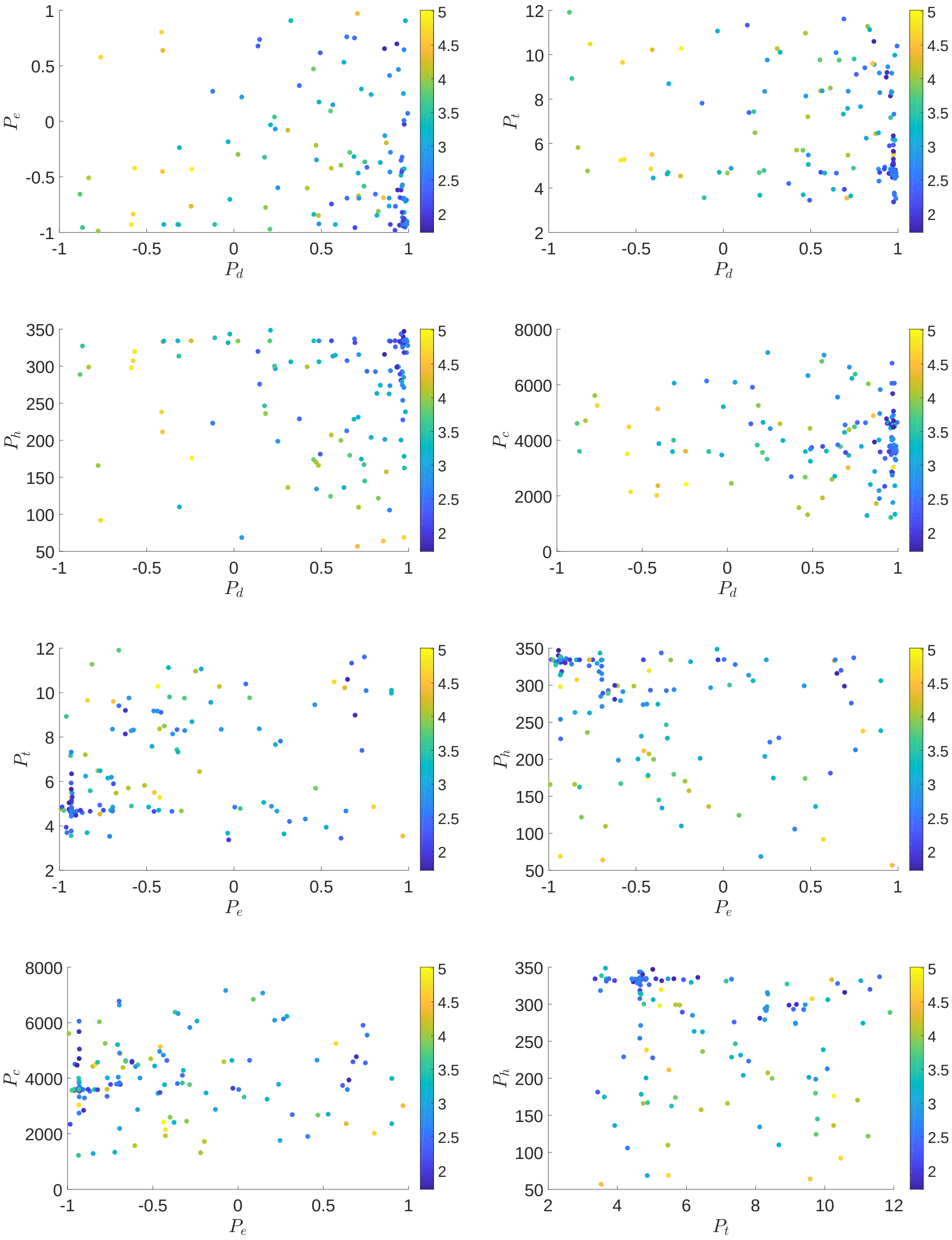


Figure S3 : Design plot for 50 × 50km

S.5 Codes and data set information

The MATLAB codes and the data-sets used to produced the simulations in the case studies in Section , and the application in Section 4 are available from GitHub repository `https://github.com/georgios-stats/ABTCK`.

Description

MATLAB scripts implementing our proposed Bayesian procedure,'Augmented Bayesian treed co-kriging'. This procedure can be used for the statistical analysis of computer models when the available simulations are in multiple fidelity, they are not necessarily based on hierarchically nested experimental designs, and they may present local features, non-stationarities, or discontinuities.

This is a MATLAB code has been used to produce the numerical results and the statistical analysis reported in the paper "Bayesian analysis of multifidelity computer models with local features and non-nested experimental designs: Application to the WRF model", by Alex Konomi and Georgios Karagiannis; submitted in Technometrics for publication.

Requirements

MATLAB compiler (R2017a or later)

Files

- cov_functions: Covariance functions necessary for the GP
- function_fold: Functions used in the simulation study .
- Multi_GP_operations: MCMC operations for the co-kriging Gaussian process hyperparamter updates.
- Multi_prediction: Functions used for prediction.
- ABCK_M: Functions used for imputation and predictions.
- Multi_tree_nested: MCMC treed update of the model when the design is nested. All the necessary RJ-MCMC steps in the nested case including split, merge, change, swap and rotate.
- Multi_tree_NON_nestedB: MCMC treed update of the model when the design is nested. All the necessary RJ-MCMC steps in the non-nested case: split, merge, change, swap and rotate.
- Random_var: Various algorithms for generating data from distributions necessary in the BTC

- help_tree_operations: Operations that help to define the treed form

- FirstSimulationStudy: Contain the example 1 of the paper (section 3.1). It also contain different variations of that example. The user can choose to change the function. This is not a fixed example -- you generate every time new data and check the performance of the algorithm. The performance may depend also on the simulated data.

- SecondSinulationStudy: Contain the example 2 of the paper (section 3.2). Here the user can also change the function and the setting

- Online_ExampleP: Contain the example in the supplementary material S.3 of the paper.

- Application_WRF2: Contain the application to WRF dataset in the paper (section 4). The user can change the settings for testing/training datasets.

References

B. Konomi, H. Sang, and B. Mallick. Adaptive bayesian nonstationary modeling for large spatial datasets using covariance approximations. *Journal of Computational and Graphical Statistics*, 1:In press, 2013.

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 A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.