Cokriging for multivariate Hilbert space valued random fields. Application to multifidelity computer code emulation

Ognjen Grujic[1*], Alessandra Menafoglio[2], Guang Yang[1] and Jef Caers[1]
[1]Stanford Center for Reservoir Forecasting, Stanford University, USA
[2]MOX-Department of Mathematics, Politecnico di Milano, Milano, Italy
[*]ogyg@stanford.edu

Abstract

In this paper we propose Universal trace co-kriging (UTrCoK), a novel methodology for interpolation of multivariate Hilbert space valued functional data. Such data commonly arises in multi-fidelity numerical modeling of the subsurface and it is a part of many modern uncertainty quantification studies. Besides theoretical developments we also present methodological evaluation and comparisons with the recently published projection based approach by Bohorquez et al. [2016]. Our evaluations and analyses were performed on synthetic (oil reservoir) and real field (Uranium contamination) subsurface uncertainty quantification case studies. Monte Carlo analyses were conducted to draw important conclusions and to provide practical guidelines for all future practitioners.

Keywords: Co-kriging of functions, Hilbert Space, computer code emulation

1 Introduction

Numerical reservoir modeling is an irreplaceable component of all modern subsurface uncertainty quantification studies. The reservoir models used in these studies are featured by high dimensional inputs and they often produce multiple outputs that come as any combination of scalars, time series, images or 3D surfaces. Uncertainty quantification, conducted through numerical reservoir models, entails exploration of high dimensional input spaces and production of statistical summaries on the produced outputs. Computational and temporal requirements of uncertainty quantification studies vary depending on the amount of modeled physics. In the reservoir modeling community, there is a general agreement that more modeled physics is better, since it increases the fidelity of the model. However, the higher the amount of modeled physics the higher the computational time, which is often unfavorable in practice due to usually

tight temporal constraints. For these reasons, modelers often build statistical emulators or meta-models as fast replacements for computationally expensive numerical models, or they construct faster numerical models by dropping certain physical aspects of the modeled system.

The idea of statistical emulators is very simple. First statistical design of experiments is used on the input space, then a high fidelity numerical model is employed to compute a set of outputs and finally a regression model is fitted aiming to predict the numerical models output from a given set of inputs. One of the most commonly used statistical emulators for numerical models with scalar outputs is kriging for computer experiments [Sacks et al., 1989, Rasmussen and Williams, 2006, Roustant et al., 2012]. This emulator generalizes the concept of universal kriging [Chiles and Delfiner, 1999] to high dimensional input spaces. Kriging is very convenient in this kind of application because it exactly reproduces the scalar outputs of the training set (i.e., it is an interpolator). Kriging-based emulation of computer experiments that produce functional outputs (time series) is a very active area of research. The recently published meta-modeling methodology by Bottazzi and Della Rossa [2017] utilizes ordinary co-kriging of basis coefficients by Nerini et al. [2010] to construct a functional meta-model. In the same vein, one can also generalize the state-of-the-art non-stationary methods, such as universal trace-kriging (UTrK) by Menafoglio et al. [2013] and universal co-kriging of functional principal component scores (UCoK) by Menafoglio et al. [2016], to construct non-stationary functional meta models. To the best of our knowledge this application has not yet been explored and evaluated in the literature.

Numerical models of lower fidelity are rarely used as a stand-alone replacement for their high-fidelity counterparts. Instead, modelers often use the low fidelity numerical solution in conjunction with the statistical emulators to construct the so-called error models. The idea of error modeling is analogous to the idea of statistical emulation of high fidelity models. One also starts from a training set that is in this case simulated with both high and low fidelity solutions, then proceeds to model the discrepancies ("errors") between the two solutions with some form of regression. Uncertainty quantification then proceeds to explore the input space with the low fidelity solution, and the statistical error model corrects its outputs to best resemble the unevaluated high fidelity counterpart. There are many applications and expansions of this concept up to date. Some of the most notable ones for computer models that produce scalar outputs [Scheidt et al., 2011, Ginsbourger et al., 2013, among others], and for computer models that produce functional outputs [Josset et al., 2015, Trehan et al., 2017, Pagani et al., 2017].

An interesting alternative to error models, that tries to jointly utilize both high and low fidelity models for input space exploration, comes from Kennedy and O'Hagan [2000], who generalized the concept of universal co-kriging [Chiles and Delfiner, 1999] to high dimensional input spaces. The idea underlying the method consists in treating the high-fidelity solution as the "primary" variable,

and the low fidelity solution as the "secondary" variable in a co-kriging setting. This emulator exactly reproduces the high-fidelity training data as the kriging for computer experiments we mentioned previously. The method is advantageous over error models since it can incorporate information from multiple models of different levels of fidelity. In addition, unlike error models, it does not require simulation of all training points with all levels of fidelity. Instead, it allows for completely non-coincident training sets, one for each level of fidelity. Due to its ability to incorporate several numerical solutions of different degrees of fidelity, the method is commonly referred to as "multi-fidelity" meta-modeling. The original work by Kennedy and O'Hagan [2000] considered a subsurface reservoir modeling application with scalar outputs, later Le Gratiet [2012] expanded the method and applied it to multi-fidelity modeling of aeromechanical numerical experiments. Co-kriging for functional data, as well as multi-fidelity modeling for computer experiments that produce functional data, are both very active research areas. Recently, Bohorquez et al. [2016] developed a co-kriging method for multivariate functional data based upon a dimensionality reduction of the data (functional principal component analysis). [Thenon et al., 2016] constructs a multi-fidelity functional meta model with similar, although simplified, ideas as Bohorquez et al. [2016].

In this work, we propose a novel method called "Universal Trace co-Kriging" for interpolation of multivariate functional data. Unlike existing co-kriging methods Bohorquez et al. [2016], Thenon et al. [2016] that are based on prior dimensionality reductions of the data, the method we proposed is fully functional, and developed around the assumption that functional data takes values in an infinite dimensional separable Hilbert space. These developments extend the concepts presented in Menafoglio et al. [2013] to the functional multivariate setting. Although for our application we focus on square-integrable data (i.e., on the embedding into the Hilbert space L^2), our method is entirely general, and allows dealing with different kinds of data. For instance, it allows accounting for differential properties of the data if the considered Hilbert space is a Sobolev space, or to consider distributional data in the form of probability density functions, through the embedding in a Bayes space (see van den Boogaart et al. [2014], Hron et al. [2016], Menafoglio et al. [2014]). Our findings enables one to predict primary observations by jointly using the entire information content embedded in both the primary and the secondary data. The model we propose also allows considering non-stationary response variables, modeled in a universal kriging setting. Although the method is new and of general application is diverse environmental settings, we here consider its development and application in continuity with the concepts originally developed by Kennedy and O'Hagan [2000].

We develop two case studies and three extensive Monte Carlo analyses to compare the performance of our multi-fidelity functional meta models with the multi-fidelity functional meta models constructed with the methodology of Bohorquez et al. [2016]. Here, we also investigate the performances of the functional

meta models that do not account for secondary data, constructed with UTrK by Menafoglio et al. [2013] and UCoK by Menafoglio et al. [2016]. To the best of the authors' knowledge, the conducted case studies are the first contribution that extensively applies and evaluates universal kriging and co-kriging methods for Hilbert data in the context of functional meta-modeling. The important conclusions drawn from the case studies provide valuable guidelines for future practitioners and identify new avenues for future research.

The remaining part of the paper is organized as follows. In Section 2, we present detailed theoretical derivations of universal trace co-kriging. In Section 3 we provide a brief overview of universal co-kriging of functional principal component scores by Bohorquez et al. [2016]. Section 4 investigates the performances of kriging and co-kriging method on simulated data, whereas Section 5 illustrates the application of the method to a case study dealing with a numerical model of uranium bio-remediation experiment in Rifle Colorado. Section 6 summarizes the paper and outlines the opportunities for future work.

2 A Trace-Cokriging predictor for multivariate Hilbert data

We here consider the problem of optimal spatial prediction for multivariate functional random fields, and develop a Universal Trace-Cokriging method which represent the first novel contribution of this work.

In the following developments, we will always consider as ambient space for the analysis a Hilbert space. The use of a Hilbert-space embedding for functional geostatistics is well-documented in the literature (see e.g., [Menafoglio and Secchi, 2017] for a review). Its mathematical and application-oriented convenience is twofold: (i) it allows working by analogy with the scalar setting, providing strong intuitions and interpretations to the concepts involved (e.g., for the concepts of variogram); and (ii) it allows working in a very general setting, which may even involve functional constrained data (e.g., PDFs, [Menafoglio et al., 2014]).

We thus denote by H_k, $k = 1, ..., K$, a separable Hilbert space endowed with the inner product $\langle \cdot, \cdot \rangle_{H_k}$, and call D a Euclidean spatial domain in $\mathbb{R}^d$, $d \geq 1$. Given a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$, we indicate by $\mathcal{X}^{(k)}$ (possibly with a spatial index $\boldsymbol{s} \in D$) a random element on $(\Omega, \mathfrak{F}, \mathbb{P})$ in H_k.

We here consider multivariate random fields: we denote by $\{\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}, \boldsymbol{s} \in D\}$ a multivariate random process on $(\Omega, \mathfrak{F}, \mathbb{P})$, that is valued in the Cartesian space $H^K = H_1 \times H_2 \times \cdots \times H_K$: each element $\boldsymbol{\mathcal{X}}_{\boldsymbol{s}}$ is a vector of K random elements $\mathcal{X}^{(1)}_{\boldsymbol{s}_1}, ..., \mathcal{X}^{(K)}_{\boldsymbol{s}_n}$ in $H_1, ..., H_K$, respectively:

$$\boldsymbol{\mathcal{X}}_{\boldsymbol{s}} = \left(\mathcal{X}^{(1)}_{\boldsymbol{s}}, ..., \mathcal{X}^{(K)}_{\boldsymbol{s}}\right)^T.$$

To define the first and second order properties of the field, we proceed by analogy with the classical framework and define multivariate spatial mean and covari-

ogram structure. We call $\boldsymbol{m_s} \in H^K$ the spatial mean of the process at $\boldsymbol{s}$ in D, that is

$$\boldsymbol{m_s} = \mathbb{E}[\boldsymbol{\mathcal{X}_s}] = \left(m_{\boldsymbol{s}}^{(1)}, ..., m_{\boldsymbol{s}}^{(K)}\right)^T, \quad m_{\boldsymbol{s}}^{(k)} = \mathbb{E}[\mathcal{X}_{\boldsymbol{s}}^{(k)}].$$

To define a measure of multivariate spatial dependence, we generalize to the multivariate setting the concept of trace-covariogram previously introduced by Giraldo [2009], Menafoglio et al. [2013]. Hereafter, we assume that the spaces $H_1, ..., H_K$ coincide, and thus omit the subscript H_k in the notation of the inner product. In case they actually do not coincide, one can apply the trace-kriging strategy of Menafoglio et al. [2013] in the Cartesian space H^K, endowed with the inner product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{H_K} = \sum_{k=1}^{K} \langle x^{(k)}, y^{(k)} \rangle_{H^k}$. This case is not considered further here.

We thus consider the map $C : D \times D \to \mathbb{R}^{K \times K}$, that determines the trace-covariograms and cross-trace-covariograms of the field as follows

$$\begin{aligned} (\boldsymbol{s}, \boldsymbol{u}) &\mapsto C(\boldsymbol{s}, \boldsymbol{u}) \in \mathbb{R}^{K \times K} \\ C_{kl}(\boldsymbol{s}, \boldsymbol{u}) &= \mathbb{E}[\langle \mathcal{X}_{\boldsymbol{s}}^{(k)} - m_{\boldsymbol{s}}^{(k)}, \mathcal{X}_{\boldsymbol{u}}^{(l)} - m_{\boldsymbol{u}}^{(l)} \rangle]. \end{aligned}$$

Note that this quantity cannot be defined in case of non-coincident $H_1, ..., H_K$, as the inner product between elements of different spaces is not defined.

In this work, we assume that every element $\mathcal{X}_{\boldsymbol{s}}^{(k)}$ of the multivariate process $\boldsymbol{\mathcal{X}_s}$ is non stationary, and that it can be represented by a sum of deterministic mean (drift) and zero-mean globally second order stationary residual:

$$\mathcal{X}_{\boldsymbol{s}}^{(k)} = m_{\boldsymbol{s}}^{(k)} + \delta_{\boldsymbol{s}}^{(k)}. \tag{1}$$

Here, the drift is assumed to be non-constant in space D and, analogously to Menafoglio et al. [2013], modeled as a functional linear model

$$m_{\boldsymbol{s}}^{(k)} = \sum_{l=0}^{L} a_l^{(k)} f_l(\boldsymbol{s}) \tag{2}$$

where $a_l^{(k)}$ are coefficients in H_k, and $f_l(.)$ are scalar regressors known over the entire domain D. Further, the residual is assumed to be globally second order stationary in the sense of Menafoglio et al. [2013]. That is, we assume that the multivariate trace-covariogram structure depends only on the increment between locations, i.e., there exists $\widetilde{C}$ such that $\widetilde{C}(\boldsymbol{s} - \boldsymbol{u}) = C(\boldsymbol{s}, \boldsymbol{u})$, for all $\boldsymbol{s}, \boldsymbol{u} \in D$. For ease of notation, hereafter we denote $\widetilde{C}$ simply by C.

We call $\boldsymbol{s}_1, ..., \boldsymbol{s}_{N_j}$ $(j = 1, ..., K)$ the measurement locations (or *design of experiment*), and $\mathcal{X}_{\boldsymbol{s}_1}^{(j)}, ..., \mathcal{X}_{\boldsymbol{s}_{N_j}}^{(j)}$ the partial observation of the j-th element of the multivariate process at these locations. Within the former assumptions, we aim to predict the k-th element $\mathcal{X}_{\boldsymbol{s}_0}^{(k)}$ of $\boldsymbol{\mathcal{X}_{s_0}}$ at a target location $\boldsymbol{s}_0$ in D. To this

end, we consider the Trace-Cokriging predictor, that is the best linear unbiased predictor within the class of linear predictors

$$\mathcal{X}_{\boldsymbol{s}_0}^{(k)\lambda} = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \lambda_{ji} \mathcal{X}_{\boldsymbol{s}_i}^{(j)} \tag{3}$$

To find the optimal weights, λ_{ji}^*, $j = 1, ..., K$, $i = 1, ..., N_j$, we minimize the mean squared error of prediction under the unbiasedness constraint, that is

$$\begin{aligned} \min_{\substack{\lambda_{ji} \in \mathbb{R}, \\ j=1,...,K, i=1,...,N_j}} \quad & \mathbb{E}\left[\| \mathcal{X}_{\boldsymbol{s}_0}^{(k)\lambda} - \mathcal{X}_{\boldsymbol{s}_0}^{(k)} \|^2 \right] \\ \text{subject to} \quad & \mathbb{E}[\mathcal{X}_{\boldsymbol{s}_0}^{(k)\lambda}] = m_{\boldsymbol{s}_0}^{(k)}. \end{aligned} \tag{4}$$

It is straightforward to see that the unbiasedness constraint reads as

$$\begin{aligned} &\sum_{i=1}^{N_k} \lambda_{ki} f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\ &\sum_{i=1}^{N_j} \lambda_{ji} f_l(\boldsymbol{s}_i) = 0, \quad \text{for} \quad j \neq k, \quad \forall l; \end{aligned} \tag{5}$$

Indeed,

$$\mathbb{E}[\mathcal{X}_{\boldsymbol{s}_0}^{(k)\lambda}] = \sum_{j=1}^{K} \sum_{i=1}^{N_j} \lambda_{ji} m_{\boldsymbol{s}_i}^{(j)}$$

and the latter quantity is equal to $m_{\boldsymbol{s}_0}^{(j)}$ if and only if condition (5) is fulfilled.

Developing the functional in the first line of Eq. 4 yields:

$$\begin{aligned} \mathbb{E}\left[\| \mathcal{X}_{\boldsymbol{s}}^{(k)\lambda} - \mathcal{X}_{\boldsymbol{s}}^{(k)} \|^2 \right] = C_{kk}(\mathbf{0}) + \\ \sum_{j=1}^{K} \sum_{i=1}^{N_j} \sum_{j'=1}^{K} \sum_{i'=1}^{N_{j'}} \lambda_{ji} \lambda_{j'i'} C_{jj'}(\boldsymbol{s}_i - \boldsymbol{s}_{i'}) - \\ 2 \sum_{j=1}^{K} \sum_{i=1}^{N_j} \lambda_{ji} C_{jk}(\boldsymbol{s}_i - \boldsymbol{s}_0) \end{aligned} \tag{6}$$

Introducing Kx$(L+1)$ Lagrange multipliers to account for the unbiasedness

constraints in Eq. (5) leads to the following objective functional

$$\begin{aligned}\Phi(\lambda) = C_{kk}(\mathbf{0}) + \sum_{j=1}^{K}\sum_{i=1}^{N_j}\sum_{j'=1}^{K}\sum_{i'=1}^{N_{j'}}\lambda_{ji}\lambda_{j'i'}C_{jj'}(\boldsymbol{s}_i-\boldsymbol{s}_{i'}) - \\ 2\sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jk}(\boldsymbol{s}_i-\boldsymbol{s}_0) + \\ 2\sum_{l=0}^{L}\mu_{kl}\left(\sum_{i=1}^{N_k}\lambda_{ki}f_l(\boldsymbol{s}_i) - f_l(\boldsymbol{s}_0)\right) + \\ 2\sum_{l=0}^{L}\sum_{\substack{j=1\\ j\neq k}}^{K}\mu_{jl}\left(\sum_{i=1}^{N_j}\lambda_{ji}f_l(\boldsymbol{s}_i)\right)\end{aligned} \tag{7}$$

After taking partial derivatives of equation (7) with respect to λ's and μ's we obtain the following system of linear equations:

$$\begin{aligned}\sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jj'}(\boldsymbol{s}_i-\boldsymbol{s}_{i'}) + \sum_{l=0}^{L}\mu_{j'l}f_l(\boldsymbol{s}_i) = C_{j'k}(\boldsymbol{s}_{i'}-\boldsymbol{s}_0), \\ (j'=1,...,K; i'=1,...,N_{j'};); \\ \sum_{i=1}^{N_k}\lambda_{ki}f_l(\boldsymbol{s}_i) = f_l(\boldsymbol{s}_0), \quad \forall l; \\ \sum_{i=1}^{N_j}\lambda_{ji}f_l(\boldsymbol{s}_i) = 0, \quad j\neq k, \quad \forall l;\end{aligned} \tag{8}$$

The trace-variance associated with predictor $\mathcal{X}_{\boldsymbol{s}_0}^{(k)*} = \sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}^{*}\mathcal{X}_{\boldsymbol{s}_i}^{(j)}$ is given by

$$\sigma_k^2(\boldsymbol{s}_0) = C_{kk}(\mathbf{0}) - \sum_{j=1}^{K}\sum_{i=1}^{N_j}\lambda_{ji}C_{jk}(\boldsymbol{s}_i-\boldsymbol{s}_0) + \sum_{l=0}^{L}\mu_{kl}f_l(\boldsymbol{s}_0).$$

System (8) can be expressed in a matrix form as follows (for $k=1$):

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} & \cdots & \boldsymbol{C}_{1K} & \boldsymbol{F}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} & \cdots & \boldsymbol{C}_{2K} & \mathbf{0} & \boldsymbol{F}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{C}_{K1} & \boldsymbol{C}_{K2} & \cdots & \boldsymbol{C}_{KK} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F}_K \\ \boldsymbol{F_1}^T & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{F}_2^T & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{F}_K^T & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \vdots \\ \boldsymbol{\lambda}_K \\ \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \boldsymbol{\mu}_3 \\ \vdots \\ \boldsymbol{\mu}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_{10} \\ \boldsymbol{c}_{20} \\ \vdots \\ \boldsymbol{c}_{K0} \\ \boldsymbol{f}_{01} \\ \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix} \tag{9}$$

where:

$$[\boldsymbol{C}_{mn}]_{ij} = Cov\left(\mathcal{X}_{\boldsymbol{s}_i}^{(m)}, \mathcal{X}_{\boldsymbol{s}_j}^{(n)}\right) = C_{mn}(\boldsymbol{s}_i - \boldsymbol{s}_j)$$

$$\boldsymbol{c}_{j0} = \begin{bmatrix} C_{jk}(\boldsymbol{s}_1 - \boldsymbol{s}_0) \\ C_{jk}(\boldsymbol{s}_2 - \boldsymbol{s}_0) \\ \vdots \\ C_{jk}(\boldsymbol{s}_{N_j} - \boldsymbol{s}_0) \end{bmatrix}, \boldsymbol{\lambda}_j = \begin{bmatrix} \lambda_{j1} \\ \lambda_{j2} \\ \vdots \\ \lambda_{jN_j} \end{bmatrix}, \boldsymbol{\mu}_j = \begin{bmatrix} \mu_{j0} \\ \mu_{j1} \\ \vdots \\ \mu_{jd} \end{bmatrix},$$

$$\boldsymbol{F}_j = \begin{bmatrix} f_0(\boldsymbol{s}_1) & f_1(\boldsymbol{s}_1) & \cdots & f_L(\boldsymbol{s}_1) \\ f_0(\boldsymbol{s}_2) & f_1(\boldsymbol{s}_2) & \cdots & f_L(\boldsymbol{s}_2) \\ \vdots & \vdots & \vdots & \vdots \\ f_0(\boldsymbol{s}_{N_j}) & f_1(\boldsymbol{s}_{N_j}) & \cdots & f_L(\boldsymbol{s}_{N_j}) \end{bmatrix}, \boldsymbol{f}_{0j} = \begin{bmatrix} f_0(\boldsymbol{s}_0) \\ f_1(\boldsymbol{s}_0) \\ \vdots \\ f_L(\boldsymbol{s}_0) \end{bmatrix}.$$

The system given in equation (9) is analogous to the system of universal co-kriging equations outlined in Chiles and Delfiner [1999].

Parameter inference. The strategy for parameter inference can be analogous to that performed in conventional co-kriging. First functional regression (Ramsay and Silverman [2005]) is used to estimate the functional drift of each of the elements of the multivariate functional data, e.g., via ordinary least squares. Then, the estimates of the trace-auto and trace-cross covariances are computed on the estimated functional residuals and admissible covariance structures are fitted with the linear model of coregionalization (LMC, Goovaerts [1997]). Improved estimates of the drift and of the residuals can be obtained by using a generalized least square approach. However, the latter is associated with the need to use iterative algorithms to jointly estimate the drift and the spatial dependence (see Menafoglio et al. [2013]). In all these cases, note that regression-based drift estimates yields biased estimates of the variogram, due to the use of estimated residuals in place of the true ones, similarly as in the scalar geostatistical case. Menafoglio et al. [2013] discuss the point and show via simulation that such bias does not have a strong influence on the results.

For the purpose of our work we here focus on the procedure to estimate the dependence structure in the stationary case, the non-stationary setting being obtained by replacing to the observations the estimated residuals of model (1). In this case, the auto-covariance estimation can be performed simply by means of trace-variography, introduced by Giraldo [2009], Menafoglio et al. [2013]. The trace variogram estimator is formulated as follows

$$\gamma_{k,k}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(\boldsymbol{h})} \|\mathcal{X}_{\boldsymbol{s}_i}^{(k)} - \mathcal{X}_{\boldsymbol{s}_j}^{(k)}\|^2 \tag{10}$$

where $N(\boldsymbol{h})$ denotes the set of pairs (i, j) approximately separated by a vector $\boldsymbol{h}$, i.e., such that $\boldsymbol{s}_i - \boldsymbol{s}_j \sim \boldsymbol{h}$. To find the cross-covariance estimators we proceed analogously to the multivariate case by generalizing the very well known cross-variogram (Goovaerts [1997]) and pseudo cross-variogram (Clark et al. [1987]) estimators:

1. The trace cross-variogram estimator:

$$\gamma_{k,l}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(\boldsymbol{h})} \langle \mathcal{X}_{\boldsymbol{s}_i}^{(k)} - \mathcal{X}_{\boldsymbol{s}_j}^{(k)}, \mathcal{X}_{\boldsymbol{s}_i}^{(l)} - \mathcal{X}_{\boldsymbol{s}_j}^{(l)} \rangle \tag{11}$$

2. The pseudo trace-cross-variogram estimator:

$$\gamma_{k,l}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(i,j)\in N(h)} \|\mathcal{X}_{\boldsymbol{s}_i}^{(k)} - \mathcal{X}_{\boldsymbol{s}_j}^{(l)}\|^2 \tag{12}$$

The properties of the trace cross-variograms are the same as their scalar counterparts. Pseudo trace cross-variogram is always positive and applicable to both isotopic and heterotopic data sampling, while the trace cross-variogram is only applicable in the case of isotopic data sampling (Wackernagel [2010]). In practice, inference and fitting of trace variograms over high dimensional input spaces is limited to omni-directional variograms due to difficulties with unidirectional (marginal) variogram estimation in high dimension (i.e., curse of dimensionality De Cesare et al. [2001]).

It should be noted that, when considering formulas (11) and (12), all elements of multivariate functional data need to be in the same units and scale. In scalar geostatistics a simple rescaling or normalization is often employed to overcome this problem (Goovaerts [1997]). However, for functional data there is no general consensus on what should be consider as the most appropriate rescaling method. For instance, a possible generalization of scalar standardization is a point-wise standardization (i.e., point-wise subtraction of a sample mean, followed by a point-wise division by sample standard deviation, both mean and standard deviation estimated point-wise). However, this kind of standardization may have detrimental effects on the functional form of the data, besides being not well-defined from the mathematical viewpoint. Instead, a sensible notion of standardization of the element $\mathcal{X}_{\boldsymbol{s}_l}^{(j)}$, $j = 1, ..., K$, $l = 1, ..., N_j$, is

$$\frac{\mathcal{X}_{\boldsymbol{s}_l}^{(j)} - \overline{\mathcal{X}}^{(j)}}{\frac{1}{N_j}\sum_{i=1}^{N_j} \left\| \mathcal{X}_{\boldsymbol{s}_i}^{(j)} - \overline{\mathcal{X}}^{(j)} \right\|}.$$

Here, $\overline{\mathcal{X}}^{(j)} = \frac{1}{N_j}\sum_{i=1}^{N_j} \mathcal{X}_{\boldsymbol{s}_i}^{(j)}$ denotes the sample mean of the data, which is subtracted to the observation prior to the normalization with respect to the (trace-)standard deviation of the data.

Note that, currently, the method of moments and least squares fitting approaches appear as the most viable procedures, as the concept of density for functional data is not mathematically well-defined (Delaigle and Hall [2010]), preventing the use of automated maximum likelihood based parameter inference procedure.

The range of applicability. As mentioned previously, the Hilbert spaces within which the element of vector $\boldsymbol{\mathcal{X}}_s$ are embedded must be coincident in

order to compute the cross covariances. In multi-fidelity modeling this is almost always the case since low fidelity simulations produce the same type of output data as their high fidelity counterparts. Another requirement for this modeling strategy to work is that discrepancies between functional data be in amplitude rather than in phase. Phase shifted interpolation is more complex and it would require modeling with warping functions (Ramsay and Li [1998]) that is beyond the scope of this work. Nonetheless, both the former and the latter requirements could be overcome through extensions of the proposed setting.

3 Projection based co-kriging for multivariate Hilbert data

An alternative approach to interpolation of multivariate functional data was recently proposed by Bohorquez et al. [2016] as an extension of the method introduced by Nerini et al. [2010] to non-stationary multivariate functional data. The method relies on a functional principal component decomposition (FPCA, Ramsay and Silverman [2005]) of each element $\mathcal{X}^{(j)}$ of the multivariate functional vectors $\boldsymbol{\mathcal{X}}$, followed by modeling and prediction of functional principal component scores, in a co-kriging setting. For sake of clarity, we recall the method by extending the notation of Bohorquez et al. [2016], valid for square-integrable data in L^2 only, to the more general setting of Hilbert-space data here considered.

For simplicity, consider a sample of bi-variate (2-levels) functional data $\boldsymbol{\mathcal{X}}_{\boldsymbol{s}_i} = \left(\mathcal{X}^{(1)}_{\boldsymbol{s}_i}, \mathcal{X}^{(2)}_{\boldsymbol{s}_i}\right)^T$ fully observed over a set of design points $\boldsymbol{s}_i$ where $i = 1, 2, ..., N$ and where $\boldsymbol{s}$ is a vector in $\mathbb{R}^d$. Let $e^{(1)} = \{\phi^{(1)}_1, \phi^{(1)}_2, ..., \phi^{(1)}_P\}$ and $e^{(2)} = \{\phi^{(2)}_1, \phi^{(2)}_2, ..., \phi^{(2)}_Q\}$ be ortho-normal sets of functional principal components of $\mathcal{X}^{(1)}_{\boldsymbol{s}_i}$'s and $\mathcal{X}^{(2)}_{\boldsymbol{s}_i}$'s, respectively, and let $\xi^{(k)}_p(\boldsymbol{s}_i)$ be a principal component score of $\mathcal{X}^{(k)}_{\boldsymbol{s}_i}$ on $\phi^{(k)}_p$. By definition, every function in the ensemble can be reconstructed from its mean, its principal component scores and its functional principal components. For a new design point $\boldsymbol{s}_0$, the predictions of $\mathcal{X}^{(k)}_{\boldsymbol{s}_0}$ are thus sought in the following form

$$\mathcal{X}^{(k)}_{\boldsymbol{s}_0} = \mu^{(k)} + \sum_{p=1}^{P} \xi^{(k)}_p(\boldsymbol{s}_0)\phi^{(k)}_p, \tag{13}$$

where $\mu^{(k)}$ is the mean function of k-th level.

The only unknowns in equation (13) are the principal component scores $\xi^{(k)}_p(\boldsymbol{s}_0)$'s that are assumed to be non-stationary

$$\begin{aligned} \xi^{(k)}_p(\boldsymbol{s}) &= m^{(k)}(\boldsymbol{s}) + r^{(k)}(\boldsymbol{s}); \\ m^{(k)}(\boldsymbol{s}) &= \sum_{l=0}^{d} \beta_l f_l(\boldsymbol{s}), \quad \beta_l \in \mathbb{R}. \end{aligned} \tag{14}$$

and sought as the best linear unbiased combination of the principal component scores of *all* the observed curves:

$$\xi_p^{(k)}(\boldsymbol{s}_0) = \sum_{i=1}^{N}\sum_{p=1}^{P} \lambda_{i,p}^{(1)} \xi_p^{(1)}(\boldsymbol{s}_i) + \sum_{i=1}^{N}\sum_{q=1}^{Q} \lambda_{i,q}^{(2)} \xi_q^{(2)}(\boldsymbol{s}_i). \tag{15}$$

The weights $\lambda_{p,q}^{(k)}$ are found by solving the well known system of universal cokriging equations (Chiles and Delfiner [1999], pg. 300):

$$\begin{bmatrix} \boldsymbol{C}_{11}^{11} & \boldsymbol{C}_{12}^{11} & \boldsymbol{C}_{11}^{12} & \boldsymbol{C}_{12}^{12} & \boldsymbol{F}_1^1 & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}_{21}^{11} & \boldsymbol{C}_{22}^{11} & \boldsymbol{C}_{21}^{12} & \boldsymbol{C}_{22}^{12} & \mathbf{0} & \boldsymbol{F}_2^1 & \mathbf{0} & \mathbf{0} \\ \boldsymbol{C}_{11}^{21} & \boldsymbol{C}_{12}^{21} & \boldsymbol{C}_{11}^{22} & \boldsymbol{C}_{12}^{22} & \mathbf{0} & \mathbf{0} & \boldsymbol{F}_2^1 & \mathbf{0} \\ \boldsymbol{C}_{21}^{21} & \boldsymbol{C}_{22}^{21} & \boldsymbol{C}_{21}^{22} & \boldsymbol{C}_{22}^{22} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{F}_2^2 \\ (\boldsymbol{F}_1^1)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\boldsymbol{F}_2^1)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (\boldsymbol{F}_1^2)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & (\boldsymbol{F}_2^2)^T & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1^1 \\ \boldsymbol{\lambda}_2^1 \\ \boldsymbol{\lambda}_1^2 \\ \boldsymbol{\lambda}_2^2 \\ \boldsymbol{\sigma}_1 \\ \boldsymbol{\sigma}_2 \\ \boldsymbol{\sigma}_3 \\ \boldsymbol{\sigma}_4 \end{bmatrix} = \begin{bmatrix} \boldsymbol{c}_0^{11} \\ \boldsymbol{c}_0^{21} \\ \boldsymbol{c}_0^{12} \\ \boldsymbol{c}_0^{22} \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Where $[\boldsymbol{C}_{pq}^{lm}]_{ij} = Cov(\xi_p^l(\boldsymbol{s}_i), \xi_q^m(\boldsymbol{s_j})) = C_{pq}^{lm}(|\boldsymbol{s}_i - \boldsymbol{s}_k|)$, and where $\boldsymbol{F}_p^l(i,.) = \{f_0(\boldsymbol{s}_i), f_1(\boldsymbol{s}_i), ..., f_L(\boldsymbol{s}_i,)\}$.

Parameter inference. Given that this approach effectively transforms a multivariate functional interpolation problem into a multivariate (vector) interpolation problem, many parameter inference procedures developed in multivariate geostatistics are available. Both variogram fitting procedures with the linear model of co-regionalization (LMC, Goovaerts [1997]), as well as automated maximum likelihood approaches are applicable (Gelfand et al. [2004], Fricker et al. [2013], Zhang [2007]). The size of the model depends on the size of the training dataset and the number of kept functional principal components on every level of multivariate functional data. Bohorquez et al. [2016] reported numerical difficulties with the linear model of co-regionalization for large numbers of kept principal components.

The range of applicability. The projection based approach is applicable to a variety of modeling situations. The method may not be limited to amplitude shifted curves; indeed, it may also account for phase variability if at least one principal component captures a shift in phase. This may have an impact on the complexity required to represent the functional data, hence may entail the increase of the dimensionality of the model that ultimately affects parameter inference. One attractive feature of this approach is that it does not require that Hilbert spaces of functional data are coincident; for instance, the secondary data does not even need to be functional. This opens perspectives to more advanced multi-fidelity emulation approaches by combining information from low fidelity responses that are not necessarily functional. One example of such low fidelity models are flow diagnostics proxies (Shahvali et al. [2012]) that produce one dimensional summaries of flow characteristics of Earth models. With projection

based approaches, flow diagnostics responses can be combined with other functional low fidelity models (i.e. upscaled models) to construct an emulator that predicts high fidelity flow responses.

4 Performance analysis on synthetic data-sets

In this section we set out to explore and assess the performance of the previously presented emulation techniques on a purely synthetic numerical model of the subsurface. For this purpose we developed a homogeneous 3D oil-water reservoir model with 4 producer wells at the top of the reservoir structure, and an aquifer connected at the bottom left corner for pressure support (Figure 1 left). The four wells produce two types of fluid, oil and water. Initially, the reservoir is saturated with oil and wells do not produce any water until the reservoir pressure becomes low enough to allow water encroachment from the aquifer. The speed of encroachment is dependent on the reservoir properties and the viscosity of the present fluids. One typical field water production rate (FWPR) response is given in Figure 1 right, while the model parameters that are the most influential on FWPR are summarized in Table 1.

Given that all of the presented computer code emulation techniques aim to make use of both computationally expensive (high fidelity), and computationally cheap (low fidelity) simulations two levels of numerical abstractions were considered. High fidelity flow simulations were computed on a finely gridded reservoir volume (150x100x25), while the low fidelity flow simulations were computed on a coarsely gridded reservoir volume (150x100x13). In both cases we used Eclipse E100 black oil reservoir simulator to simulate subsurface flow. The two solutions produced somewhat different, but highly correlated ($\rho = 0.91$) flow responses (Figure 1 - right). The discrepancies between the responses are a consequence of numerical dispersion caused by coarser vertical discretization of the reservoir volume.

We used the reservoir model to develop two datasets for methodological comparisons and assessment. The first dataset considered only two input parameters, PERMZm and PORVm (Table 1). The second dataset considered three input parameters: PERMZm, PORVm, and PERM (Table 1). Both datasets consist of training and testing subsets. The training subsets were produced by latin hypercube sampling and were evaluated with both high and low fidelity flow simulations. The test sets were produced with uniform sampling and were evaluated only with the high fidelity flow solution. The two datasets are summarized in Table 2.

Output data pre-processing The training ensemble of FWPR curves of the two parameter dataset is given in Figure 3 - right. Graphical inspection of the figure clearly suggests that the data are shifted in both phase and amplitude. Embedding the data in the space L^2 of square-integrable functions (i.e., setting $H = L^2$ in the notation of Section 2) and applying trace based co-kriging is not

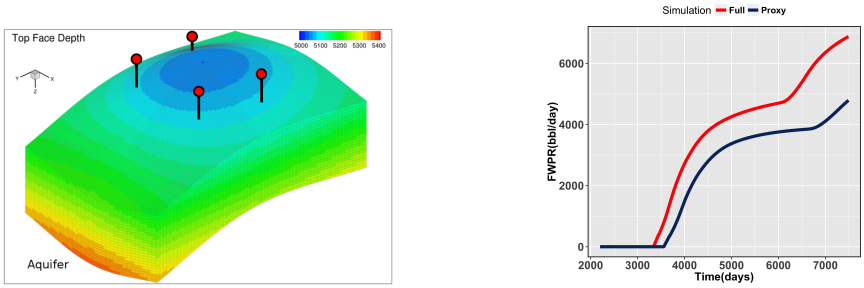

Figure 1: **Left** - 3D reservoir model with four producer wells. Cells are colored by vertical depth in feet. Model dimensions are: X=500 ft, Y=750 ft, Z=150 ft. **Right** - An example of discrepancies between proxy and full physics flow simulations.

Table 1: Simulation parameters

Parameter	**Value**	**Description**
PORVm (-)	1-1000	Aquifer Strength
PERMZm (-)	0-1	Vertical Perm. (K mult.)
K (md)	25	Reservoir permeability
ϕ (frac)	0.2	Reservoir porosity
μ_o (cp)	0.0002	Oil Viscosity
μ_w (cp)	0.00001	Water viscosity

appropriate in this case, since the L^2 geometry is only suitable for unconstrained data with amplitude shift. However, the ensemble of phase-amplitude shifted FWPR curves can be transformed into an ensemble of amplitude shifted curves with a simple ad-hoc procedure. For one curve, the procedure consists of identification of the water breakthrough time, followed by a simple regression fit to the early post water breakthrough rates and substitution of the zero production rates with regressions solution. This procedure is explained visually in Figure 2.

Table 2: Summary of the produced datasets

Dataset type	**# Proxy**	**# Full**	**# Test**
2 parameter	189	176	400
3 parameter	466	462	400

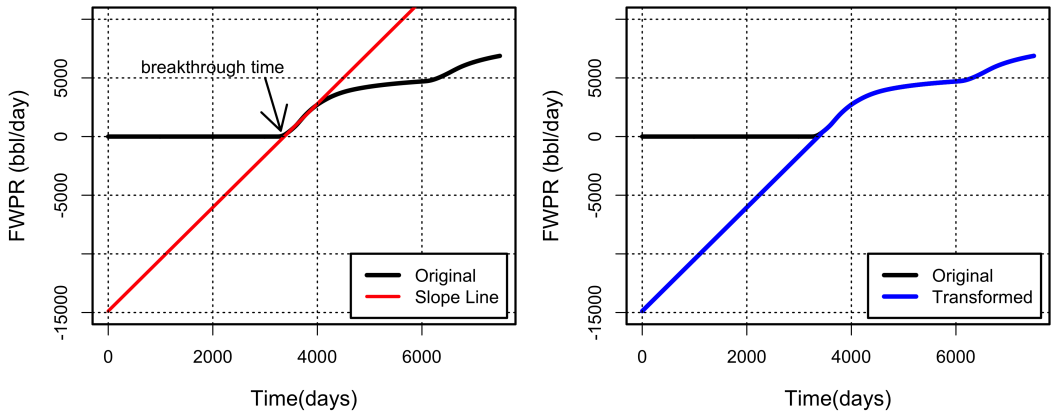

Figure 2: Curve transformation procedure. Left - Original curve with a straight line fitted through the early breakthrough rates. Right - The resulting "transformed" curve.

4.1 Analysis: Computer experiment with 2 parameters

Our first analysis focuses on the two parameter dataset. In this exercise, only a portion of the available training data was used, namely 50 high fidelity flow (fine) simulations and 150 low fidelity simulations (proxy). The sub-sampled training dataset was used to fit the following models: (i) Universal Trace Kriging (**UTrK**) by Menafoglio et al. [2013], (ii) Universal Trace Co-Kriging (**UTrCoK**) introduced in this work, (iii) projection based Universal co-kriging for functional data (**UCoK**) by Menafoglio et al. [2016], and (iv) projection based Universal co-kriging with secondary functional data (**UCoK2**) by Bohorquez et al. [2016] (described in Section 3). We recall that UTrK corresponds to the univariate version of the method proposed in Section 2 (i.e., setting $K = 1$). Similarly, method UCoK is the univariate counterpart of UCoK2; it was developed by Menafoglio et al. [2016] as an extension to the non-stationary setting of Nerini et al. [2010]. Hence, models (i) and (iii) represent the situation in which no secondary data is available together with the target response; however, they pursue different approaches, the former following a trace-approach, the latter a projection-based approach. Note that UCoK and UTrK were thus fitted only on the full physics responses (i.e., without considering the low-fidelity model) since they are univariate functional interpolation methods.

Given that the projection based methods UCoK and UCoK2 can be fitted with a variable number of principal components, we produced models with two (suffix: ".K2"), and three (suffix: ".K3") leading principal components. For parameter inference we used variogram fitting and linear model of coregionalization (LMC, Goovaerts [1997]) on omni-directional variograms computed over the unit cube of re-scaled input parameters, as proposed by Sacks et al. [1989].

The produced statistical models were then used to predict the test set (400

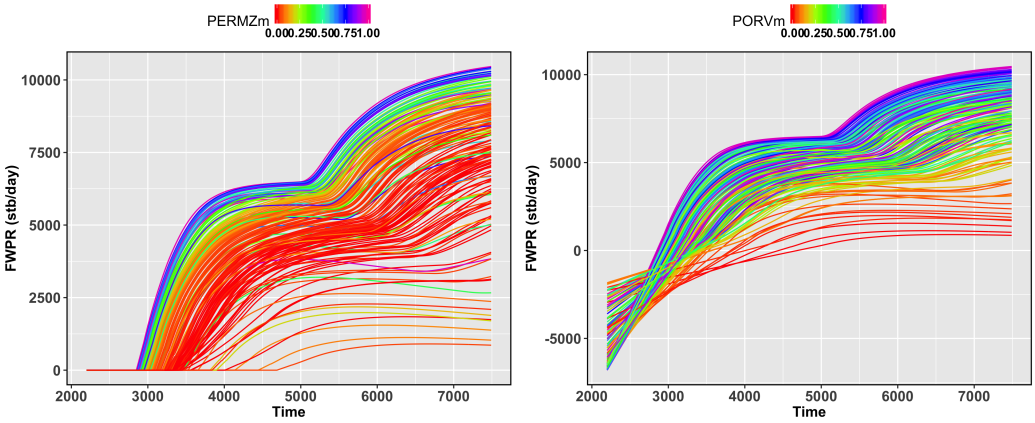

Figure 3: Raw and transformed FWPR curves from 2 parameter dataset. Left - Raw curves colored by PERMZm, Right - Transformed curves colored by PORVm

curves) and summarize the predictions by computing the sum of squared errors (SSE) of each prediction.

$$SSE_i = \|\mathcal{X}_i^{(k)} - \hat{\mathcal{X}}_i^{(k)}\|^2. \tag{16}$$

To better appreciate the magnitude of the error all SSE's were normalized by the average squared norm of the entire test set (400 simulations).

$$SSE_i^n = \frac{SSE_i}{\frac{1}{400}\sum_{i=1}^{400}\|\mathcal{X}_i^{(k)T} - \mu^{(k)T}\|^2} \tag{17}$$

where $\mu^{(k)T}$ is the mean of the test set.

Empirical variograms of the trace based co-kriging and universal co-kriging with secondary data are given in Figures 4 and 5 along with the fits produced with LMC.

Test sets error summary is given in Table 3, and visually in Figure 7. We observe that trace based methods performed slightly better than projection based approaches, and we also observe that incorporation of the secondary data in a form of proxy solution improved the overall SSE.

4.2 Monte Carlo Analysis

To assess the performance under variable training set sizes and different ratios of full physics to proxy simulations we set up a Monte Carlo study. For variable numbers of proxy and full physics simulations we repeated the previous forecasting study one hundred times, and at each step we computed the mean and the median of the test sets SSE's. Distribution of the mean and the median of SSE

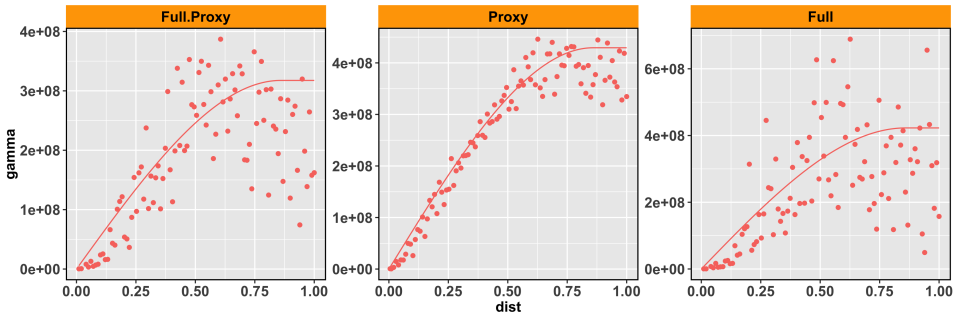

Figure 4: Empirical omni-directional trace variograms and models fitted with the LMC ($Sph(\frac{d}{0.85})$). Left - trace-cross-variogram, middle and right auto trace-variograms

Table 3: 2D dataset - Error Summary Table (normalized SSEs computed with eq (17))

	min	p0.25	p0.5	p0.75	max	mean
Projection Methods						
UcoK.K2	0.0019	0.0080	0.0205	0.0535	2.3807	0.0739
UcoK.K3	0.0004	0.0026	0.0052	0.0112	2.2712	0.0483
UcoK2.K2	0.0018	0.0088	0.0181	0.0486	0.9568	0.0482
UcoK2.K3	0.0005	0.0027	0.0049	0.0086	1.0661	0.0239
Trace Methods						
UTrCoK	0.0000	0.0001	0.0005	0.0034	0.4143	0.0175
UTrK	0.0000	0.0001	0.0007	0.0045	2.2030	0.0416

for each fitting method on the two parameter dataset is shown in Figure 8. The same analysis was performed on the three parameter dataset and its results are shown in Figure 9.

We observe that the median of the SSE was consistently lower for trace based methods compared to projection based methods. We also observe that all methods have similar SSE with a large number of full physics simulations.

5 Case Study: Uranium contamination dataset

In this section, we apply and illustrate the presented computer code emulation techniques on a real case study. The case study considers a numerical model of uranium bio-remediation experiment in Rifle Colorado (Yabusaki et al. [2007], Li et al. [2011], Kowalsky et al. [2012]). The experiment consists of acetate and tracer injection into eleven injection wells and monitoring their concentrations

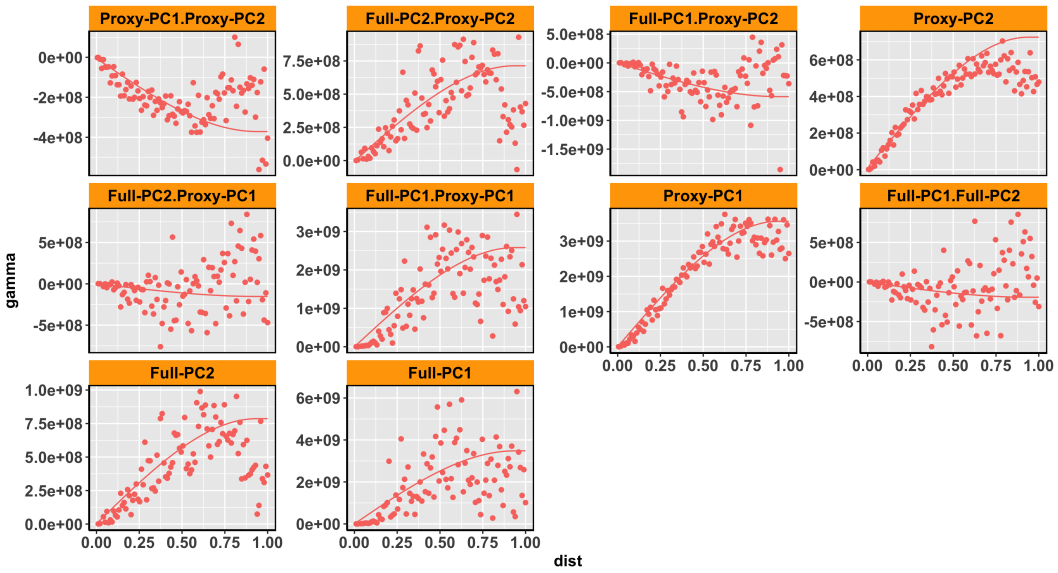

Figure 5: UCoK2: Empirical auto and cross omni-directional variograms and models fitted with the LMC for K=2. ($Sph(d/0.94)$).

at twelve monitoring wells (Figure 10 left). The presence of acetate in the subsurface is known to stimulate biochemical reactions between in-situ bacteria and mobile Uranium U(VI) ions (Williams et al. [2011]), whose product are immobile Uranium U(IV) ions. Since there is no direct way of inferring the volumes of immobilized uranium, indirect inference by means of numerical simulation and inversion is necessary. In particular, spatial distributions of immobilized uranium from the numerical models that match the measured data at monitoring wells can be used to estimate the immobilized volumes of U(VI).

Numerical modeling of bio-remediation is difficult and computationally expensive. One has to consider both geological and geochemical uncertainties and complex physics need to be simulated with advanced reactive transport numerical simulators. Simulation models used in this case study were developed with Crunchflow (Steefel et al. [2015]) a reactive transport simulator. The contaminated site is an unconfined aquifer in alluvial floodplane that was modeled as a single layer with 64x68x1 grid blocks with thickness of about 2.5 meters. We used latin hypercube sampling to vary five input parameters: three geological and two geochemical. Geological parameters are: mean log permeability (meanLogK) of the reservoir, correlation length (CorrL) of reservoir permeability and the variance of reservoir permeability (varK), while geochemical parameters are kinetic rates of microbial reactions: ferric rate (FerricRate) and microbial sulfate reduction rate (SRBrate). The parameters and their ranges are summarized in Table 4. Geological properties were modeled with sequential gaussian co-simulation (coSGS, Verly [1992]), and a total of 500 geological models were developed.

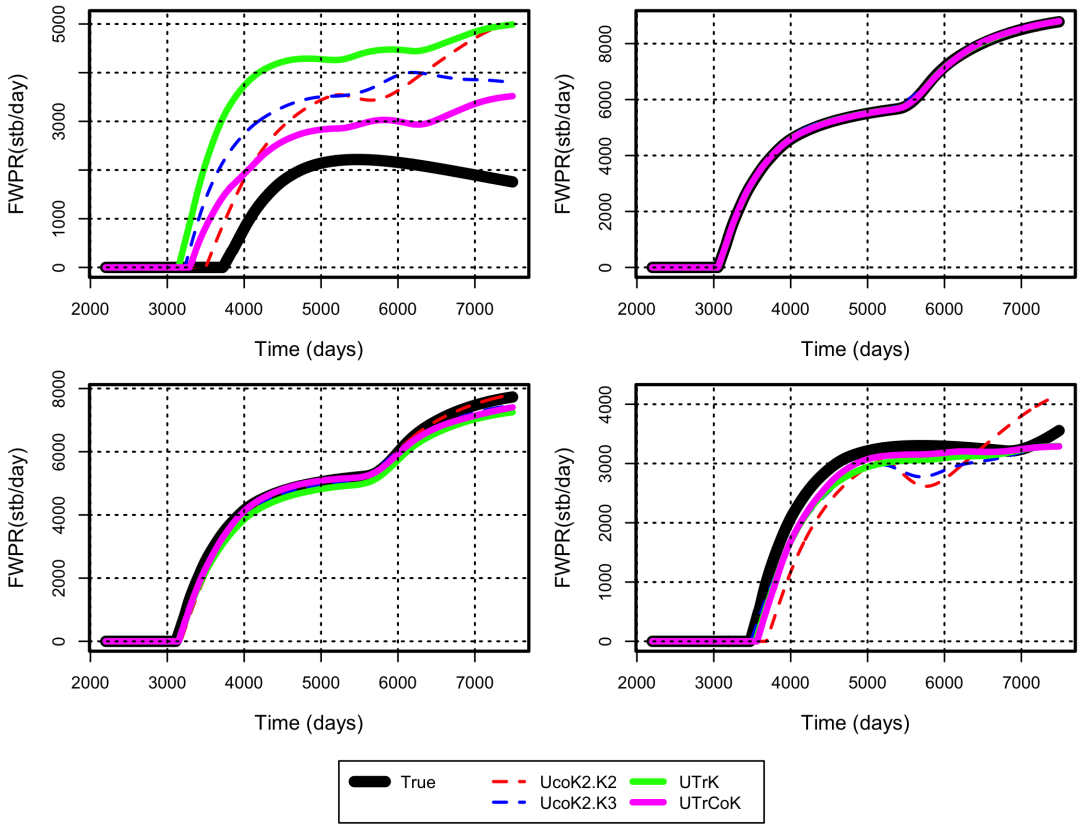

Figure 6: Two parameter dataset: An example of forecasts produced with each meta model at four randomly selected design points.

While this model is fairly small, one simulation run took around 2 hours due to the fact that the modeled physics are very complex. To demonstrate and evaluate our computer code emulation methodology we upscaled/upgridded the models to produce proxy flow simulations. Upgridded models contained 32x34x1 grid blocks and this simplification reduced simulation time to just 10 minutes.

In our analysis we considered simulated acetate concentration curves from monitoring well number 11 (Figure 11). With this data we conducted the same type of Monte Carlo study as we did before on the synthetic reservoir model. The only difference was that in this case we did not have a fixed test set, instead at every iteration we randomly sample for variable numbers of proxy and full physics reservoir models and a non overlapping test set of size 100. In all models we use variogram fitting procedure for parameter inference, and in the case of projection based methods we consider five and six principal components since they capture the most of the variance in this data (98%). A few forecasts produced with the trace based methods on this dataset are given in Figure 12, while the results of

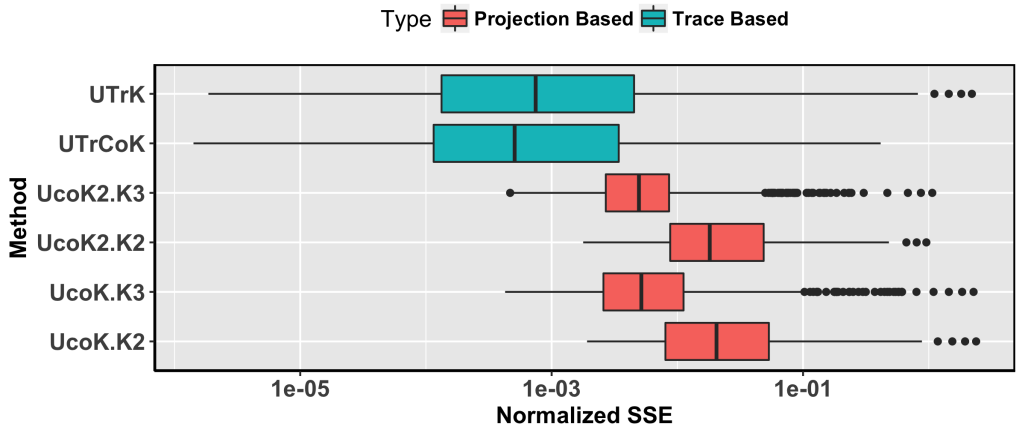

Figure 7: Normalized SSE distribution of each forecasting approach. Note: SSE = sum of squared errors, a dimensionless quantity computed with equation (16)
.

Table 4: Uranium contamination model parameters

Parameter	Range
meanLogK	-10.5 to -10
CorL	3 m - 7 m
varK	0.2 - 0.7
FerricRate	1 - 2
SRBRate	0 - 2

the Monte Carlo study are given in Figure 13.

We observe that the results of the uranium case study are very similar to the results we obtained on synthetic datasets. Trace based approach slightly outperformed the projection based approaches, and in this case there was not much difference between single variate projection based approach (UCoK) and multivariate projection based approach (UCoK2).

6 Conclusions

This paper introduced and analyzed trace co-kriging (UTrCoK), a novel and original method for interpolation of multivariate functional data. The method is useful for emulation of functional variables produced by computer codes of variable degrees of fidelity and numerical speed. The need for multifidelity modeling often arises in uncertainty quantification studies throughout various fields of Earth science, where quick exploration of high dimensional input spaces is necessary. The proposed method is applicable to situations where all computer codes produce

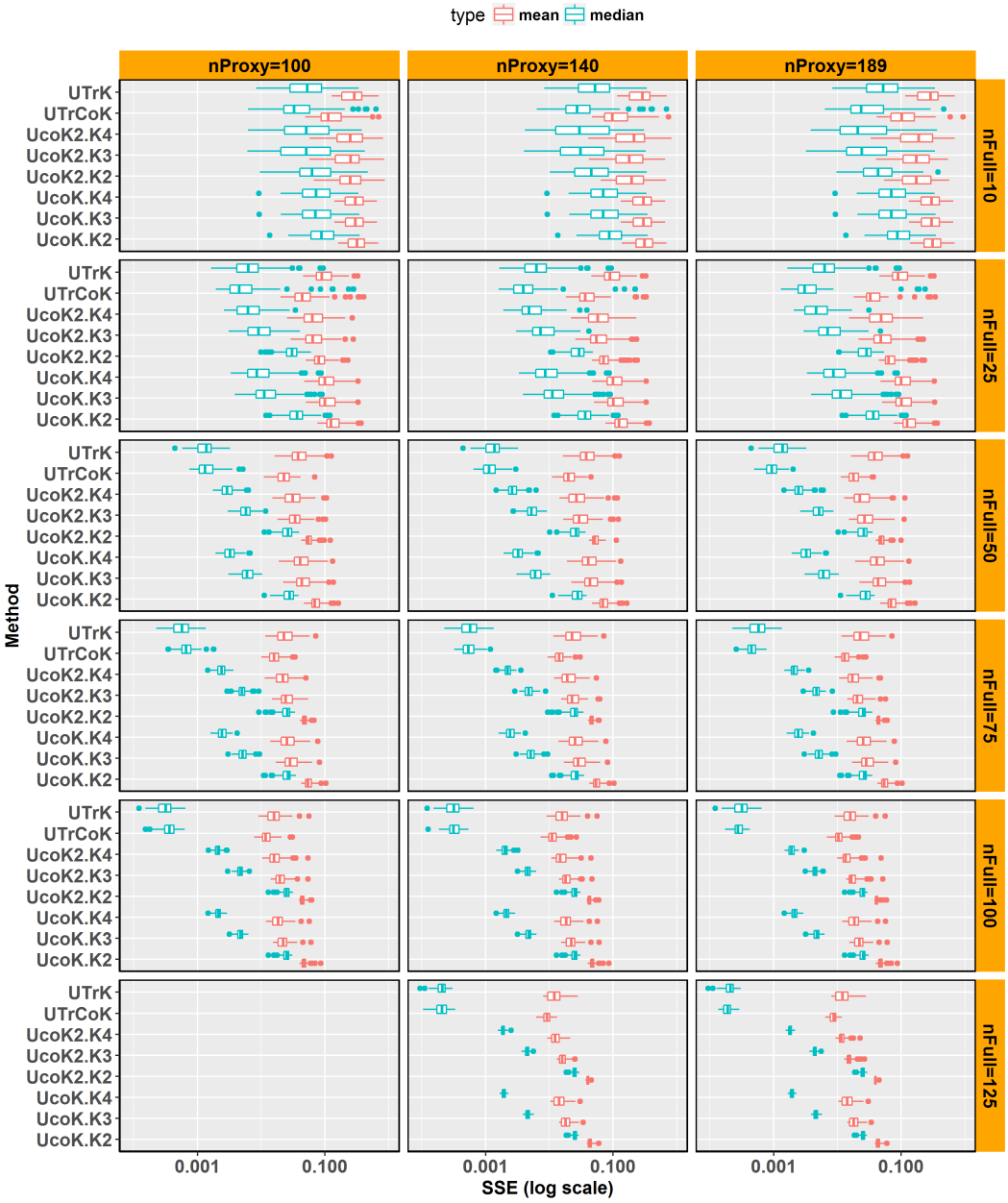

Figure 8: Error analysis of Monte Carlo results on 2 parameter dataset. (Note: SSE = sum of squared errors, a dimensionless quantity computed with equation (16))

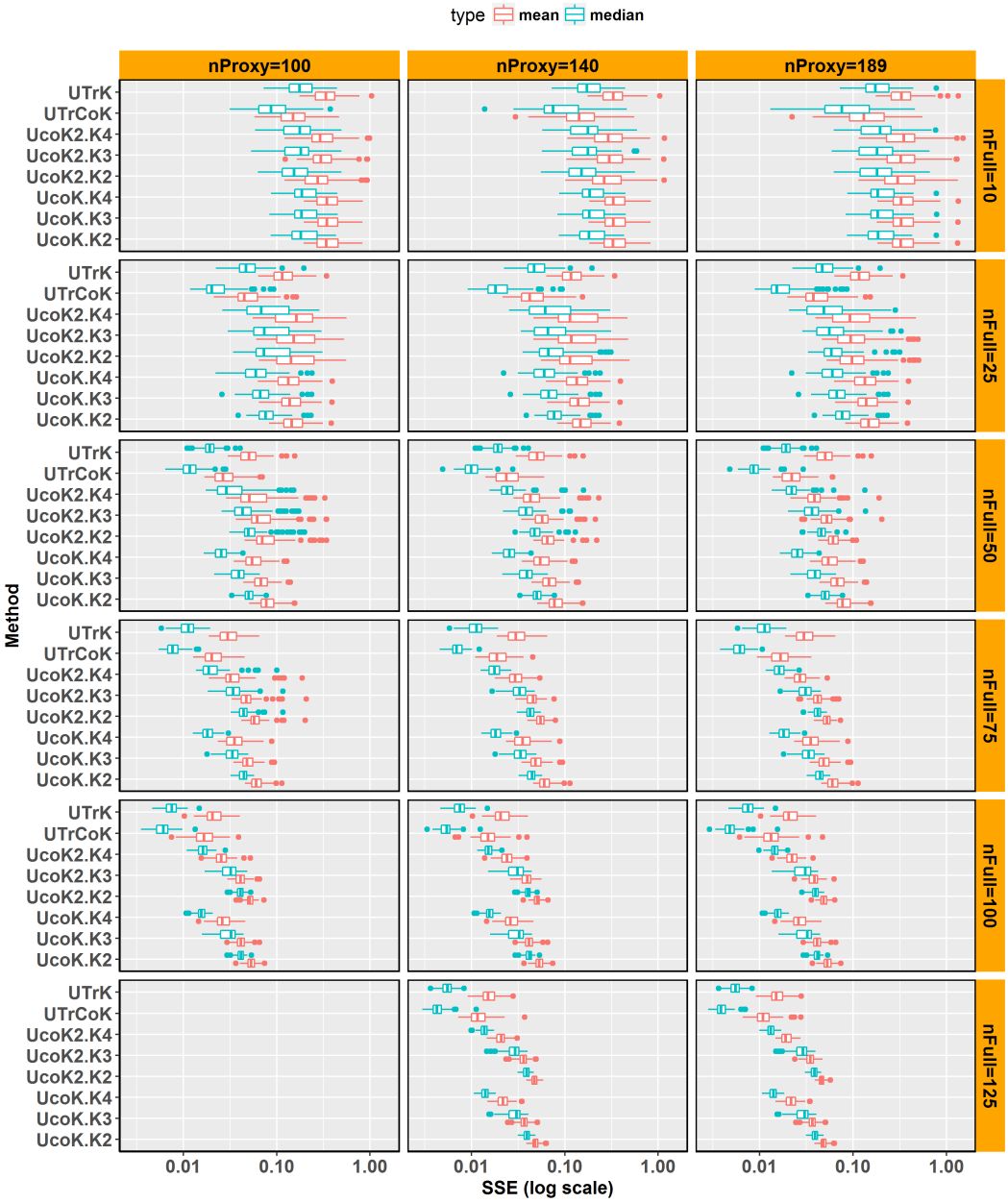

Figure 9: Error analysis of Monte Carlo results on 3 parameter dataset. (Note: SSE = sum of squared errors, a dimensionless quantity computed with equation (16))

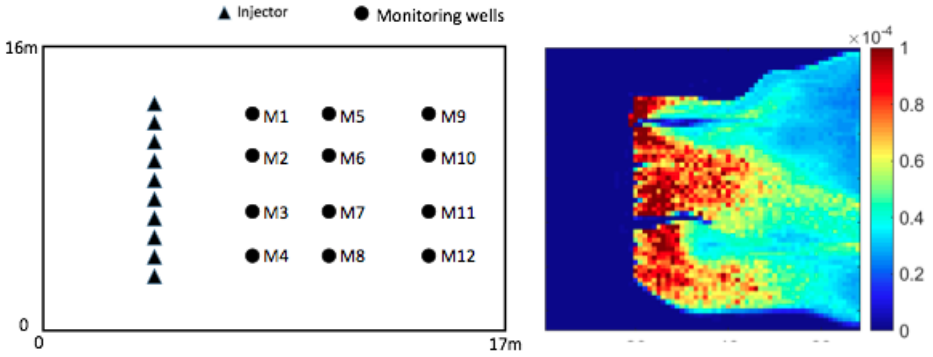

Figure 10: Uranium contamination model. Left - spatial setup (modified from Kowalsky et al. [2012]). Right - A map of immobilized uranium at the end of simulation time

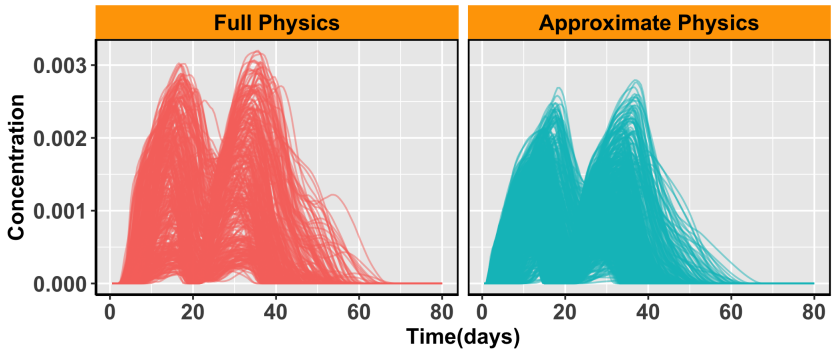

Figure 11: Uranium Dataset: Acetate concentration curves over time computed with full physics and approximate physics simulation.

the same type of functional outputs (i.e. rate vs. time), and where discrepancies between the functions are in amplitude rather than in phase. Nonetheless, the generality of the proposed Hilbert-space approach opens new venues for the meta-modeling and computer emulations of more complex responses, such as distributional responses in the form of PDFs of a target variable. This has the clear potential to offer innovative and groundbreaking perspectives to efficiently and effectively approximate entire distributions, thus avoiding expensive Monte Carlo simulations or restrictive parametric assumptions.

In addition we also introduced a recently developed projection based method for interpolation of multivariate functional data (UCoK2: Bohorquez et al. [2016]), into the context of multifidelity computer code emulation. The two methods were applied to real and synthetic subsurface flow modeling case studies and their solu-

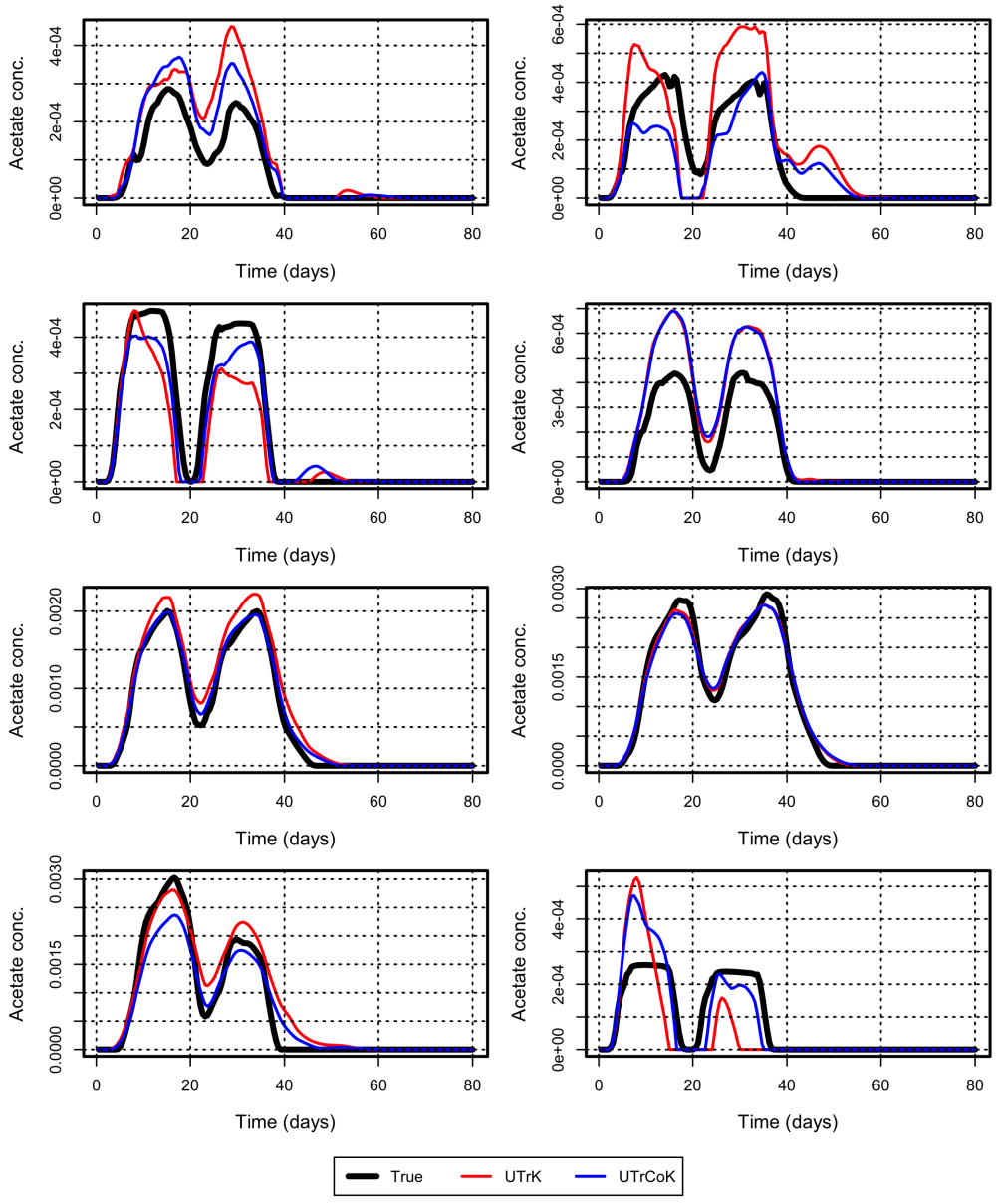

Figure 12: Uranium Dataset: UTrK and UTrCoK forecasts of eight randomly selected design points.

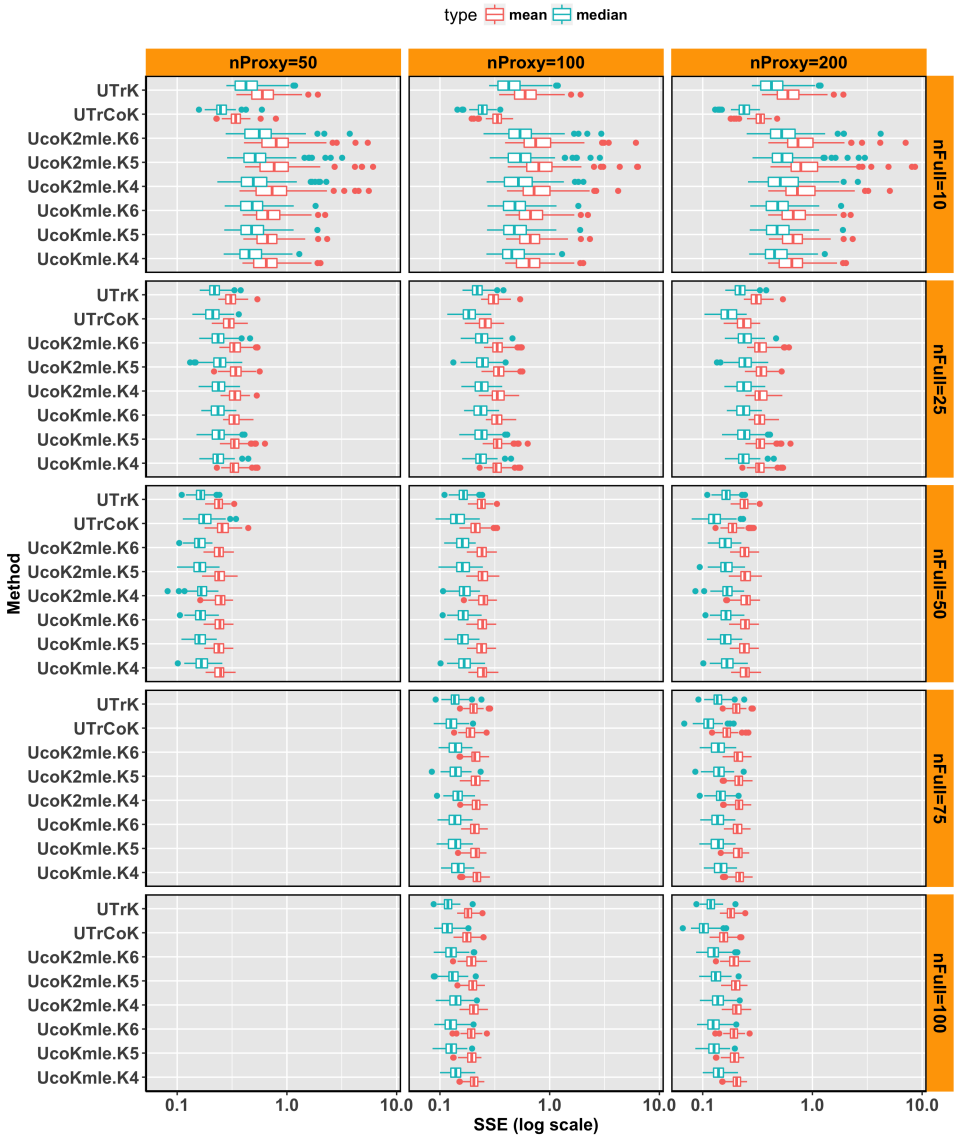

Figure 13: Error analysis of Monte Carlo results on Uranium contamination dataset.
(Note: SSE = Sum of squared errors, a dimensionless quantity computed with equation (16))

tions were then compared to the solutions of another two single variate functional interpolation methods: universal trace kriging (UTrK: Menafoglio et al. [2013]) and universal co-kriging for functional data by (UCoK: Menafoglio et al. [2016], Nerini et al. [2010]). To gain deeper understanding about the ranges of applicability of each method we set up three Monte Carlo studies in which we varied the size of the training sets and the ratios between proxy and full physics simulation runs. Based on the results of our analyses we draw the following conclusions:

- In general UTrCoK performed best out of all considered methods, and particularly better in cases when the number of high fidelity flow simulations was low. This is due to the fact that proxy flow simulations in combination with the linear model of coregionalization (LMC) helped produce better variogram fits.
- UTrCoK requires a much lower modeling effort. Trace variography required LMC fitting over three empirical variograms for two levels of computer code, while projection based method on the same data and with only two principal components on each level of computer code required computing and fitting ten variograms. Automated parameter inference procedures in the context of UCoK2 were not attempted in this work.
- All methods, single and multivariate, converged to the same solution for larger numbers of high fidelity flow simulations. This result suggests that proxy flow simulations become unimportant in the presence of enough full physics simulations, in which case one can approximate the true solution just by means of single variate functional interpolation. This result raises an important practical question of how to estimate this critical number of full physics simulations, or when to stop sampling with the proxy, which will be the scope of future work.
- Projection based methods performed worse than trace based methods for low numbers of high fidelity flow simulations. This poor performance is due to difficulties with estimation of the functional principal components with very low number of training functions.

In our analyses we relied on variogram fitting for parameter inference which was our only option in the case of trace based methods. However, we do recognize the need for the development of an automated procedure for parameter inference of trace based methods. This subject will be in the focus of our future work.

An R package that implements all of the methods discussed in this paper is freely available online at the following location: www.github.com/ogru/fdagstat.

Acknowledgements

The authors thank Alexandre Boucher and Herve Gross from Ar2Tech for valuable suggestions and discussions in early stages of this project. This research

was financially supported by Stanford Center for Reservoir Forecasting (SCRF) research consortia 2015/2016.

References

M. Bohorquez, R. Giraldo, and J. Mateu. Multivariate functional random fields: prediction and optimal sampling. *Stochastic Environmental Research and Risk Assessment*, 31(1):53–70, June 2016. ISSN 1436-3259. doi: 10.1007/s00477-016-1266-y. URL http://dx.doi.org/10.1007/s00477-016-1266-y.

Francesca Bottazzi and Ernesto Della Rossa. A functional data analysis approach to surrogate modeling in reservoir and geomechanics uncertainty quantification. *Mathematical Geosciences*, 49(4):517–540, 2017. ISSN 1874-8953. doi: 10.1007/s11004-017-9685-y. URL http://dx.doi.org/10.1007/s11004-017-9685-y.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics - Modeling Spatial Uncertainty*. Wiley, 1999.

Isobel Clark, Karen Basinger, and William Harper. Muck, a novel approach to cokriging. In Bruce Buxton, editor, *87 Conference on Geostatistical, Sensitivity and Uncertainty Methods for Ground-water Flow and Radionuclide Transport Modeling*, 1987. URL http://drisobelclark.kriging.com/publications/Battelle1987.pdf.

L. De Cesare, D. E. Myers, and D. Posa. Estimating and modeling space-time correlation structures. *Statistics and Probability Letters*, 51:9–14, January 2001. URL http://www.u.arizona.edu/~donaldm/homepage/my_papers/StProbltrs-1.pdf.

Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, April 2010. URL https://projecteuclid.org/euclid.aos/1266586626.

Thomas Fricker, Jeremy Oakley, and Nathan Urban. Multivariate gaussian process emulators with nonseparable covariance structures. *Technometrics*, 55(1):47–56, 2013. doi: 10.1080/00401706.2012.715835. URL http://www.tandfonline.com/doi/abs/10.1080/00401706.2012.715835.

Alan Gelfand, Alexandra Schmidt, Sudipto Banerjee, and C.F. Sirmans. Nonstationary multivariate process modeling through spatially varying coregionalization. *Sociedad de Estadistica e Investigacion Operativa Test*, 13(2):263–312, 2004. URL https://link.springer.com/article/10.1007/BF02595775.

David Ginsbourger, Bastien Rosspopoff, Guillaume Pirot, Nicolas Durrande, and Philippe Renard. Distance-based kriging relying on proxy simulations for inverse conditioning. *Advances in Water Resources*, 52:275 – 291, 2013. ISSN

0309-1708. doi: https://doi.org/10.1016/j.advwatres.2012.11.019. URL http://www.sciencedirect.com/science/article/pii/S0309170812003016.

Ramon Giraldo. *Geostatistical Analysis of Functional Data*. PhD thesis, University of Barcelona, 2009.

Pierre Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, 1997.

K. Hron, A. Menafoglio, M. Templ, K. Hrůzová, and P. Filzmoser. Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, 94:330 – 350, 2016.

L. Josset, D. Ginsbourger, and I. Lunati. Functional error modeling for uncertainty quantification in hydrogeology. *Water Resources Research*, 51(2): 1050–1068, February 2015. URL http://onlinelibrary.wiley.com/doi/10.1002/2014WR016028/abstract.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. URL https://www.jstor.org/stable/2673557.

M. B. Kowalsky, S. Finsterle, K. H. Williams, C. Murray, D. Commer, M. Newcomer, A. Englert, C. I. Steefel, and S. S. Hubbard. On parameterization of the inverse problem for estimating aquifer properties using tracer data. *Water Resources Research*, 48(6):1–25, June 2012. doi: 10.1029/2011WR011203. URL http://onlinelibrary.wiley.com/doi/10.1029/2011WR011203/full.

L. Le Gratiet. Recursive co-kriging model for Design of Computer experiments with multiple levels of fidelity with an application to hydrodynamic. *ArXiv e-prints*, October 2012.

Li Li, Nitin Gawande, Michael B. Kowalsky, Carl I. Steefel, and Susan S. Hubbard. Physicochemical heterogeneity controls on uranium bioreduction rates at the field scale. *Environmental Science and Technology*, 45(23):9959–9966, October 2011. doi: 10.1021/es201111y. URL https://www.ncbi.nlm.nih.gov/pubmed/21988116.

A. Menafoglio and P. Secchi. Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. *European Journal of Operational Research*, 258(2):401–410, 2017.

A. Menafoglio, A. Guadagnini, and P. Secchi. A Kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers. *Stochastic Environmental Research and Risk Assessment*, 28 (7):1835–1851, 2014.

Alessandra Menafoglio, Piercesare Secchi, and Matilde Dalla Rosa. A universal kriging predictor for spatially dependent functional data of a hilbert space. *Electronic Journal of Statistics*, 7(0):2209–2240, 2013. ISSN 1935-7524. doi: 10.1214/13-ejs843. URL http://dx.doi.org/10.1214/13-EJS843.

Alessandra Menafoglio, Ognjen Grujic, and Jef Caers. Universal kriging of functional data: Trace-variography vs cross-variography? application to gas forecasting in unconventional shales. *Spatial Statistics*, 15:39 – 55, 2016. ISSN 2211-6753. doi: http://dx.doi.org/10.1016/j.spasta.2015.12.003. URL http://www.sciencedirect.com/science/article/pii/S2211675315001141.

David Nerini, Pascal Monestiez, and Claude Manté. Cokriging for spatial functional data. *Journal of Multivariate Analysis*, 101(2):409–418, Feb 2010. ISSN 0047-259X. doi: 10.1016/j.jmva.2009.03.005. URL http://dx.doi.org/10.1016/j.jmva.2009.03.005.

Stefano Pagani, Andrea Manzoni, and Alfio Quarteroni. Efficient state/parameter estimation in nonlinear unsteady pdes by a reduced basis ensemble kalman filter. *SIAM/ASA J. Uncertainty Quantification*, 5(1):890–921, 2017.

J. Ramsay and B. Silverman. *Functional Data Analysis*. Springer, 2005.

J.O. Ramsay and Xiaochun Li. Curve registration. *Journal of Statistical Society*, 60:351–363, April 1998. URL http://onlinelibrary.wiley.com/doi/10.1111/1467-9868.00129/abstract.

C.E. Rasmussen and C.K.I. Williams. *Gaussian Process for Machine Learning*. MIT Press, 2006. URL http://www.gaussianprocess.org.

Olivier Roustant, David Ginsbourger, and Yves Deville. Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software, Articles*, 51 (1):1–55, 2012. ISSN 1548-7660. doi: 10.18637/jss.v051.i01. URL https://www.jstatsoft.org/v051/i01.

Jerome Sacks, William Welch, Toby Mitchell, and Henry Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989. URL https://projecteuclid.org/euclid.ss/1177012413.

Céline Scheidt, Jef Caers, Yuguang Chen, and Louis J. Durlofsky. A multiresolution workflow to generate high-resolution models constrained to dynamic data. *Computational Geosciences*, 15(3):545–563, 2011. ISSN 1573-1499. doi: 10.1007/s10596-011-9223-9. URL http://dx.doi.org/10.1007/s10596-011-9223-9.

Mohammad Shahvali, Bradley Mallison, Kaihong Wei, and Herve Gross. An alternative to streamlines for flow diagnostics on structured and unstructured

grids. *SPE journal*, 17, September 2012. doi: 10.2118/146446-PA. URL https://www.onepetro.org/journal-paper/SPE-146446-PA.

C.I. Steefel, C.A.J. Appelo, B. Arora, D. Jacques, T. Kalbacher, O. Kolditz, V. Lagneau, P.C. Lichtner, K.U. Mayer, J.C.L. Meeussen, S. Molins, D. Moulton, H. Shao, J. Šimůnek, N. Spycher, S.B. Yabusaki, and G.T. Yeh. Reactive transport codes for subsurface environmental simulation. *Computational Geosciences*, 19(3):445–478, Jun 2015. doi: 10.1007/s10596-014-9443-x. URL https://link.springer.com/article/10.1007/s10596-014-9443-x.

Arthur Thenon, Véronique Gervais, and Mickaële Le Ravalec. Multi-fidelity meta-modeling for reservoir engineering - application to history matching. *Computational Geosciences*, 20(6):1231–1250, 2016. ISSN 1573-1499. doi: 10.1007/s10596-016-9587-y. URL http://dx.doi.org/10.1007/s10596-016-9587-y.

Sumeet Trehan, Kevin Carlberg, and Louis J. Durlofsky. Error modeling for surrogates of dynamical systems using machine learning. *International Journal for Numerical Methods in Engineering*, July 2017. ISSN 1097-0207. doi: 10.1002/nme.5583. URL http://dx.doi.org/10.1002/nme.5583.

K. G. van den Boogaart, J. J. Egozcue, and V. Pawlowsky-Glahn. Bayes Hilbert spaces. *Australian & New Zealand Journal of Statistics*, 56:171–194, 2014.

G. W. Verly. Sequential gaussian cosimulation: A simulation method integrating several types of information. In Amilcar Soares, editor, *Geostatistics Troia 92*, 1992. URL https://link.springer.com/chapter/10.1007/978-94-011-1739-5_42.

Hans Wackernagel. *Multivariate Geostatistics*. Wiley, 2010.

Kenneth Williams, Philip E. Long, James Davis, Michael Wilkins, A. Lucie N'Guessan, Carl Steefel, Li Yang, Darrell Newcomer, Frank Spane, Lee Kerkhof, Lora McGuinness, Richard D. Dayvault, and Derek Lovley. Acetate availability and its influence on sustainable bioremediation of uranium contaminated groundwater. *Geomicrobiology Journal*, 28(5-6):519–539, July 2011. doi: 10.1080/01490451.2010.520074. URL http://dx.doi.org/10.1080/01490451.2010.520074.

Steven B. Yabusaki, Yilin Fang, Philip E. Long, Charles T. Resch, Aaron D. Peacock, John Komlos, Peter R. Jaffe, Stan J. Morrison, Richard D. Dayvault, David C. White, and et al. Uranium removal from groundwater via in situ biostimulation: Field-scale modeling of transport and biological processes. *Journal of Contaminant Hydrology*, 93(1-4):216–235, Aug 2007. ISSN 0169-7722. doi: 10.1016/j.jconhyd.2007.02.005. URL http://dx.doi.org/10.1016/j.jconhyd.2007.02.005.

Hao Zhang. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, 18(2):125–139, 2007. ISSN 1099-095X. doi: 10.1002/env.807. URL http://dx.doi.org/10.1002/env.807.