Functional error modeling for uncertainty quantification in hydrogeology

L. Josset,[1] D. Ginsbourger,[2] I. Lunati,[1]

Corresponding author: L. Josset, ISTE, University of Lausanne, Switzerland. (laureline.josset@unil.ch)

[1]ISTE, University of Lausanne,

Switzerland

[2]IMSV, University of Bern, Switzerland

Abstract.

Approximate models (proxies) can be employed to reduce the computational costs of estimating uncertainty. The price to pay is that the approximations introduced by the proxy model can lead to a biased estimation. To avoid this problem and ensure a reliable uncertainty quantification, we propose to combine Functional Data Analysis and Machine Learning to build error models that allow us to obtain an accurate prediction of the exact response without solving the exact model for all realizations. We build the relationship between proxy and exact model on a learning set of geostatistical realizations for which both exact and approximate solvers are run. Functional principal components analysis (FPCA) is used to investigate the variability in the two sets of curves and reduce the dimensionality of the problem while maximizing the retained information. Once obtained, the error model can be used to predict the exact response of any realization on the basis of the sole proxy response. This methodology is purpose-oriented as the error model is constructed directly for the quantity of interest, rather than for the state of the system. Also, the dimensionality reduction performed by FPCA allows a diagnostic of the quality of the error model to assess the informativeness of the learning set and the fidelity of the proxy to the exact model. The possibility of obtaining a prediction of the exact response for any newly generated realization suggests that the methodology can be effectively used beyond the context of uncertainty quantification, in particular for Bayesian inference and optimization.

1. Introduction

The major challenge in hydrogeology is to deal with an incomplete knowledge of aquifer properties, which are usually measured only at few, discrete locations. This lack of information makes it impossible to address hydrogeological problems in a deterministic sense. The problem is typically stated in a stochastic framework and Monte Carlo simulations are used to propagate the uncertainty on aquifer properties to the quantities of interest [*Dagan*, 2002]. A typical example is the prediction of the fate of a contaminant, which depends on the heterogeneity structure of the aquifer. The uncertainty on the contaminant breakthrough curve at a given location is estimated by solving the transport problem in a set of realizations, which represent the uncertainty on the permeability of the aquifer. The ensemble of the responses in the different realizations provides a sample of reference of the breakthrough curves.

Despite the appealing conceptual simplicity of this approach, problems arise when many realizations have to be considered and a large number of expensive flow and transport simulations have to be performed: computational cost quickly becomes prohibitive. To avoid this computational bottleneck, the problem is approximated either by coarsening the description of aquifer properties (standard upscaling techniques can be used to this end [*Wen and Gómez-Hernández*, 1996; *Renard and de Marsiliy*, 1997; *Christie*, 1996; *Durlofsky*, 2005]) or by simplifying the description of the physical processes, thus employing an approximate model or proxy (e.f., *Scheidt and Caers* [2009a]).

The price to pay for these simplifications is that inference based on the computed responses could lead to a wrong uncertainty quantification. If the approximation is phys-

ically motivated, the bias can be safely ignored. Effective computational gains, however, usually require very crude approximations whose effects on the uncertainty quantification is difficult to assess beforehand. To avoid this problem, the proxies are typically employed only to identify a representative subset of realizations for which the exact model is solved. This is the strategy of ranking methods [*McLennan and Deutsch*, 2005; *Ballin et al.*, 1992], or distance kernel methods [*Scheidt and Caers*, 2009a]. In such case, it is crucial to evaluate to which extent the proxy is informative of the exact model response.

While it is generally acknowledged that an error analysis is necessary [*Christie et al.*, 2005], it is rarely performed. Although approaches that entail a systematic analysis and the construction of error models have been applied to flow in porous media (e.g., to correct fluid-properties approximations *O'Sullivan and Christie* [2005, 2006] or approximate numerical solvers *Josset and Lunati* [2013]), in most cases the appraisal of approximate methods is performed for a very limited number of test cases, and it is assumed that they behave similarly for a wider range of applications. This approach is not exempt from problems because the informativeness of the proxy also depends on flow regimes and on the specific quantities of interest.

In this paper, we propose a novel methodology to systematically build statistical error models that describe the discrepancy between exact and approximate responses. Once the error model is constructed, it is used to correct the approximate responses and predict the responses expected from the exact model for all realizations. A characteristic of our approach is that the error model is purpose oriented, that is, it is established directly for the quantities of interest (in our case the breakthrough curve of a contaminant) and not for the state of the system (for instance, the full saturation -or concentration- and pressure

fields). This reduces the complexity of the data to be handled (e.g., time-dependent curves rather than time-dependent fields) while retaining all the relevant information.

Despite some similarities with the error models proposed by *Josset and Lunati* [2013], two additional key features characterize the present approach: the description of sparse data as continuous variables (time-dependent breakthrough curves), and the reduced dimensionality of the problem that is solved to construct the error model. To this end we employ Functional Principal Component Analysis (FPCA [*Henderson*, 2006]), which is a functional extension of PCA. The theoretical background is provided by Functional Data Analysis (FDA), a discipline that gathers mathematical tools to construct and treat continuous data. The description of continuous variables from sparse data is a problem faced in many fields of research and not only in environmental applications. While functional analysis is well established, FDA has been integrated as a whole only recently and promoted by *Ramsay* [2006]; *Ramsay et al.* [2009]. It has since been applied in various areas such as biomedical science, biomechanics, medicine or linguistic among others. We refer to *Ullah et al.* [2013] for a recent review of the application of FDA over the last 20 years. More specifically to the domain of groundwater protection problem, FPCA has been applied to interpret various contaminant concentrations in river quality [*Henderson*, 2006].

The paper is organized as follows. After a general problem statement (Sec. 2), we introduce the formalism used and describe the methodology in detail (Sec. 3). Then, the methodology is evaluated for a synthetic test case that represents a typical groundwater problem (Sec. 4). The paper ends with a discussion of the performance and of prospective applications (Sec. 5).

2. Problem statement

We consider a contamination problem in which a non-aqueous phase liquid (NAPL) is accidentally released and forms a plume that contaminates the fresh water. We are interested in predicting the breakthrough curve of the pollutant at a given location (typically a drinking well or a river that can be contaminated). Examples of NAPL contamination are hydrocarbons spills, or leakage of chlorinated solvents such as TCE. As the NAPL is not miscible with water and forms a separate phase, the evolution of the contamination plume is governed by a set of nonlinear transport equations (Appendix A), which complicates both the contaminant behaviour and the numerical resolution of the equations.

Due to sparse measurements, the properties of the aquifer are only partially known. Their uncertainty is represented by a set of N_r geostatistical realizations of the permeability and porosity fields $\{R_i\}_{i=1,\dots,N_r}$. In brute force Monte Carlo approaches, this uncertainty is propagated by solving the nonlinear multiphase transport model (hereafter "exact model") and computing the NAPL breakthrough curve in each realization. Here it is assumed that the resulting set of curves, $\{y_i(t)\}_{i=1,\dots,N_r}$, provides an accurate representation of the uncertainty on the travel time.

Our goal is to find an approximation of the uncertainty without computing the full set of exact curves $\{y_i(t)\}_{i=1,\dots,N_r}$. To this end we use a simplified model based on the linear single-phase transport equations (hereafter "approximate model" or "proxy"), which allows a relatively inexpensive calculation of the approximate breakthrough curves, $\{x_i(t)\}_{i=1,\dots,N_r}$. To provide an accurate approximation of the uncertainty, we need to learn the relationship between the proxy and the exact responses, such that an exact response can be predicted from each proxy response.

We formulate this step in a standard machine learning framework: a statistical model relating the exact response curves (treated as outputs of the statistical model) to the proxy response curves (treated as inputs of the statistical model) is postulated. The parameters are estimated based on a learning set (or training set), i.e., a collection of pairs of response curves obtained with the two models for $N_l < N_r$ geostatistical realizations, $\{(x_i(t), y_i(t))\}_{i=1,\ldots,N_l}$.

The statistical model relating the two sets of response curves (exact and proxy) is here restricted to the class of functional linear models [*Ramsay*, 2006], in which the relationships between the responses is

$$y_i = T(x_i) + \varepsilon_i \qquad i \in [1, \ldots, N_r], \tag{1}$$

where T is a bounded linear operator from the Hilbert space L_2 to itself, and the error functions ε_i are centered, independent, and typically assumed to meet further technical conditions [*Cuevas et al.*, 2002].

Since the identification of such statistical model is ill-posed, in practice further restrictions on the form of T are made introduced to enable inferring T from the learning set. Two methods are suggested by *Ramsay* [2006]; *Ramsay et al.* [2009]: the full functional regression model and the Concurrent model. The full functional regression model allows capturing complex behaviours, but it is costly and requires the fine tuning of several smoothing parameters. The Concurrent model consists of a simpler functional linear regression. This method is fast, but quite rudimentary because the model uses only concurrent features of the curves (additional details about the two models can be found in Appendix B).

In this paper, we follow a slightly different strategy: we appeal to a spectral approach and decompose the elements of the learning set on two *ad hoc* bases, one for the proxy and one for the exact responses. The response curves are then described in two spaces of dimensions $D_{ex} < N_l$ for the exact responses and $D_{app} < N_l$ for the proxy responses. A statistical model is constructed to relate the coefficients of the elements of one space, $y_i(t)$, to the coefficients of the elements of the other space, $x_i(t)$, as illustrated in Fig. 1.

Once the approximation $\hat{T}$ of T is obtained from the learning-set, it is used to predict the exact responses of all realizations from of the approximate responses, i.e.,

$$\{\hat{y}_i = \hat{T}(x_i)\}_{i=1,\dots,N_r}, \tag{2}$$

and the uncertainty is quantified from the ensemble of predicted curves.

3. Methodology

The construction of the error model consists of four steps: first, functional objects are built from the data in the learning set; second, the dimensionality of the problem is reduced by decreasing the dimensions of the two functional spaces; third, the relationship between the approximate and exact responses is constructed; fourth, the error model is used to predict the exact responses from the proxy responses. These steps are illustrated in the flowchart in Fig. 2.

3.1. Recasting discretized curves as functional data

Both exact and proxy responses are obtained from numerical simulations and are represented by contaminant breakthourgh curves defined at discrete times. Therefore, we recast the time-discrete curves into time-continuous functions. This has two practical advantages: first, it allows us to use the formalism of functional data analysis and the tools

that have been developed in this context; second, it permits to work with asynchronous information about the curves, i.e., curves that have been sampled at different times. Note that this step is essential in applications in which analytic solutions are used as proxies or if the exact responses are provided by field measurements, which are typically acquired with different temporal resolution.

Many functional bases are available to recast discretized curves into functional data. Here, we use a K-dimensional B-spline basis denoted by $\{\varphi_k(t)\}_{k\in[1,K]}$. To determine the coefficients, a linear combination of the elements of this basis is fitted to the data, which are represented as time dependent functions of the form

$$f(t) = \sum_{k=1}^{K} c_k \varphi_k(t) \tag{3}$$

Ramsay [2006] suggests two strategies to choose the basis and fit the coefficients to data: either a low-dimension basis is used and the data are plainly projected (e.g., by ordinary least squares), or a high-dimension basis is used with a roughness penalty to mitigate overfitting. Both strategies allow not only to distinguish noise from information but also to impose various constraints on the functional objects, e.g. positivity and/or monotonicity. As our data (contaminant breakthrough curves) are typically fairly smooth, a standard B-spline basis of small dimension can be used. We refer the readers to [*Ramsay*, 2006; *Ramsay et al.*, 2009] for more details about the notions of roughness penalty and incorporation of constraints.

3.2. Functional reduction of the dimensionality

The previous step allows representing each exact response and each proxy response as a continuous function, i.e., $y_i(t)$ and $x_i(t)$, respectively. To decrease the dimension of the response spaces and the size of the regression problem, we employ Functional Principal Component Analysis, which is a functional extension of standard PCA and allows highlighting the main modes of variability in a sample of functions. Beside a small computational advantage, using spaces of lower dimension reduces the risk of over-fitting and allows us to visualize the data to assess the informativeness of the proxy response with respect to the exact response.

We apply FPCA to the exact and proxy responses in the learning set. Given the sample of proxy functions in the learning set, $\{x_i(t)\}_{i=1,\dots,N_l}$, with average $\bar{x}(t) = \frac{1}{N_l}\sum_{i=1}^{N_l} x_i(t)$ and estimated covariance function

$$\nu(t',t) = \frac{1}{N_l - 1}\sum_{i=1}^{N_l}[x_i(t') - \bar{x}_i(t')][x_i(t) - \bar{x}_i(t)], \tag{4}$$

FPCA constructs a non increasing sequence of eigenvalues of the estimated covariance function, $\mu_1^\circ \geq \mu_2^\circ \geq \cdots \geq \mu_{N_l-1}^\circ$, by solving the functional eigenequation

$$\int \nu(t',t)\zeta_i^\circ(t)dt = \mu_i^\circ \zeta_i^\circ(t'). \tag{5}$$

The sequence of eigenfunctions (or harmonics) of the covariance function, $\{\zeta_1^\circ, \dots, \zeta_{N_l-1}^\circ\}$, satisfies the condition

$$\int \zeta_i^\circ(t)\zeta_j^\circ(t)dt = \delta_{ij}, \tag{6}$$

(where δ_{ij} is the Kronecker delta), and, together with the average $\bar{x}(t)$, form an orthonormal basis for the space of the sampled approximate responses. The eigenvalue μ_i is also denoted as the *probe score variance* and the eigenfunction $\zeta_i^\circ(t)$ as *harmonic* [*Ramsay*

179 *et al.*, 2009]. The dimensionality of the response space can be optimally reduced consider-
180 ing only the first D_{ex} and D_{app} for the exact response space and the proxy response space,
181 respectively. The fact that the sequence of eigenvalues is non increasing guarantees that
182 no other basis of size D_{app} can describe better the data; the total squared error introduced
183 by discarding the eigenfuncions $(\zeta_i^\circ(t))_{i>D_{app}}$ is $\sum_{i=D_{app}+1}^{N_l-1} \mu_i^\circ$.

184 The basis allows us to approximate each proxy response as

$$x_i(t) \approx \tilde{x}_i(t) = \bar{x}(t) + \sum_{j=1}^{D_{app}} b_{ij}^\circ \zeta_j^\circ(t) \tag{7}$$

where

$$b_{ij}^\circ = \int [\bar{x}(t) - x_i(t)] \zeta_j^\circ(t) dt \tag{8}$$

185 is the projection of the deviation from the mean of the i^{th} approximate curve on the j^{th}
186 harmonic ($\tilde{x}_i(t)$ denotes the approximation of $x_i(t)$ in terms of the first D_{app} harmonics).
187 As in standard PCA, these coefficients are typically referred to as *scores*.

Although it offers an optimal dimensionality reduction with respect to the total mean squared error, the orthonormal basis might not be ideal to represent the information. The *varimax* algorithm [*Kaiser*, 1958] can be applied to find a suitable rotation that improve data interpretation while preserving the optimality of the result in terms of explained variance [*Richman*, 1986; *Ramsay et al.*, 2009]. Therefore, without any further loss of information, the approximate curves can be written as

$$\tilde{x}_i(t) = \bar{x}(t) + \sum_{j}^{D_{app}} b_{ij} \zeta_j(t), \tag{9}$$

where

$$b_{ij} = \int [\bar{x}(t) - x_i(t)] \zeta_j(t) dt \tag{10}$$

188 is the projection of the deviation from the mean of the i^{th} curves on the rotated harmonic
189 $\zeta_j(t)$.

An analogous procedure is applied to the sample of exact responses in the learning set, $\{y_i(t)\}_{i=1,\ldots,N_l}$, which is approximated as

$$\tilde{y}_i(t) = \bar{y}(t) + \sum_{j}^{D_{ex}} c_{ij}\eta_j(t), \tag{11}$$

where $\bar{y}(t)$ is the average, $\eta_j(t)$ the j^{th} harmonic of the (varimax) rotated orthonormal basis $\{\eta_i(t)\}_{i=1,\ldots,D_{ex}}$, and

$$c_{ij} = \int [y_i(t) - \bar{y}(t)]\eta_j(t)dt \tag{12}$$

190 the score with respect to $\eta_j(t)$. (As for the proxy curve, the *tilde* denotes the restriction
191 to the first D_{ex} harmonics).

3.3. Regression and error model

192 Once the problem dimensionality has been reduced by FPCA, we investigate the rela-
193 tionships between the two sets of curves in the learning set approximated by considering
194 the first D_{app} and D_{ex} harmonics, $\{\tilde{x}_i(t), \tilde{y}_i(t)\}_{i=1,\ldots,N_l}$. The goal is to find a transforma-
195 tion between the spaces of exact and proxy responses. (Notice that the *varimax* rotation
196 does not affect the representation of the curves, but might affect the quality of the trans-
197 formation).

Here, we restrict ourselves to functional linear regression models of the form given in Eq. 1. Training such a functional linear model in full generality is not straightforward. A simple choice to restrict the class of linear regression models is to postulate that, at any time t, $\tilde{y}_i(t)$ depends on $\tilde{x}_i(t)$ solely through its value at that time t. This assumption

leads to the Concurrent model

$$\tilde{y}_i(t) = \beta_0(t) + \tilde{x}_i(t)\beta_i(t) + \varepsilon_i(t), \tag{13}$$

which is a particular case of the functional linear model in Eq. 1 and corresponds to $T(x_i)(t) = \beta_0(t) + x_i(t)\beta_i(t)$. The Concurrent model will be used as baseline in our numerical application, and compared to our FPCA-based prediction approach.

To simplify the exposition, in the following we assume that the same number of harmonics is retained for the two spaces, i.e., $D = D_{ex} = D_{app}$. However, the number of harmonics depends on the inherent variability of the learning set, which can be different for the exact and proxy responses. Ultimately, the number of harmonics to be employed depends on how rapidly the eigenvalues of the FPCA decomposition decrease for the specific problem. It has to be chosen large enough to guarantee an exhaustive representation of the variability of the response curves, but small enough with respect to the number of elements in the learning set to avoid over-fitting when the regression model is constructed.

Given $N_l \leq N_r$ pairs of accurate and proxy responses, $\{(\tilde{x}_i(t), \tilde{y}_i(t))\}_{i=1,\ldots,N_l}$, we postulate that there exists a $(D+1) \times D$ matrix of real-valued coefficients $\boldsymbol{\beta}$ (with line index starting at 0, by convention) and a $N_l \times D$ error matrix $\mathbf{E}$, such that for any $(i,j) \in [1, N_l] \times [1, D]$,

$$c_{ij} = \beta_{0j} + \sum_{\ell=1}^{D} b_{i\ell}\beta_{\ell j} + e_{ij}, \tag{14}$$

where β_{ij} and e_{ij} are the components of $\boldsymbol{\beta}$ and $\mathbf{E}$, respectively. The errors, e_{ij}, are implicitly assumed to be Gaussian with zero mean and variance σ_j^2, which depends only

on j. In matrix notation, the statistical model reads

$$\mathbf{C} = \mathbf{B}\boldsymbol{\beta} + \mathbf{E}, \tag{15}$$

where $\mathbf{C}$ is the $N_l \times D$ matrix containing the scores of the exact responses, c_{ij}, and $\mathbf{B}$ is the $N_l \times (D+1)$ with elements of the first column $b_{i0} = 1$ by convention, and containing the scores of the proxy responses $b_{i(j-1)}$.

In the statistics literature, solving Eq. 15 for the coefficient matrix $\boldsymbol{\beta}$ is referred to as a *multivariate multiple regression* problem ([*Fox and Weisberg*, 2011; *Hastie et al.*, 2009]). A simpler regression problem can be obtained by separating the regression models for the D responses, hence solving D independent regression problems

$$\mathbf{C}_{(j)} = \mathbf{B}\boldsymbol{\beta}'_{(j)} + \mathbf{E}'_{(j)} \quad (1 \leq j \leq D), \tag{16}$$

where $\mathbf{C}_{(j)}$ is the j^{th} column of the score matrix $\mathbf{C}$. A very convenient fact is that the columns of the Ordinary Least Squares (OLS) estimator of $\boldsymbol{\beta}$ coincides with the concatenated OLS estimators of $\boldsymbol{\beta}'_{(j)}$ [*Hastie et al.*, 2009], that is

$$\widehat{\boldsymbol{\beta}}_{(j)} = \widehat{\boldsymbol{\beta}'}_{(j)} \quad (1 \leq j \leq D), \tag{17}$$

where $\widehat{\boldsymbol{\beta}}_{(j)}$ are the columns of the OLS estimator $\widehat{\boldsymbol{\beta}}$ (hereafter, the hat denotes the OLS estimator of the quantity). However, test statistics and confidence bands of the multivariate regression model cannot be directly derived from those obtained for the multiple linear regressions in Eq. 16 and have to be computed for the general regression model in Eq. 15. The formula of the simultaneous confidence bands is given in appendix C, together with a brief outline of the derivation.

3.4. Prediction of the exact response from the proxy response

Once the OLS estimator $\hat{\boldsymbol{\beta}}$ has been obtained, the regression model is used to predict the exact response for all N_r geostatistical realizations on the basis of the corresponding proxy responses $\tilde{x}_i(t)$. The predicted exact response for the i^{th} realization is

$$\hat{y}_i(t) = \bar{y}(t) + \sum_{j=1}^{D} \hat{c}_{ij}\eta_j(t). \tag{18}$$

where

$$\hat{c}_{ij} = \hat{\beta}_{0j} + \sum_{\ell=1}^{D} \hat{\beta}_{j\ell} b_{i\ell}, \tag{19}$$

are the estimates of the scores with respect to the rotated harmonics.

The estimator of the linear regression model allows us to predict the $\hat{c}_{ij}$ scores solely from the scores b_{ij} of the proxy responses, hence predicting $\tilde{y}_i(t)$ without solving the exact model. We emphasize the difference between the proxy response $x_i(t)$ (or $\tilde{x}_i(t)$, which is the projection onto the lower dimensional space defined by the first D harmonics, $\{\zeta_j\}_{j=1,\ldots,D}$), and the predicted exact response $\hat{y}_i(t)$: they both approximate the "true" response $y_i(t)$, but, while $x_i(t)$ is simply the result of the proxy model and lives in the space defined by the basis of the proxy curves, $\hat{y}_i(t)$ results from applying the error models to the proxy response and lives in the space of the exact responses (more precisely: in the subspace defined by the orthonormal basis formed by the first D harmonics, $\{\eta_j\}_{j=1,\ldots,D}$).

Surrogating $y(t)$ by $\hat{y}(t)$ is prone to errors: first, $\{\eta_i(t)\}_{i=1,\ldots,N_l}$ depends on the quality of the learning set; second, the subspace of the prediction is further reduced by considering only the first D harmonics; third, the coefficients $\hat{c}_{ij}$ are predicted through the OLS estimator of a linear regression model, and thus entails statistical uncertainties and possibly systematic errors due to the choice of a simple linear model.

4. Numerical test case: An idealized NAPL pollution problem

We consider an idealized groundwater pollution problem in which the fate of a NAPL plume has to be predicted. We model a portion of aquifer as a vertical 2D domain of length $10.8m$ and depth $5.1m$ discretized into cells of size $10cm \times 10cm$. Gravity effects are neglected, which implies that the density of the NAPL phase is equal to the water density. No-flow boundary conditions are imposed at the upper and lower boundaries, whereas the pressure is fixed at the right boundary. The contaminant is released at the left boundary (a constant influx is assigned) and displaces the water initially present in aquifer. We are interested in the time evolution of NAPL saturation at the right boundary. Two cases are investigated; first, we estimate the uncertainty on the contaminant breakthrough curve computed by averaging the saturation along the right boundary; then, we consider a single-point breakthrough curve obtained by sampling the saturation in a single cell (Sec. 4.5.2). As the NAPL is immiscible with water, the exact model solves the multiphase flow and transport equations, which require solving a pressure equation and a highly nonlinear phase-transport equation [see, e.g., *Marle*, 1981; *Helmig*, 1997]. The two equations are highly coupled and characterized by fluxes that exhibit a non-linear dependence on NAPL saturation. (The full system of equations is described in Appendix A.)

The uncertainty on the transport properties of the aquifer (permeability and porosity) is represented by a set of $N_r = 1000$ geostatistical realizations that are generated by a multipoint geostatistical method (DeeSse) [*Mariethoz et al.*, 2010] with a training image obtained from data of facies-distribution collected at the Herten site (Germany) [*Bayer et al.*, 2011]. As an example, three realizations are shown in Fig. 3.

4.1. The proxy model

The proxy model simplifies the physics of the problem by treating the NAPL as an ideal tracer, thus solving a linear transport problem. Although it is possible to further improve the computational efficiency by simplifying the description of the heterogeneity (e.g., by some upscaling or multiscale methods [see, e.g., *Josset and Lunati*, 2013]), here we do not approximate the aquifer properties.

In practical situations, replacing a multiphase flow problem by a single-phase (tracer-transport) problem considerably reduces the computational costs. Indeed, a large part of the cost of solving the flow and transport system stems from the solution of the elliptic (or parabolic) equation that governs the pressure. Due to the effects of the saturation on the fluxes, this equation has to be solved at every time step in multiphase problems. In contrast, if the pollutant is considered as an ideal tracer, the saturation does not impact the velocity, and the pressure equation has to be solved only once. The NAPL transport equation becomes linear and can be solved very efficiently by streamline methods (here, we use a Finite-Volume upwind scheme that can be seen, in some sense, as a very rudimentary streamline method without sub-grid interpolation of the velocity field).

4.2. The learning dataset

After the proxy responses have been obtained by solving the ideal transport problem and computing the contaminant breakthrough curves for the whole sample of 1000 realizations, we construct the learning set by identifying a subset of $N_l = 20$ realizations. The realizations can be selected in several ways, including a simple random choice. Here, we use a clustering technique to group the proxy responses based on their l_2-distance, and we choose the k-medoid curves as representative of the clusters (Distance Kernel Method

[*Scheidt and Caers*, 2009a]). The medoids define the subset of realizations, $\{R_i\}_{i=1,\ldots,N_l=20}$, for which the exact responses are computed by solving the multiphase transport problem. Additional tests (not reported here) with learning sets consisting of $N_l = 50$ and $N_l = 100$ realizations did not show a significant improvement of the quality of the learning set. This suggests that only 20 realizations are sufficient to obtain a satisfactory error model for the present test case. Cross validation tests can be performed to identify the optimal size of the learning set.

As the numerical NAPL breakthrough curves are discrete in time, a spline basis is defined to interpolate the discrete data and construct the functional objects. In the present test case, data points are fairly smooth and a rather small number of basis functions is necessary for an accurate representation of the data (here, only 50 splines are used as basis functions). The 20 pairs of spline-interpolated proxy and exact curves in the learning set, $\{(x_i(t), y_i(t))\}_{i=1,\ldots,N_l=20}$, are shown in Fig. 4.

4.3. Understanding the data using FPCA

To extract the relevant information from the data and to reduce the problem dimensionality, we apply FPCA independently to both sets of approximate and exact curves in the learning set. As in standard PCA, if all the components (harmonics) are considered, no approximation is made and the data are represented exactly. However, the eigenvalues of higher order harmonics decrease so fast that the first three components describe more than 97% and 99% of the variability of proxy and exact curves, respectively. In the subspaces defined by the first three harmonics, each curve is described by the corresponding three scores and by the sample means. To improve the interpretability of the data, a rotation

297 is sought with the varimax algorithm [*Ramsay et al.*, 2012]. The rotated harmonics for
298 both sets of curves are shown in Fig. 5.

299 In the subset of the exact responses, the first rotated component explains the devi-
300 ation from the mean behavior measured at late time. The second rotated component
301 describes the variation at the beginning of the breakthrough curve, thus enlightening high-
302 connectivity paths. The third component explains the variation observed at intermediate
303 time. In the proxy subset, the first rotated component describes the initial variability; the
304 second component highlights the variation at high saturation; and the third component
305 explains the variation observed at intermediate time. By analyzing the projection of the
306 curves on these components, it is possible to gain information about the data, for instance
307 about the link between the early-time responses and the late-time variations. We refer to
308 *Henderson* [2006] for an example in hydrology.

4.4. Regression model and evaluation of the proxy

309 The linear regression model is built between the scores of proxy and exact curves, which
310 represent their coordinates with respect to the two orthonormal bases formed by the first
311 three harmonics. Three linear regression problems (one for each exact-response score,
312 $j = 1, 2, 3$) are solved to establish a relationship with the three proxy-response scores.
313 The resulting coefficients of the three regression models are

	β_{0j}	β_{1j}	β_{2j}	β_{3j}	R^2	p-value
$j=1$	$-2.3 \cdot 10^{-16}$	0.42	0.18	-0.37	0.99	$< 2 \cdot 10^{-16}$
$j=2$	$4.4 \cdot 10^{-17}$	0.82	-0.02	0.37	0.99	$< 2 \cdot 10^{-16}$
$j=3$	$1.6 \cdot 10^{-16}$	0.51	0.03	0.08	0.97	$1.3 \cdot 10^{-12}$

(20)

314 Notice that the R^2 values are quite high and that $\beta_{0j} \approx 0$, which suggests that the linear
315 regression model preserves the mean. The dependency among scores is illustrated in Fig.

316 6. The relationships between the scores of the three harmonics of the exact curves and the
317 scores of the first harmonic of the proxy curves are rather well approximated by the linear
318 regression. The scores of the second harmonic of the proxy curves are less important as
319 it is indicated by the low values of β_{22} and β_{23}. This might be due to the fact that the
320 proxy second harmonic explains the variability of the curves for saturations close to one,
321 a situation that is not observed in the two-phase responses.

4.5. Performance of the regression model as error model

322 In general, the proxy-curve scores are informative of the exact-curve scores, at least
323 for the curves pairs in the learning set. This suggests that, despite the rather primitive
324 physical model employed, the regression model can be effectively used to predict the
325 exact responses of the realizations for which only the proxy solution is available. The
326 exact response is predicted on Eqs. 18 and 19.

327 4.5.1. Prediction of the average breakthrough curve at the outlet

328 We start by considering the prediction of the breakthrough curve calculated by averaging
329 the saturation at the right-hand boundary. Examples of two predicted curves are shown in
330 Fig. 7a and b. Despite the fact that the curves are very different for the two realizations,
331 both predictions are in good agreement with the exact responses. In general, the behaviour
332 of the exact response is well predicted, with the exception of some fluctuations at early
333 times. The error model greatly improves the proxy solution and provides a much better
334 prediction than the Concurrent model, which is unable to significantly modify the shape
335 of the curves due to the use of only concurrent information.

336 The differences between predicted and exact curves are illustrated in Fig. 7c for all
337 $N_r = 1000$ realizations, together with the mean error. The maximum differences in the

saturation are observed at early time and are about 10%; later, the saturation discrepancy remains below ±1.8% for 68% of the realizations and below ±4% in the worst cases. The mean error is very close to zero, which shows that the predicted curves conserve the mean behaviour of the exact curves, and that the subset of 20 realizations selected in the learning set is representative of the whole sample to describe the mean behaviour.

Fig. 8a shows the histograms for the l_2-norm and the l_∞-norm of the errors. We compare the performance of the error model based on FPCA with the Concurrent functional linear regression model. The histogram of the l_∞-norm shows that on average the maximum deviation is 4.5% for FPCA, and about 8% for the Concurrent model. The l_2-error is on average more than three times lower for the FPCA-based model.

In many applications, the uncertainty is quantified in terms of the quantiles of the responses. Fig. 7d displays the quantile curves obtained using the different models. The Concurrent model fails to reproduce the 90^{th} percentiles, because it is unable to modify the plateau of the proxy curves close to saturation one; it performs better for the other quantiles. The quantiles curves computed using only the learning set of exact responses (as suggested by [*Scheidt and Caers*, 2009a, b]), are slightly biased estimates of the exact quantiles. An excellent estimate is obtained with the functional error model, which is able to correct the approximate responses and predicts quantiles close to the exact ones.

4.5.2. Prediction of single-point breakthrough curve

In this second test case, we are interested in predicting the breakthrough curve of the contaminant at a precise location, defined by a single cell of the numerical grid, which is located at mid-depth at the outlet. In contrast to the breakthrough curves averaged over the whole outlet, in which the effects of extreme permeability structures (flow barriers

or preferential pathways) are smoothed, the single-point breakthrough curves display a variety of shapes. The large contrast in permeability and in connectivity at the sampling location leads to important differences, particularly in the first arrival time.

In this case, it is useful to apply a translation in time to redefine the origin, which is chosen to be the first arrival time. This procedure is referred to as registration in the FDA literature *Ramsay* [2006]; *Ramsay et al.* [2009]. For the translated responses in the learning set FPCA is then applied and the dimensionality is reduced as described above. Again, we use the first three harmonics, which describe more than 98% of the variability of the shape of the curves after the registration. An example of proxy, predicted and exact curves after registration is shown in Fig. 9a for a realization that does not belong to the learning set.

Beside the prediction of the shape, it is now necessary to predict the first arrival time and translate back the predicted curves. The first arrival time is predicted jointly to the scores of the harmonics by solving a 4×4 regression model, where the 4^{th} dimension is the first arrival times of the proxy responses, which have been used for the registration. Fig. 9b compares the proxy and exact curves with the predicted curve after translation by the predicted arrival time (these curves correspond to the registered curves in Fig. 9a). For the whole sample of realizations, the mean saturation error is close to zero and with a standard deviation that remains below 0.04 (Fig. 9c).

The predicted quantile curves (shown in Fig. 9d) are in good agreement with the exact quantile curves for P50 and P90, but P10 is biased. As the concurrent model would perform very poorly in this case because it is unable to deal with curves characterized by different arrival times, we compare our methodology with the quantile curves obtained

directly from the exact response in the learning set (this procedure corresponds to the classical DKM). As both the functional error model and the DKM estimates depend on the clustering, we have applied both methodologies 200 times. The example shown in Fig. 9 is representative of the typical behaviors of the methods (i.e., the quantiles are close to the average quantiles obtained form the 200 applications of the methods shown in Fig. 9e and d. In average, the functional error model is more robust than DKM and provides a better prediction of the P10 quantile curve.

4.5.3. Effects of the number of harmonics

Here, we investigate the effects of the number of harmonics on the prediction of single-point breakthrough curves. In order to increase the difficulty of the problem, we do not apply the registration as in the previous section (i.e., the breakthrough curves are not translated by their first arrival times). On one hand this requires more harmonics to describe the variability of the curves; on the other hand it allows us to demonstrate that the functional error model is able to correct for different arrival times also without registration.

We consider 200 different learning sets, which are selected by DKM clustering with different initialization. For each learning set we apply FPCA and then construct the functional error models by employing a different number of harmonics. The quality of the prediction is measured by the l_2 distance between the predicted and exact responses for all 1000 realizations.

The performance of the method (expressed as median error and confidence interval of the responses of the 200 learning sets) is presented in Fig. 10 as a function of the number of harmonics. The error exhibits a minimum around 5-7 harmonics. Indeed,

when the number of harmonics is increased from 2 to 5, the variability of the learning set represented increases from 92% to 99%, leading to an improved error model. If the number of harmonics is increased further, the error increases quite rapidly. For 12 harmonics errors are very large and fluctuate greatly depending on the choice of the learning set. This behavior is a clear signature of over-fitting, as the large number of harmonics is not balanced by the size of the learning set (consisting of 20 pairs of curves) and the parameters of the regression model are not constrained enough by the data.

5. Conclusions

We have presented a novel methodology that combines elements of Functional Data Analysis and Machine Learning to construct error models that improve uncertainty quantification. The approach is purpose-oriented as it is formulated directly on the quantity of interest (in the case considered here, the contaminant breakthrough curve) rather than on the state of the system (e.g., the entire saturation and pressure fields).

The core idea of the method is to construct an error model from a learning set containing pairs of proxy and exact responses of a subset of realizations, and to predict the exact responses of the entire sample without solving the exact model for all realizations. FPCA is employed to separately reduce the dimensionality of the spaces of exact and proxy responses in the learning set. The advantage is twofold: on one hand, the small dimension allows a diagnostic of the regression model on scores to assess the informativeness of the proxy for the application at hand; on the other hand, using spaces of lower dimension reduces the risk of over-fitting when the regression model is constructed.

The method has been tested for a synthetic contamination problem, in which the breakthrough curve of a NAPL contaminant is predicted with the help of a tracer transport

simulation (as proxy model). We have obtained excellent results with a learning set consisting of 20 pairs of curves (corresponding to 20 realizations out of a sample of 1000) and considering only the first three harmonics, which describe more than 97% of the variability. Visual inspection of the score scatter plots shows that the proxy is indeed potentially very informative of the exact response (this is confirmed by a linear determination coefficient $R^2 = 0.97$). Notice that this is not necessarily an indication of the quality of the predictions as the size of the learning set and the number of harmonics also influence the accuracy of the prediction. For both test case, the error model allows us to solve a two-phase problem only for the 20 realizations, whereas a simple tracer transport problem is solved for all realizations in the sample. The gain in computational efficiency is evident as multiphase transport requires solving the pressure problem at every time step, in contrast to ideal tracer transport, which requires solving the pressure equation only once.

In comparison to the Concurrent model (an existing methodology used to correct proxy responses), we have demonstrated an error reduction by a factor 3 when the functional error model is employed. Also, the error model improves the uncertainty quantification with respect to the estimate obtained solely on the basis of the 20 exact responses in the learning set (this approach corresponds to the DKM, which uses the proxy responses only to cluster the realizations). Beside an increase in accuracy, the methodology presents two advantages over the DKM. First, the error model allows us to use the proxy response to predict the exact response for any new geostatistical realization that might be successively generated; this clearly opens new possibilities to use the model beyond the context of uncertainty quantification, and in particular for Bayesian inference, model calibration and optimization. Second, simultaneous confidence bands of the predicted curves can be

defined by propagating the errors of the multivariate regression model. Notice that the residual uncertainty due to the size of the learning set and to the truncation of the basis should be taken into account.

Combining FPCA and machine learning can be seen as a general framework in which each component can be modified and improved, if it is required to improve accuracy. For instance, the rather crude linear regression model between the three-dimensional spaces of exact and proxy responses can be made more complex by increasing the dimensions (possibly with different truncations for the proxy and the exact model) or by refining the mathematical form of the statistical model to predict the scores. Possible enhancements include linear regression models with more complex basis functions (polynomials or others), but may also entail kernel methods like co-kriging. Almost any multivariate prediction may be adapted to this problem once the dimensionality reduction is performed. Another potential improvement is to perform the dimensionality reduction jointly for the proxy and the exact spaces, in order to optimize the informativeness of the proxy rather than the description of the variability of each response space independently. Indeed, in very complex test cases, it might occur that some small-eigenvalue harmonics of the proxy response might explain large-eigenvalue characteristics of the exact curves. This can be done by replacing FPCA by Functional Canonical Correlation Analysis [*Ramsay*, 2006] or by Functional Partial Least Squares [*Cuevas*, 2014].

Finally, we observe that the proposed framework can be applied far beyond the contamination example that we have presented. It can be useful in virtually any situation in which the most reliable technique has to be surrogated by an approximate method. Applications are not limited to the case in which evaluating exact response involves the

solution of a complex numerical model, but also to situations in which the proxy or the exact responses consist of experimental data. The FDA framework would be then important to compare information with different temporal resolutions. Also, the error model can potentially be very useful in the context of Bayesian inference, when the number of responses that have to be evaluated (e.g., in Metropolis-Hastings algorithms and alike) is typically of the order of 10^5. In this case, a functional error model capable to predict the exact responses only on the basis of the proxy responses can substantially speed up MCMC algorithms, as it reduces the cost of likelihood estimation. This would improve the efficiency of the calibration and optimization algorithms, which are often used in hydrogeological applications.

Appendix A: Multiphase and single-phase transport equations

Assuming that both phases are incompressible and neglecting gravity and capillary effects, the saturation of the NAPL, S, is governed by the following system of equations:

$$\nabla \cdot \left[\left(\frac{k_n(S)}{\mu_n} + \frac{k_w(1-S)}{\mu_w} \right) k \nabla p \right] = 0, \tag{A1}$$

$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(\frac{k_n(S)}{\mu_n} k \nabla p \right) = 0, \tag{A2}$$

where the absolute permeability, k, and the porosity, ϕ, are aquifer properties; p is the pressure; μ_n and μ_w are the viscosities of NAPL and water, respectively; and k_n and k_w are the relative permeabilities of NAPL and water, respectively, which are nonlinear functions of the saturation. Together with the constitutive relationships for the permeabilities (here, they are assumed quadratic i.e., $k_n(S) = S^2$ and $k_w(S) = (1-S)^2$), the two equations above form a complete system of equations that can be solved for p and S to calculate the NAPL breakthrough curves. These curves are the responses of the exact (multiphase) model.

Due to the nonlinearity of the relative permeability, the system above is computationally expensive because the two equations are coupled and the pressure equation has to be solved at any time step. This problem can be avoided by neglecting the nonlinearity of the permeabilities, hence approximating the system above as

$$\nabla \cdot \left(\frac{k}{\mu_w} \nabla p \right) = 0, \tag{A3}$$

$$\frac{\partial}{\partial t}(\phi S) - \nabla \cdot \left(S \frac{k}{\mu_w} \nabla p \right) = 0, \tag{A4}$$

which corresponds to a simple tracer transport problem without mechanical dispersion.

Appendix B: Linear models for functional responses with functional predictors

A simple class of linear models is the Concurrent model [*Ramsay*, 2006]. The value of the response variable $y(t)$ is predicted solely by the value of the functional covariate at the same time t

$$y_i(t) = \alpha(t) + x_i(t)\beta(t) + \varepsilon_i(t), \tag{B1}$$

where $\varepsilon_i(t)$ are the functional errors and the functions $\alpha(t)$ and $\beta(t)$ are estimated by minimizing the sum of squares under some penalty on the roughness of the functions to avoid overfitting and loose predictability power. Despite the rather arbitrary choice of the degree of smoothness of the functional parameters, this method is quite fast but also rudimentary because there is a priori no reason to assume that only concurrent features of the curves are relevant (this is well illustrated by the synthetic test to predict the single-point breakthrough curve in Sec. 4.5.2).

A generalized formulation is when the functional variable contributes to the prediction for all possible time values

$$y_i(t) = \alpha(t) + \int x_i(s)\beta(s,t)ds + \varepsilon_i(t) \tag{B2}$$

which allows the predicted response to depend on the functional covariate at all times, but $\beta(s,t)$ is now bivariate. The application of this model is known to be particularly challenging as the smoothing constraints to be imposed is of paramount importance.

Appendix C: Simultaneous confidence bands for multiple multivariate linear regression

To take into account the uncertainty stemming from the linear regression, we derive simultaneous confidence bands for the predicted curve $\hat{y} = \mathbf{b}'\hat{\beta}\boldsymbol{\eta}(t)$, where $1-\alpha$ is the level of confidence that the exact curve $\tilde{y}(t) = \mathbf{b}'\beta\boldsymbol{\eta}(t)$ is within the confidence bands for all t, that is

$$Pr\Big(\tilde{y}(t) \in [\hat{y}(t) - w_\alpha(t), \hat{y}(t) + w_\alpha(t)] \text{ for all } t\Big) = 1-\alpha \tag{C1}$$

and, following the sketch of proof below, where $D_{ex} + D_{app} < N_l$ is assumed,

$$\begin{aligned} w_\alpha(t) \quad &= \sqrt{\tfrac{D_{ex}(N_l - D_{app} - 1)}{N_l - D_{ex} - D_{app}} F_{D_{ex}, N_l - D_{ex} - D_{app}}(\alpha)} \\ &\times \sqrt{(1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}) \tfrac{N_l}{N_l - D_{app} - 1} \boldsymbol{\eta}'(t)\hat{\boldsymbol{\Sigma}}\boldsymbol{\eta}(\boldsymbol{t})}, \end{aligned} \tag{C2}$$

where $\boldsymbol{\eta}(t)$ the values of the exact harmonics; $F(\alpha)$ Fisher's α-quantile; and $\hat{\boldsymbol{\Sigma}}$ the covariance matrix of the errors estimated on the learning set.

The key step of the derivation is the use of Scheffe's Lemma that states that, for a symmetric and positive definite matrix $\Gamma \in \mathbb{R}^{p\times p}$, the following statements are equivalent for any vector $\mathbf{v} \in \mathbb{R}^p$ and constant $c > 0$

$$\Big(\mathbf{v}'\Gamma\mathbf{v} \leqslant c^2\Big) \iff \Big(|\boldsymbol{\psi}'\mathbf{v}| \leqslant c\sqrt{\boldsymbol{\psi}'\Gamma^{-1}\boldsymbol{\psi}} \qquad \forall \boldsymbol{\psi} \in \mathbb{R}^p\Big) \tag{C3}$$

Sketch of proof

The residuals $\hat{\mathbf{E}} = \hat{\mathbf{C}} - \mathbf{C}$ are centred and with covariance $\mathbb{E}[\hat{\mathbf{E}}'\hat{\mathbf{E}}] = (N_l - D_{app} - 1)\boldsymbol{\Sigma}$, where $(\boldsymbol{\Sigma})_{jk} = \sigma_{jk}$. Assuming that $\mathbf{E}$ is Gaussian entails that $\hat{\boldsymbol{\beta}}$ is Gaussian, whereof $\mathbf{c} \sim \mathcal{N}_{D_{ex}}\Big(\mathbf{b}'\beta, (1+\mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}')\boldsymbol{\Sigma}\Big)$. Then $\Big(\frac{\mathbf{b}'\hat{\beta}-\mathbf{b}'\beta}{\sqrt{1+\mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}}\Big)'\Big(\frac{1}{N_l - D_{app} - 1}\boldsymbol{\Sigma}\Big)^{-1}\Big(\frac{\mathbf{b}'\hat{\beta}-\mathbf{b}'\beta}{\sqrt{1+\mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}}\Big)'$ follows a Chi-squared distribution $\chi^2_{D_{ex}}$. On the other hand, the usual estimator $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$ follows a Wishart distribution independently from $\hat{\beta}$. We then obtain the following

$$t^2 = \left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}}\right)'\Big(\frac{N_l}{N_l - D_{app} - 1}\hat{\boldsymbol{\Sigma}}\Big)^{-1}\left(\frac{\mathbf{b}'\hat{\beta} - \mathbf{b}'\beta}{\sqrt{1 + \mathbf{b}'(\mathbf{B}'\mathbf{B})^{-1}\mathbf{b}}}\right) \sim T^2_{D_{ex}, N_l - D_{app} - 1}. \tag{C4}$$

As the Hotelling T^2-distribution can be expressed in term of the F-distribution, we can write that, with probability $1-\alpha$,

$$t^2 \leqslant \frac{D_{ex}(N_l - D_{app} - 1)}{N_l - D_{ex} - D_{app}} F_{D_{ex}, N_l - D_{ex} - D_{app}}(\alpha), \tag{C5}$$

where $F_{p,q}(\alpha)$ stands for the α-quantile of the Fisher-Snedecor distribution with parameters p and q.

Using Scheffe's Lemma (eq. C3) for $\mathbf{v} = \mathbf{b}'\hat{\boldsymbol{\beta}}$ the vector of predicted scores and $\boldsymbol{\psi}$ the vector of the exact harmonics values $\boldsymbol{\eta}(t)$, the second statement gives us the simultaneous confidence bands on the prediction.

Acknowledgments.

The data to support this article result from numerical simulations (multiphase and tracer transport) performed with the open source code MaFlot (*Matlab Flow and Transport* [*Künze and Lunati*, 2012, 2013]). The data treatment is performed with the *fda* package [*Ramsay et al.*, 2012] implemented in R [*R Core Team*, 2013]. Upon request by email, the authors would provide the simulated data and codes.

The authors thank Rouven Künze for his assistance with the flow simulations, Guillaume Pirot and Philippe Renard for providing the realizations of the hydraulic conductivity, and Céline Scheidt for sharing her DKM code. Many thanks are due to L. Dmbgen for his teachings, to V. Demyanov and A.H. Elsheikh for many useful discussions and to the reviewers for their suggestions and careful editing.

This project is supported by the Swiss National Science Foundation as a part of the ENSEMBLE project (Sinergia Grant No. CRSI22-132249/1). David Ginsbourger acknowledges support from the Institute of Mathematical Statistics and Actuarial Science,

University of Bern. Ivan Lunati is Swiss National Science Foundation (SNSF) Professor at the University of Lausanne (SNSF grant numbers PP00P2-123419/1 and PP00P2-144922/1).

References

Ballin, P., A. Journel, and K. Aziz (1992), Prediction of uncertainty in reservoir performance forecast, *Journal of Canadian Petroleum Technology*, *31*(4).

Bayer, P., P. Huggenberger, P. Renard, and A. Comunian (2011), Three-dimensional high resolution fluvio-glacial aquifer analog: Part 1: Field study, *Journal of Hydrology*, *405*(1), 1–9.

Christie, M. (1996), Upscaling for reservoir simulation, *Journal of Petroleum Technology*, *48*(11), 1004–1010.

Christie, M. A., J. Glimm, J. W. Grove, D. M. Higdon, D. H. Sharp, and M. M. Wood-Schultz (2005), Error analysis and simulations of complex phenomena, *Los Alamos Science*, *29*(6).

Cuevas, A., Febrero, M., and Fraiman, R. (2002). Linear functional regression: the case of fixed design and functional response, *Canadian Journal of Statistics*, 30(2), 285-300.

Cuevas, A. (2014), A partial overview of the theory of statistics with functional data, *Journal of Statistical Planning and Inference*, *147*, pp. 1–23.

Dagan, G. (2002), An overview of stochastic modeling of groundwater flow and transport: From theory to applications, *Eos, Transactions American Geophysical Union*, *83*(53), 621.

Durlofsky, L. (2005), Upscaling and gridding of fine scale geological models for flow simulation, in *8th International Forum on Reservoir Simulation Iles Borromees, Stresa,*

Italy, pp. 20–24.

Fox, J., and Weisberg, S. (2011). Multivariate Linear Models in R.

Hastie, T., R. Tibshirani, and J. Friedman (2009), *The Elements of Statistical Learning*, Springer

Helmig R. (1997) *Multiphase Flow and Transport Processes in the Subsurface*, Springer Verlag, Berlin-Heidelberg

Henderson, B. (2006), Exploring between site differences in water quality trends: a functional data analysis approach, *Environmetrics*, *17*(1), 65–80.

Josset, L., and I. Lunati (2013), Local and global error models to improve uncertainty quantification, *Mathematical Geosciences*, pp. 1–20.

Kaiser, H. (1958), The varimax criterion for analytic rotation in factor analysis, *Psychometrika*, *23*(3), 187–200, doi:10.1007/BF02289233.

Künze, R., and I. Lunati (2013), Adaptive multiscale simulations of density-driven instabilities, *Journal of Computational Physics*, 255, 502–520, 10.1016/j.jcp.2012.02.025.

Künze, R., and I. Lunati (2012), MaFloT - Matlab Flow and Transport. Published under the GNU license agreement on `www.maflot.com`

Mariethoz, G., P. Renard, and J. Straubhaar (2010), The direct sampling method to perform multiple-point geostatistical simulations, *Water Resources Research*, *46*(11), W11,536.

Marle, C.M. (1981), *Multiphase flow in porous media*, Institut Francais du Petrole, Paris, France.

McLennan, J., and C. Deutsch (2005), Ranking geostatistical realizations by measures of connectivity, in *SPE International Thermal Operations and Heavy Oil Symposium,*

98168, Alberta, Canada.

O'Sullivan, A., and M. Christie (2005), Error models for reducing history match bias, *Computational Geosciences*, *9*(2-3), 125–153.

O'Sullivan, A., and M. Christie (2006), Simulation error models for improved reservoir prediction, *Reliability Engineering & System Safety*, *91*(10), 1382–1389.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Ramsay, J. J. O., G. Hooker, and S. Graves (2009), *Functional data analysis with R and MATLAB*, Springer.

Ramsay, J. O. (2006), *Functional data analysis*, Wiley Online Library.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2012), *fda: Functional Data Analysis*, r package version 2.3.2.

Renard, P., and G. de Marsiliy (1997), Calculating equivalent permeability: a review, *Adv. Water Res.*, *20*(5-6), 253–278.

Richman, M. B. (1986), Rotation of principal components, *Journal of climatology*, *6*(3), 293–335.

Scheidt, C., and J. Caers (2009a), Representing spatial uncertainty using distances and kernels, *Mathematical Geosciences*, *41*(4), 397–419.

Scheidt, C., and J. Caers (2009b), Uncertainty quantification in reservoir performance using distances and kernel methods–application to a west africa deepwater turbidite reservoir, *SPE Journal*, *14*(4), 680–692.

Ullah, S., C. F. Finch, et al. (2013), Applications of functional data analysis: A systematic review, *BMC medical research methodology*, *13*(1), 43.

Wen, X., and J. Gómez-Hernández (1996), Upscaling hydraulic conductivities in heterogeneous media: An overview, *J. Hydr.*, *183*(1-2), R9–R32.

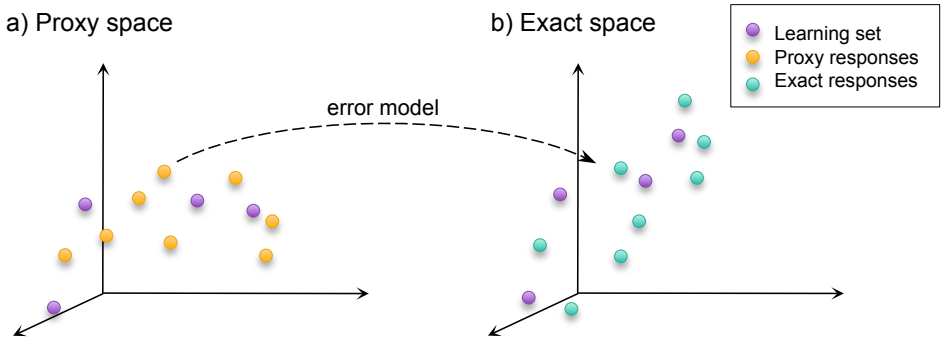


Figure 1. A statistical model is built on the learning set to relate the coefficients of the elements $x_i(t)$ in the proxy-response space to the coefficients of the elements $y_i(t)$ in the exact-response space. It is used as error model to predict the exact response from the proxy response.

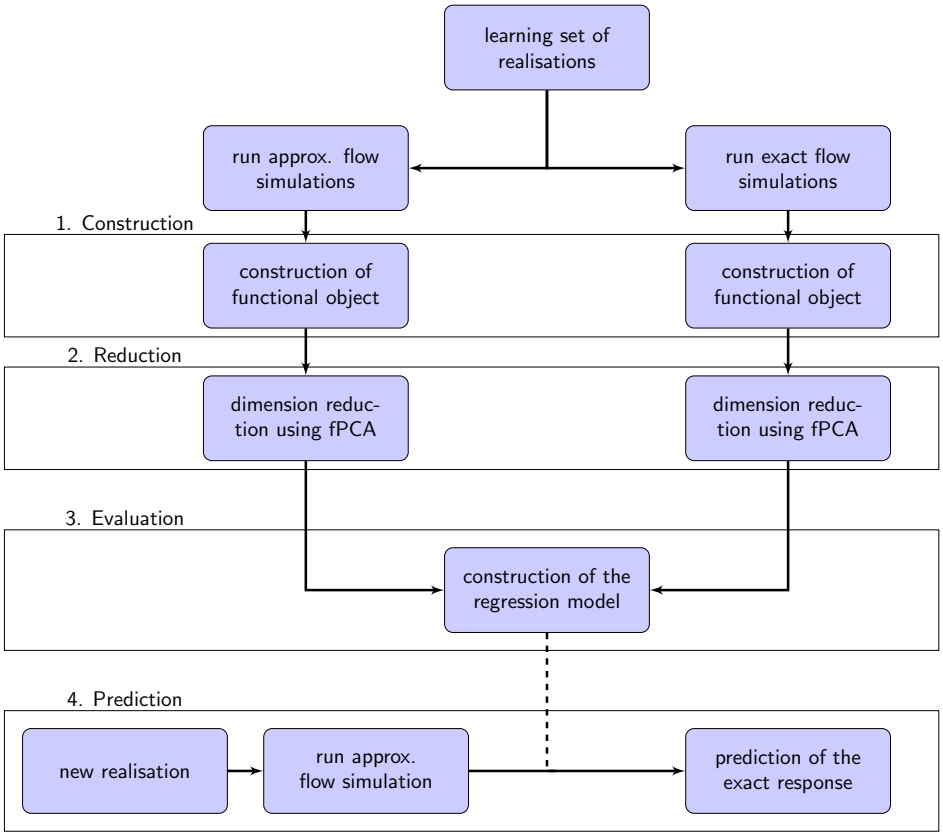
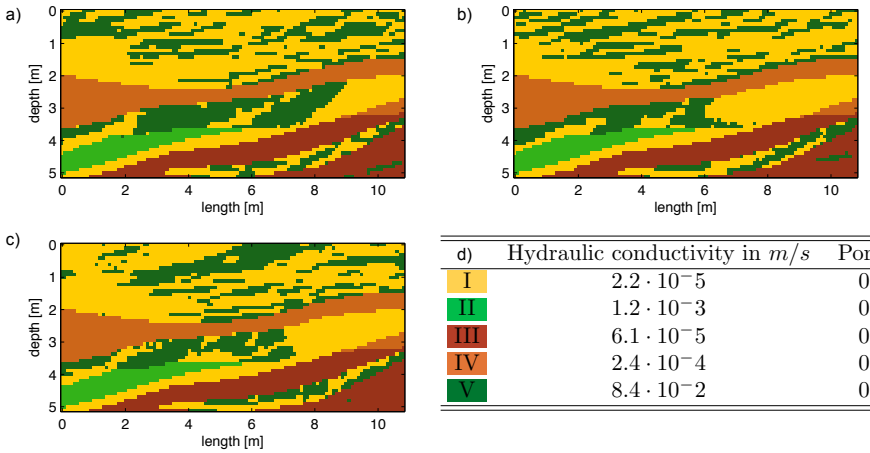


Figure 2. Flowchart of the methodology. After a learning set of realization has been constructed by selecting a subset of realizations and calculating pairs of proxy- and exact-response curves, the exact responses for the realizations that are not in the learning set can be predicted in four steps: 1. first, the functional objects are constructed by spline interpolation, 2. then, the dimensions of the subspaces of exact and proxy responses are reduced by means of FPCA, 3. next, a regression model is constructed between the proxy and the exact scores; 4. finally, the regression model is used to predict the exact responses of the realizations that are not in the learning set.



d)	Hydraulic conductivity in m/s	Porosity
I	$2.2 \cdot 10^{-}5$	0.21
II	$1.2 \cdot 10^{-}3$	0.32
III	$6.1 \cdot 10^{-}5$	0.13
IV	$2.4 \cdot 10^{-}4$	0.16
V	$8.4 \cdot 10^{-}2$	0.25

Figure 3. (a), (b) and (c): three examples of geostatistical realizations generated by a multipoint methods (DeeSSe, [*Mariethoz et al.*, 2010]) with training image from the Herten site (Germany) [*Bayer et al.*, 2011]. The different colors correspond to 5 different facies, whose properties are reported in (d). The three realizations belong to the set of realizations used to construct the learning set; the corresponding NAPL breakthrough curves obtained with the exact and with the approximate models are highlighted in Fig. 4.

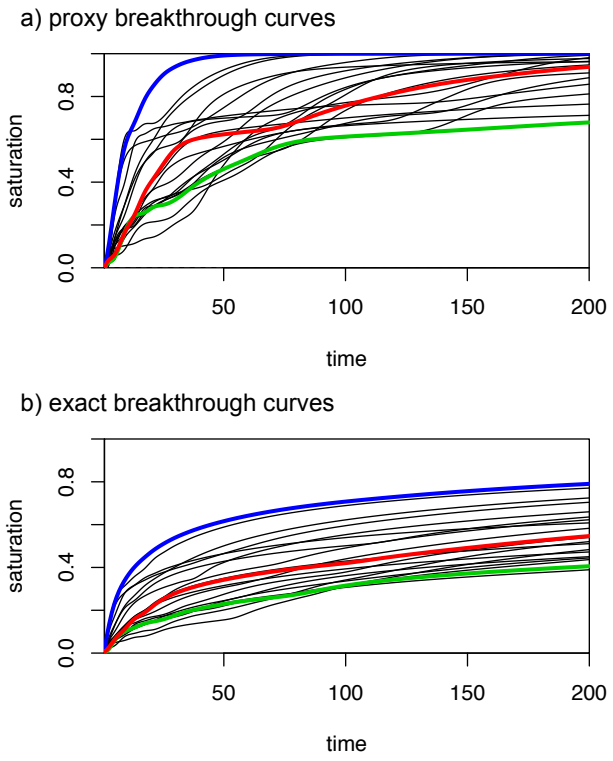


Figure 4. The learning set: (a) proxy curves and (b) exact curves recast as functional objects for the $N_l = 20$ realizations in the learning set. The thicker blue curves correspond to the realization in Fig. 3a), the red curves to 3b), and the green curves to 3c).

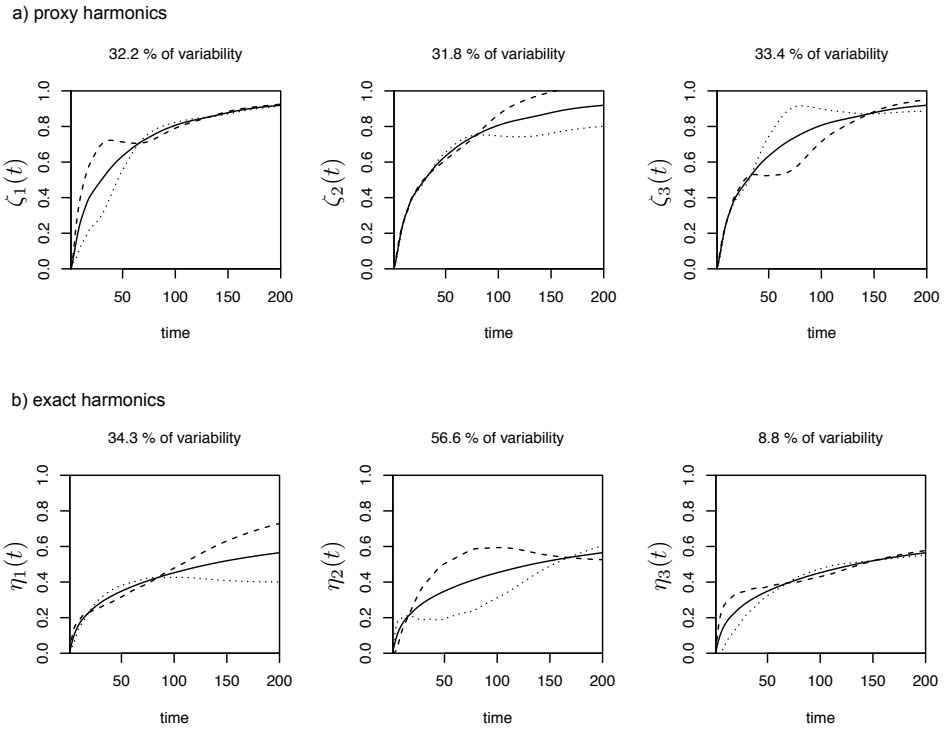


Figure 5. The three first rotated functional principal components (harmonics) extracted from the learning set are plotted for the proxy curves (top) and for the exact curves (bottom). The solid line is the mean curve and the dotted lines represent the variability around the mean described by the corresponding harmonic.

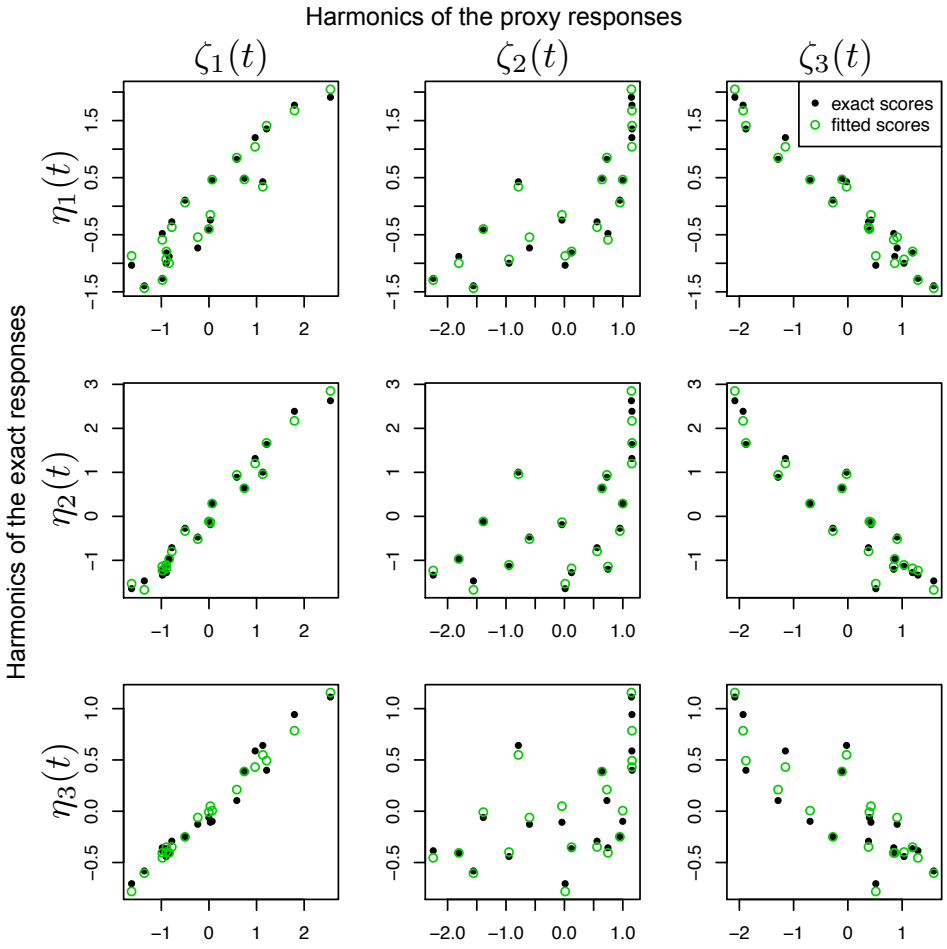


Figure 6. The scores, with respect to the first three harmonics $\{\eta_i(t)\}_{i=1,2,3}$, of the exact curves are plotted as functions of the scores for the approximate curves with respect to the harmonics $\{\zeta_i(t)\}_{i=1,2,3}$. The filled (black) circles correspond to the exact score, the empty circles (green) to the prediction of the scores by the OSL linear regression. The visualization is helpful to assess whether the linear regression model describes the relationship between proxy and exact curves in the learning set.

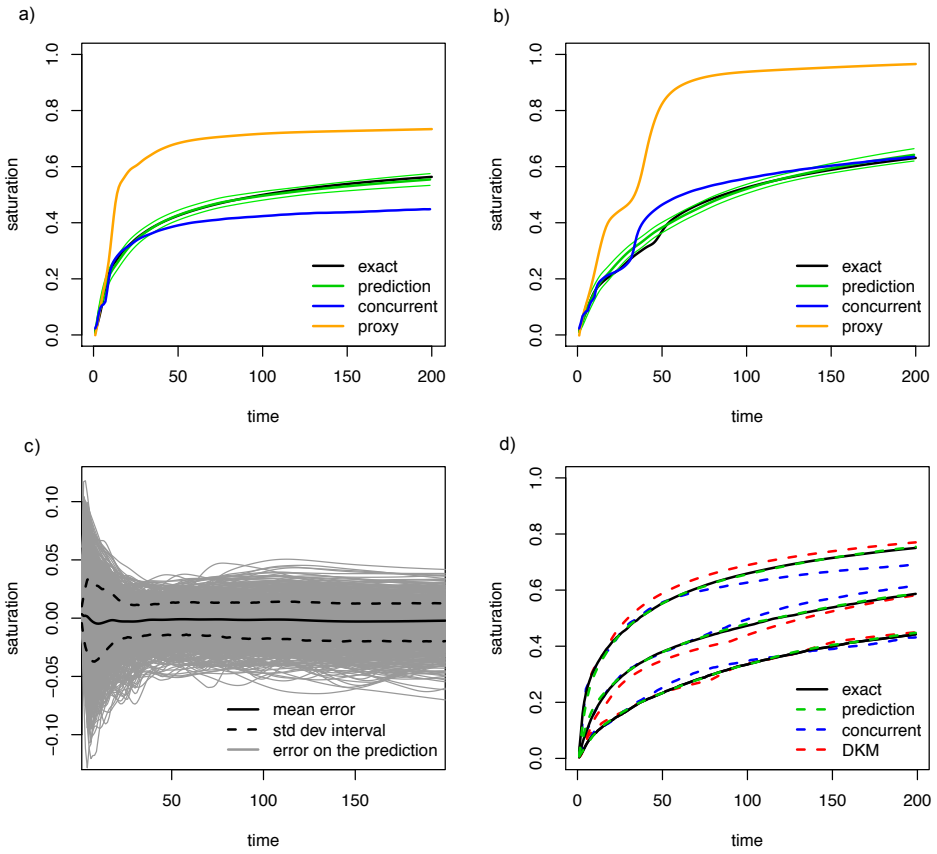


Figure 7. (a and b) the predicted responses (with 2σ-confidence intervals) of two realizations that are not in the learning set. (c) Prediction error of all $N_r = 1000$ realizations (gray curves), the mean error (continuous line), and the mean $\pm$ one standard deviation (dotted lines) are represented. (d) P10, P50 and P90 quantiles curves obtained with the different models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line).

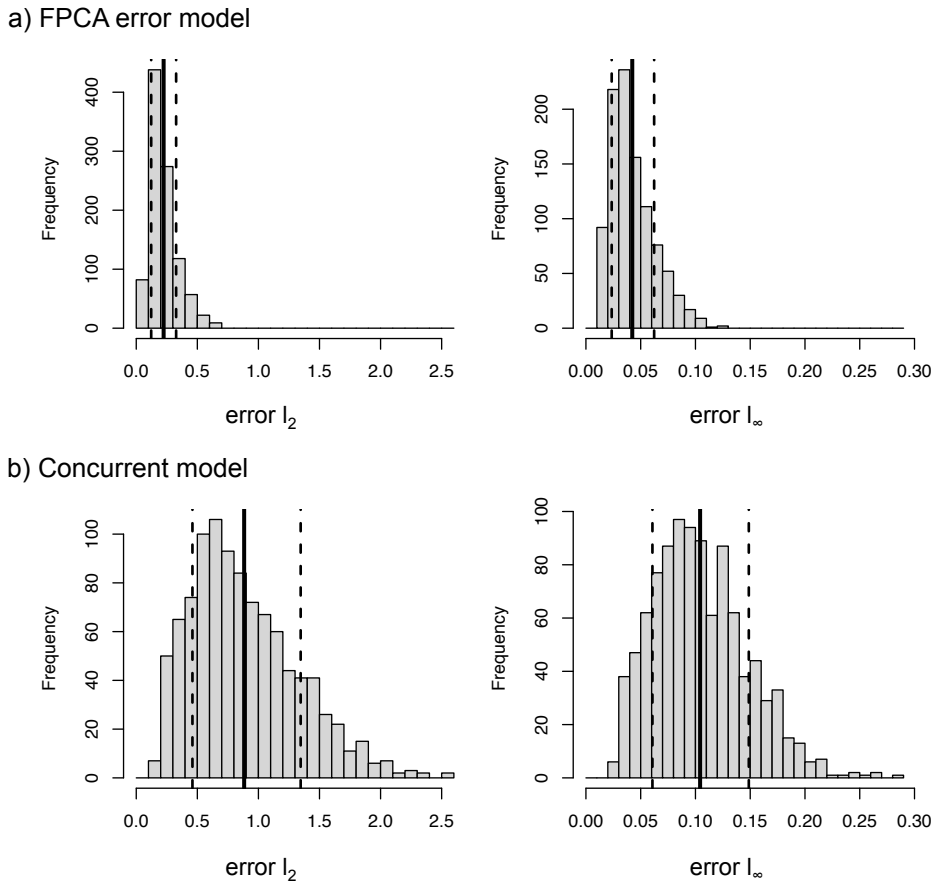


Figure 8. Histograms of the distribution of the l_2 error (left) and l_∞ error (right), (a) for the predictions of the FPCA model and (b) for the predictions of the concurrent model. The mean (continuous line) together with the mean $\pm$ one standard deviation (dotted lines) are represented.

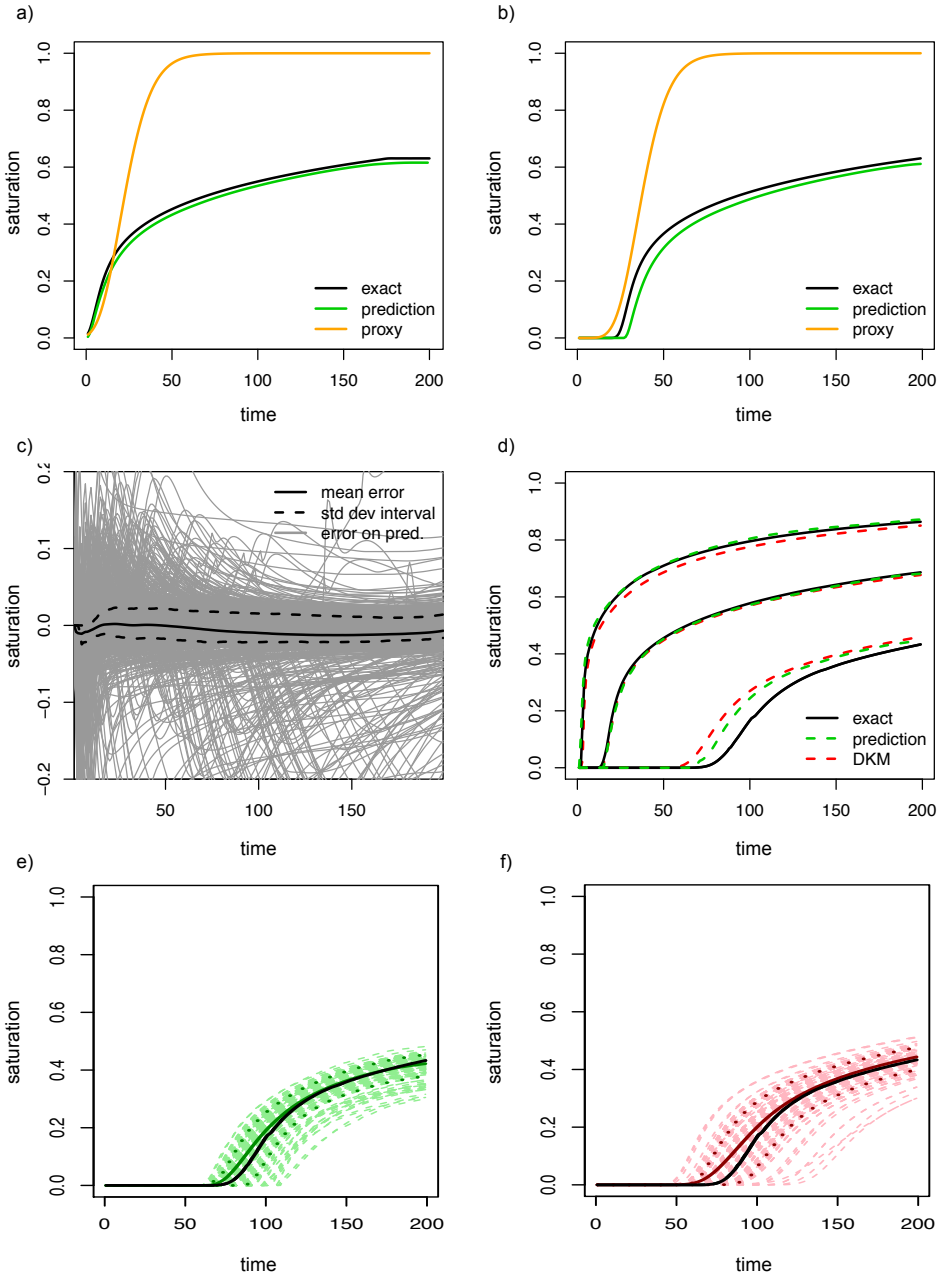


Figure 9. (a and b) predicted responses (before and after translation) of a realization that is not in the learning set. (c) Prediction error of all $N_r = 1000$ realizations (grey curves), the mean error (continuous line), and the mean $\pm$ one standard deviation (dotted lines) are represented. (d) P10, P50 and P90 quantiles curves obtained with the different models and compared to the reference quantile curves computed using the whole set of exact responses (solid black line). (e), respectively (f), shows the P10 FPCA, respectively DKM, predictions of the P10 quantile for the 200 clusterings.

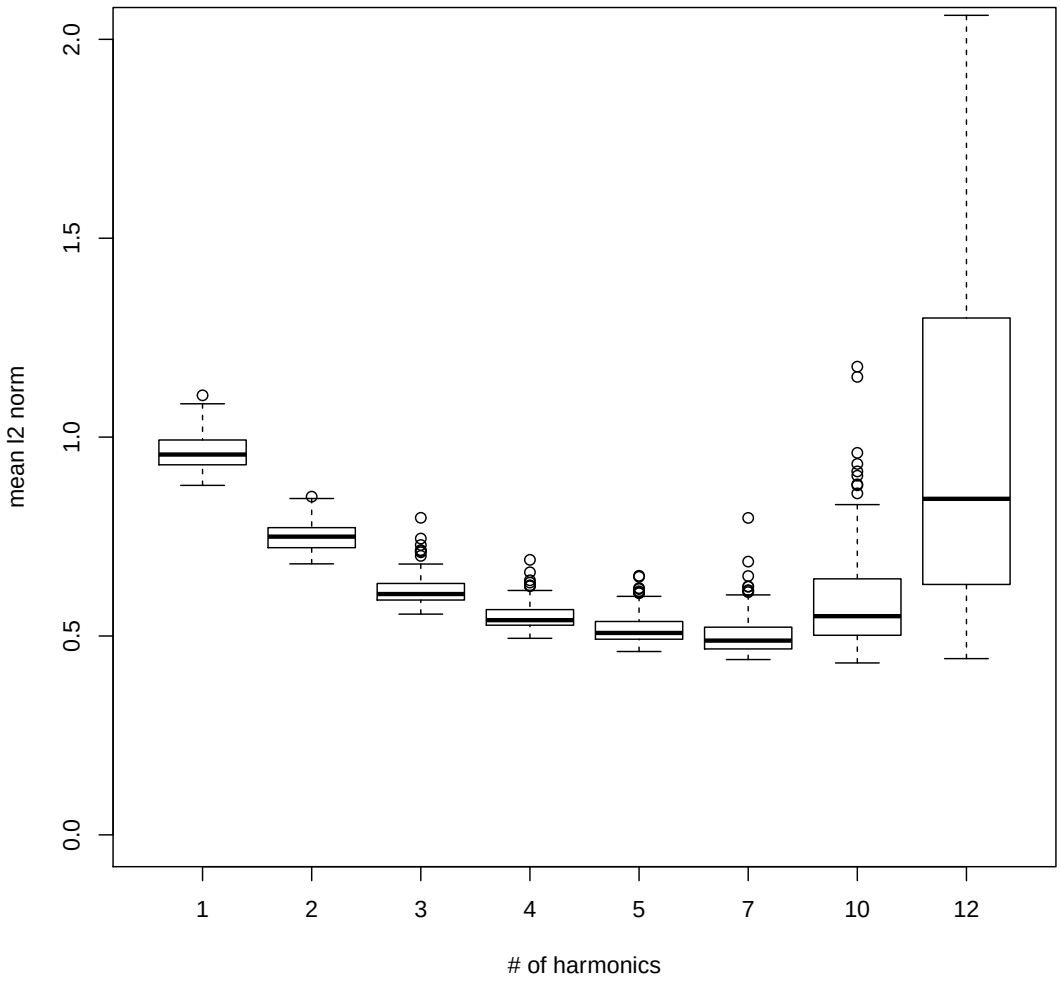


Figure 10. Boxplots of the prediction error (calculated as mean l_2 norm of the error of the predicted curves) as a function of the number of harmonics used to describe the proxy and exact curves in the learning set. The boxplots represent the statistics of the prediction errors over 200 clusterings in function of the number of harmonics. The thick line indicates the median error; the box the 1σ interval; the bars the 2σ interval; and the circles are the outliers (for 12 harmonics they are out of scale).