Math Geosci (2015) 47:173–191
DOI 10.1007/s11004-014-9521-6

Prediction-Focused Subsurface Modeling: Investigating the Need for Accuracy in Flow-Based Inverse Modeling

Céline Scheidt · Philippe Renard · Jef Caers

Received: 12 October 2012 / Accepted: 10 January 2014 / Published online: 21 February 2014

Abstract The objective of most formulations of inverse modeling in the Earth Sciences is to estimate the model parameters given the observation data. In this paper, an additional element to these formulations is considered, namely, the prediction for which the models are built. An example of such modeling is the prediction of solute transport using geological models of the subsurface constrained to existing geophysical, flow dynamic and solute observations. The paper then illustrates and addresses a fundamental question relating data, model and prediction: does model inversion reduce uncertainty in prediction variables given the observed data? To investigate this question, a diagnostic tool is proposed to assess whether matching the observed data significantly reduces uncertainty in the prediction variables. In addition, for some cases, a quick estimate of uncertainty can be obtained without applying inverse modeling. It relies on a dimensionality reduction method using non-linear principal component analysis (NLPCA) calibrated from evaluating the prediction and data response variables on a few prior Earth models. The proposed diagnostic tool is applied on a simple example of tracer flow and, for the cases investigated, the NLPCA provided an accurate diagnostic and a satisfactory pre-estimation of uncertainty in the prediction variables.

Keywords Uncertainty quantification · Inverse problems · Prior models · Subsurface geological modeling · Groundwater · Relationship data prediction

C. Scheidt (✉) · J. Caers
Department of Energy Resources Engineering, Stanford University, Stanford, CA 94305, USA
e-mail: scheidtc@stanford.edu

P. Renard
Université de Neuchâtel, Neuchâtel, Switzerland

1 Introduction

Most problems in the Earth sciences involving data and a model constitute an inverse problem: an Earth model needs to be inferred from data. The importance of calibrating the Earth model to data has led to inverse modeling as a very active domain, and explains why many review papers have been published on this topic (Mosegaard and Tarantola 2002; Carrera et al. 2005; Oliver and Chen 2011; Zhou et al. 2013). However, in addition to the field data and model, a third element is often of equal importance, namely, the quantification of uncertainty in some prediction (typically a future response) evaluated on the inverse model solutions. A typical example is the prediction of reservoir production performance based on historical production data and geological information; the data being the historical production, the model being a three-dimensional gridded Earth model constrained to well and seismic data. Uncertainty quantification is desired on some future response such as future oil and water production of existing wells, four-dimensional saturation changes in the field, or production rates of planned wells. Similar problems exist in groundwater management where piezometric records and water chemistry data are used in an inverse modeling framework to constrain a numerical model, which is then used to estimate uncertainty and provide management decisions (Freeze et al. 1990; Gallagher and Doherty 2007; Alcolea et al. 2009). What is common in many of these practical problems is that the final desired product is not always the inverted parameters or Earth model itself, but the prediction made on such an Earth model as well as some uncertainty quantification for risk assessment. The issue when using inverse methods followed by an uncertainty analysis on the forecast is that most current inverse methods require considerable modeling and computational effort. Thus, a relatively quick estimation of the relationship between the data and the prediction could be very useful. Note that this idea is different from the use of a surrogate model to accelerate uncertainty analysis: in these approaches (Mugunthan and Shoemaker 2006; Ferraille and Marel 2012), the model forecast is approximated via the surrogate model, but not the relationship between the field data and model forecasts.

In broad terms, this paper investigates the relations between inverse modeling and prediction variables. Some observed data are assumed to be available, and the objective is to assess uncertainty in a prediction variable. Both data and prediction are computed by forward simulations from Earth models (and are denoted as simulated data variables and prediction variables, respectively). Note that in this paper, the study is limited to flow variables; other inverse problems are not investigated. Characteristic of flow inverse problems is that the dynamic data ($\mathbf{d}$) may not constrain the permeability model ($\mathbf{m}$) well, because of the ill-posedness of the problem. The primary objective is to elaborate an approach to asses when model inversion does not significantly reduce uncertainty in the prediction variables given the observed data. Depending on the location where the data have been acquired and depending on the value of the observed data themselves, it may occur that those data are not constraining the model in such a way that the prediction is affected. To give an extreme example, assume that one is interested in the protection of a groundwater well and in evaluating travel times between this well and a waste disposal upstream. It is quite intuitive to think that the estimation of this travel time will not be influenced significantly by a

water level measurement far downstream of the groundwater well. In most practical cases, the situation will not be so clear and obvious, in particular if the relationship between simulated data and prediction variables is highly non-linear, and the data are only partially informative about such prediction. This suggests that there is a need to define some tools to diagnose if performing a time-consuming inverse modeling is relevant or not for a given specific practical prediction. In that perspective, the paper introduces a method allowing a fairly rapid analysis of the situation. In addition, in certain cases, one can use the proposed framework to obtain a reasonable estimation of the uncertainty in prediction without having to run the full traditional inversion approach. The approach introduced in this paper is termed prediction-focused analysis (PFA). PFA does not require iteration, but rather relies on estimating a direct relationship between the simulated data variables and prediction variables. This analysis requires the creation of prior Earth models to calibrate that relationship, but these Earth models need not match any data. PFA is illustrated on a simple hydrogeological example.

2 Problem Setup

In this section, the notations and problem setup are introduced. A Bayesian framework is used to assess the impact of inverse modeling on the uncertainty in some prediction variable: if the prior and posterior distributions of the prediction variable are similar, then uncertainty is not reduced and applying inverse modeling may not be useful. The notations of Mosegaard and Tarantola (1995) are used for the inverse problem formulation, including an additional variable: the forecast/prediction. The final aim is to get a reasonable estimation of the posterior distribution of the prediction variable. To start, $\mathbf{m}$ is the vector representing one Earth model, it is usually a gridded three-dimensional Earth model obtained by geostatistical or other simulation techniques. The prior distribution of $\mathbf{m}$ is denoted $\rho(\mathbf{m})$. $\mathbf{d}_{\text{obs}}$ represents measurements of state variables in the field, or observed field data. The data are assumed to be a flow-based (time-series) response (pressure, concentration, fluxes, etc.). The Earth model and the state variables are related via a forward model that represents the physics of the problem. This model allows computing the state variables at the location of the measurements for a given value of $\mathbf{m}$. The values of the computed state variables are denoted $\mathbf{d}$ and the forward model is denoted $\mathbf{g}(\mathbf{m})$. Therefore, $\mathbf{d} = \mathbf{g}(\mathbf{m})$. The computation of $\mathbf{g}(\mathbf{m})$ involves solving numerically a set of partial differential equations and is therefore highly CPU demanding. The solution of the inverse problem is the posterior distribution $\sigma(\mathbf{m})$ of the Earth models $\mathbf{m}$ knowing $\mathbf{d}_{\text{obs}}$. Formally, the posterior distribution can be expressed as $\sigma(\mathbf{m}) = \text{k}\, \rho(\mathbf{m})\, \text{L}(\mathbf{m})$, where k is a proportionality constant, and $\text{L}(\mathbf{m})$ is the likelihood function which is itself a function of the misfit O between the observed $\mathbf{d}_{\text{obs}}$ and calculated $\mathbf{d}$ state variables. The likelihood is usually expressed as an exponential model of the mismatch O between the observed $\mathbf{d}_{\text{obs}}$ and the simulated data variables $\mathbf{d}$

$$L(\mathbf{m}) = k\exp\left(-\frac{1}{2}O(\mathbf{m})\right). \tag{1}$$

The resolution of the inverse problem consists then not in giving a formal expression of $\sigma(\mathbf{m})$ because $\sigma(\mathbf{m})$ is too difficult to express analytically, but in generating a set of samples of $\mathbf{m}$ that are distributed according to $\sigma(\mathbf{m})$. There are various ways to generate such samples (e.g., Mosegaard and Tarantola 2002). Here, rejection sampling is used because of its simplicity. More advanced methods could be used as well but even if they are more efficient numerically, they are not really needed here because of the simplicity of the example (flow problem, few wells, etc.). The traditional rejection sampling algorithm (Ripley 1987) to generate a sample $\mathbf{m}$ of a distribution $f(\mathbf{m})$ is the following:

1. Sample $\mathbf{m}$ from a uniform distribution.
2. Sample u from a uniform distribution over [0,1].
3. Accept $\mathbf{m}$ as a sample of $f(\mathbf{m})$ if $u \leq \frac{f(\mathbf{m})}{S}$, where S is the supremum of $f(\mathbf{m})$.

However, when sampling from the posterior distribution $\sigma(\mathbf{m}) = \mathrm{k}\rho(\mathbf{m})\,\mathrm{L}(\mathbf{m})$ in a Bayesian context, models $\mathbf{m}$ from the prior $\rho(\mathbf{m})$ must be generated, and not from a uniform distribution. As a consequence, rejection sampling is applied as follows in the examples:

1. Sample $\mathbf{m}$ from its prior distribution $\rho(\mathbf{m})$.
2. Sample u from a uniform distribution over [0,1].
3. Accept $\mathbf{m}$ as a sample of $\sigma(\mathbf{m})$ if $u \leq \frac{L(\mathbf{m})}{S_L}$, where S_L is now the supremum of the likelihood function, usually $\mathrm{S}_L = 1$.

The second component of the problem is the forecast $\mathbf{h}$. It is a vector that can be computed from $\mathbf{m}$. It represents the variable of interest in a perspective of decision making. For example, it can be just a single value representing the travel time between a drinking water well and a contamination source or a complete oil recovery curve. The dimension of $\mathbf{h}$ depends on the problem but is usually much smaller than the one of $\mathbf{m}$. The computation of $\mathbf{h}$ requires running a forward simulator which is rather heavy in terms of CPU. The computation of $\mathbf{h}$ can also include a rather complex cost analysis. This overall prediction-forward model is denoted $\mathbf{r}(\mathbf{m})$ and $\mathbf{h} = \mathbf{r}(\mathbf{m})$. As before, the prior and posterior distributions of $\mathbf{h}$ cannot be expressed by closed form formulas; instead, the aim is to draw samples for uncertainty quantification on the forecast. To determine the prior $\rho(\mathbf{h})$, models $\{\mathbf{m}_1, \ldots, \mathbf{m}_l\}$ from the prior distribution $\rho(\mathbf{m})$ should be generated and then evaluated: $\mathbf{h} = \mathbf{r}(\mathbf{m})$. The resulting ensemble $\{\mathbf{h}_1, \ldots, \mathbf{h}_l\}$ is distributed according to $\rho(\mathbf{h}) = \rho(\mathbf{r}(\mathbf{m}))$. The classical method to determine the posterior $\sigma(\mathbf{h})$ is first to solve the inverse problem to obtain a set of samples $\{\mathbf{m}_1, \ldots, \mathbf{m}_l\}$ from the posterior distribution $\sigma(\mathbf{m})$. Then $\mathbf{h} = \mathbf{r}(\mathbf{m})$ is computed for each of these samples to get the resulting ensemble $\{\mathbf{h}_1, \ldots, \mathbf{h}_l\}$ which is distributed according to $\sigma(\mathbf{h}) = \sigma(\mathrm{r}(\mathbf{m}))$. All the statistics of interest (P10–P50–P90 in this paper) can be derived from samples of $\rho(\mathbf{h})$ and $\sigma(\mathbf{h})$ and can be compared for decision making.

Here the aim is to propose an analysis tool to check if it is necessary to apply the classical method. A second objective of the paper is, in certain situations, to obtain more rapidly a set of samples $\{\mathbf{h}_1, \ldots, \mathbf{h}_l\}$ that are approximately distributed according to $\sigma(\mathbf{h})$. The most important idea in the proposed approach is to introduce a technique to reduce the dimension of the problem and render possible a direct analysis of the relation

between the data **d** and the forecast **h**. Two new variables are therefore introduced: $\mathbf{d}^*$ is the representation of **d** in a low-dimensional space, $\dim(\mathbf{d}^*) \ll \dim(\mathbf{d})$, and similarly, $\mathbf{h}^*$ is the representation of **h** in a low dimensional space, $\dim(\mathbf{h}^*) \ll \dim(\mathbf{h})$. The computation of $\mathbf{d}^*$ and $\mathbf{h}^*$ will be described in Sect. 4. Note that compared to other approaches, the reduction of dimension is not applied on the Earth model **m**.

3 Illustrative Example

The example is constructed to illustrate the questions raised in the introduction. It is derived from an aquifer analog located in Germany (Herten site, Bayer et al. 2011; Comunian et al. 2011). To create the prior, a multiple-point geostatistical (MPS) approach is used, with the training image (TI) in Fig. 1 (top), showing the binary spatial distribution of a depositional feature of high hydraulic conductivity. Prior Earth models **m** are generated using the MPS code IMPALA (Straubhaar et al. 2011). A tracer is injected on the left side of each Earth model. The observed data $\mathbf{d}_{\text{obs}}$ consist of measurements of that tracer at three different depths (Fig. 1, bottom) for a period of 3.5 days. The purpose is to predict the concentration (vector **h** in our notation) of that tracer later in time (12 days) at a drinking well, located further downstream (right side of the model). Evaluation of the models is done using a flow and transport multi-scale finite volume model (Künze and Lunati 2011, 2012).

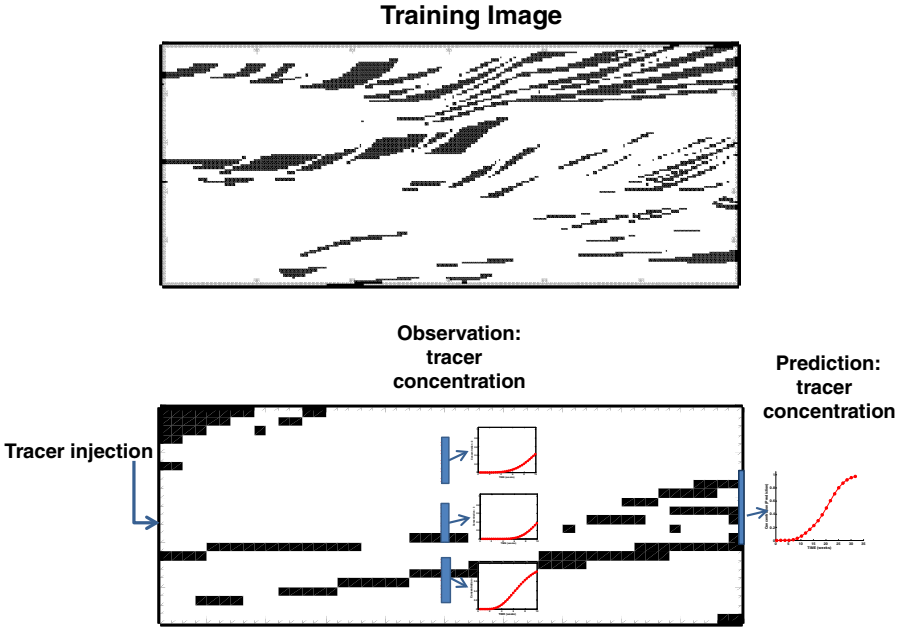

Fig. 1 Test case set-up. Cross-sectional training image (*top*), example of one realization (*bottom*). The grid is a two-dimensional model with 100×25 grid cells. Tracer is injected on the *left side*; the concentration is observed at 3 depths in the *center*, and predicted on the *right*

Three different (synthetic) observed data sets, $\mathbf{d}_{\text{obs},1}$, $\mathbf{d}_{\text{obs},2}$, $\mathbf{d}_{\text{obs},3}$ are used to analyze different situations. They have been chosen purposely to exhibit different possible situations of contaminant arrival times, namely, early, middle and late. The three various $\mathbf{d}_{\text{obs}}$ have been obtained by generating a different Earth model from the prior (same TI and hydraulic conductivities) and by running the forward simulator to obtain the tracer concentration at the 3 depths. In this sense, the prior is consistent with the data and guaranteed to cover the data. In a real case, one would evidently have only one of these data sets available. Note that noise-free observed data are used in this study for simplicity.

For each of the three cases, the reference posterior distribution of $\mathbf{h}$ is obtained by applying rejection sampling in the Bayesian framework outlined in Sect. 2. Prior models are created without using any data and a full posterior probability model of the Earth model is formulated based on the likelihood distribution and the prior.

The objective function or mismatch O in the likelihood (Eq. 1) is defined as a weighted squared distance of the difference between the observed and simulated data variables at each well and for each time steps

$$O(\mathbf{m}) = \frac{1}{N_{\text{wells}} \times N_{\text{times}}} \sum_{w=1}^{N_{\text{wells}}} \sum_{i=1}^{N_{\text{times}}} \left(\frac{d_{\text{obs}}^{i,w} - g_w^i(\mathbf{m})}{\min(0.01, \alpha d_{\text{obs}}^{i,w})} \right)^2, \tag{2}$$

where N_{wells} and N_{times} represent the number of wells and time steps in the simulations and α a tolerance around $\mathbf{d}_{\text{obs}}$. The posterior models are thus accepted based on a tolerance in the average misfit over the wells and time-steps. The reference uncertainty is subsequently obtained by estimating the P10, P50 and P90 quantiles of the tracer concentration at the prediction well from the posterior models. Again, note that rejection sampling would not be feasible in normal circumstances because of its high CPU demand.

Data Case 1 corresponds to a late arrival of the tracer at the three observation wells. Rejection sampling is applied to obtain a set of 30 models matching the data (Fig. 2), within the given tolerance defined by the likelihood (Eqs. 1, 2). A total of 14,820 model evaluations were required. In Fig. 2 (top row and bottom left), the red dots correspond to the observed tracer concentration at the three observation wells, the blue lines to the data variables simulated from the matched models. The bottom-right plot shows prediction uncertainty (P10, P50, P90) corresponding to 200 prior models (dashed black line) compared with uncertainty for the 30 posterior (matched) models (solid blue line). A late tracer arrival at the three observation wells and, not surprisingly, at the prediction well is observed. The uncertainty in the predicted tracer concentration is significantly reduced compared to the prior uncertainty.

Data Case 2 corresponds to a late tracer arrival for the observation wells 1 and 2, whereas an early arrival is obtained for well 3 (see Fig. 3). Rejection sampling was applied to obtain a total of 30 models matching the data (requiring 19,462 flow simulations). Uncertainty in the prediction is not significantly reduced compared to the prior uncertainty (bottom right).

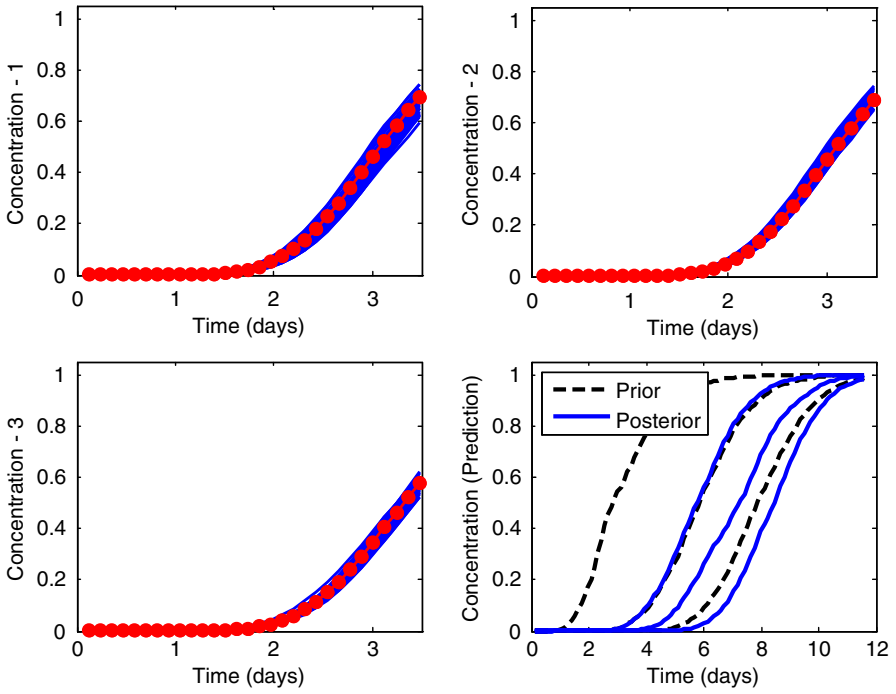

Fig. 2 Rejection sampling results for case 1. Tracer concentration at the three observation wells (*top and bottom-left*); uncertainty (P10, P50 and P90) at the prediction well (*bottom-right*) for the prior models (*dashed black*) and posterior models (*solid blue*)

Data Case 3 corresponds to an early tracer arrival at the observation wells 2 and 3 and subsequently at the prediction well. Figure 4 shows that the uncertainty in prediction for the 30 models matching the data (24,810 simulations were needed for rejection sampling) is significantly reduced, compared to the uncertainty resulting from the prior models.

In the three data cases presented above, different observed data lead to different reduction of the uncertainty space. For example, for Data Case 2, only a small reduction of uncertainty is obtained by applying inverse modeling. For this example, it may not be worth applying time-consuming inverse modeling, as the data do not appear to be influential on the prediction variable. In the next section, an analysis tool is presented which allows assessing how much the observed data are informative on the prediction variable, without the need for running costly iterative inverse methods.

4 Prediction-Focused Analysis (PFA)

The idea behind PFA is to provide a tool to analyze the need of time-consuming inverse modeling. PFA is not an inverse modeling technique, therefore it is not designed to replace inverse modeling. It can be seen as a sensitivity analysis on the relationship between the measured field data and the prediction variables.

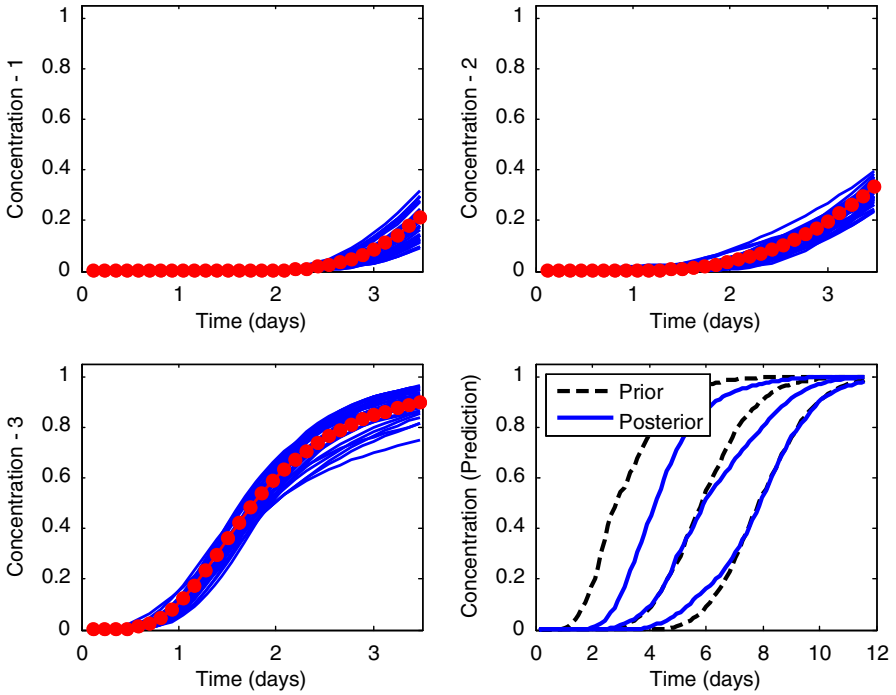

Fig. 3 Rejection sampling results for case 2. Tracer concentration at the three observation wells (*top and bottom-left*); uncertainty (P10, P50 and P90) at the prediction well (*bottom-right*) for the prior models (*dashed black*) and posterior models (*solid blue*)

4.1 Main Idea

The idea of PFA is to investigate the following questions: (1) is there a relationship between $\mathbf{d}$ and $\mathbf{h}$ and, (2) if there is a (most likely non-linear) relationship, can it be estimated. The first question can also be phrased as a diagnostic to verify if significant changes between prior $\rho(\mathbf{h})$ and posterior $\sigma(\mathbf{h})$ distributions are observed. If little or no relationship is observed, then there may be no need to perform time-consuming model-based inversion. To establish such a relationship, PFA requires only the construction of a set of *prior* Earth models. Both forward models $\mathbf{g}$ and $\mathbf{r}$ are evaluated on these Earth models to get the simulated data $\mathbf{d}$ and predictions $\mathbf{h}$. $\mathbf{d}$ and $\mathbf{h}$ are vectors of potentially large dimension, although, as typical for flow inverse problems, much smaller than the model dimension $\mathbf{m}$. Any analysis of a direct relationship between $\mathbf{d}$ and $\mathbf{h}$ will only be possible if a sufficient dimension reduction is obtained (an assumption made in the PFA approach). By applying such dimensionality reduction techniques individually on the simulated data and prediction, low-dimensional representations of $\mathbf{d}$ and $\mathbf{h}$ are created, which are denoted $\mathbf{d}^*$ and $\mathbf{h}^*$, respectively.

Having the compressed field $\mathbf{d}^*$ and $\mathbf{h}^*$, the objective is now to analyze the relationship between $\mathbf{d}^*$ and $\mathbf{h}^*$ and to compare approximate distributions of $\rho(\mathbf{h})$ and $\sigma(\mathbf{h})$ in reduced space, denoted, respectively, $\rho^*(\mathbf{h}^*)$ and $\sigma^*(\mathbf{h}^*)$. Note that this relationship will be informative on the data prediction relationship in physical space, if the dimen-

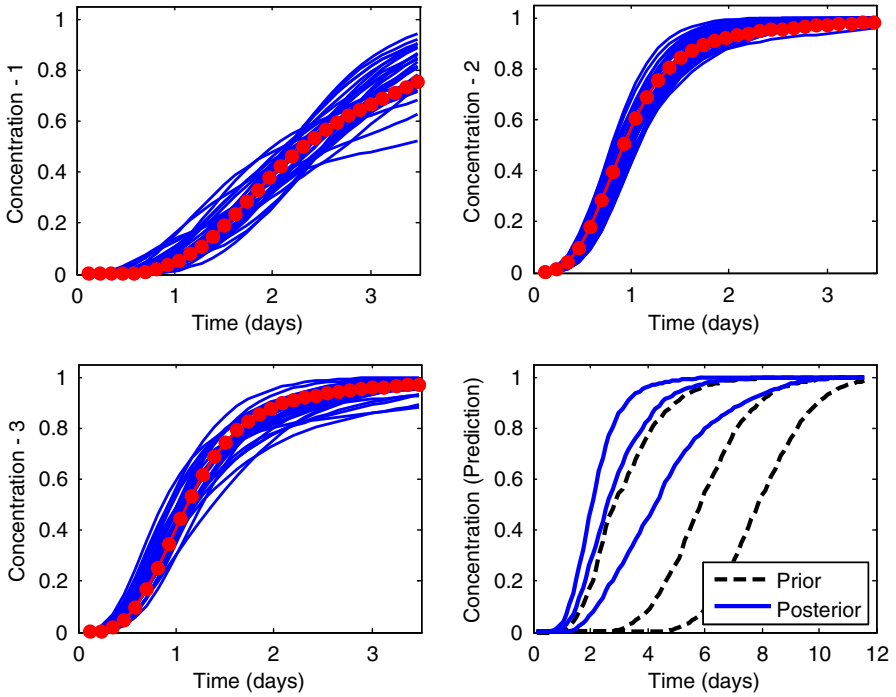

Fig. 4 Rejection sampling results for case 3. Tracer concentration at the three observation wells (*top and bottom-left*); uncertainty (P10, P50 and P90) at the prediction well (*bottom-right*) for the prior models (*dashed black*) and posterior models (*solid blue*)

sionality reduction technique possesses a few properties. First, points close to each other in physical space (in this case concentration curves) should remain close in low-dimensional space (proximity relationship should be preserved). Secondly, the transformation from the compressed field back to the initial data should be unique. If those properties are met, then differences in distributions in reduced space will be indicative of differences in distributions in physical space. The dimensionality reduction technique employed in this paper is non-linear principal component analysis (NLPCA, Kramer 1991). NLPCA has the advantage of dealing with the non-linear aspect of both **d** and **h**, and has a straightforward reconstruction of **d** and **h** from $\mathbf{d}^*$ and $\mathbf{h}^*$. Other candidates for dimensional reduction methods such as multi-dimensional scaling (Borg and Groenen 1997) or KPCA (Schöelkopf and Smola 2002) are not possible, due to the non-unique reconstruction from the reduced space to the physical space.

4.2 PFA as a Diagnostic Tool

NLPCA is a non-linear generalization of principal component analysis (PCA). While PCA is restricted to mapping only linear correlations among variables, NLPCA can reveal the non-linear association present (Caers and Srinivasan 2002, for an example

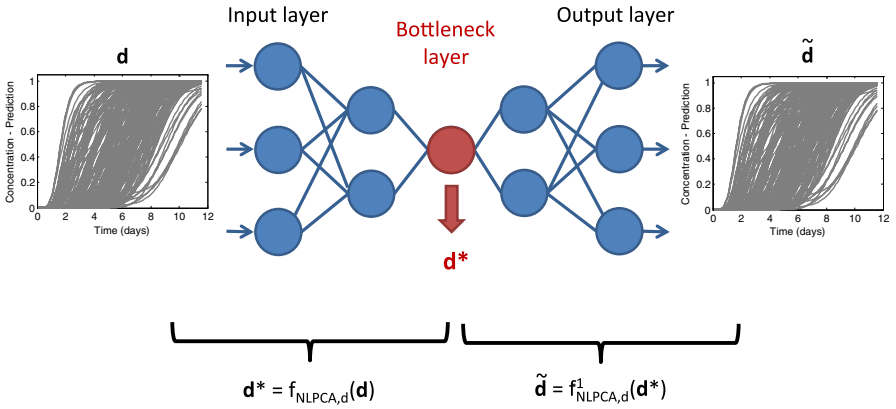

Fig. 5 Schematic illustration of the non-linear principal component analysis technique based on a neural network with a hidden bottleneck layer of low dimension

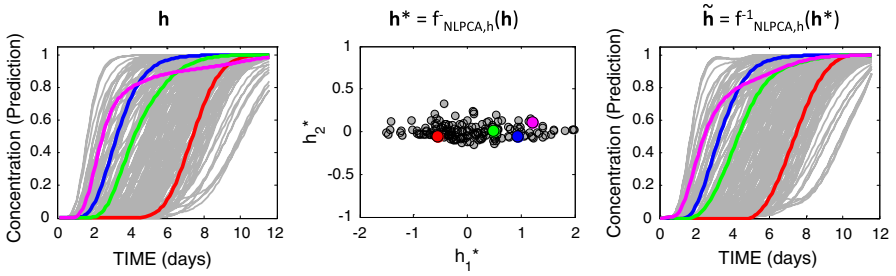

Fig. 6 NLPCA. Concentration curves that were used to train the neural network (*left*). Projection of each curve in two-dimensional space (*middle*). Back-transformation from reduced space to physical space (*right*)

of NLPCA on pressure data obtained from geostatistical realizations). NLPCA can be implemented using a neural network (Bishop 1995). The training of the network is done to obtain an identity mapping: the network outputs are simply the reproduction of network inputs (simulated data or prediction variables in this case). In the middle of the network is a layer that works as a bottleneck in which a reduction of the dimension of the data is enforced (Fig. 5). This bottleneck layer provides the desired component values (scores). A hidden layer in each part enables the network to perform non-linear mapping through a non-linear function. NLPCA requires a training data set which is simply obtained by generating a few (hundred) prior Earth models **m** and then evaluating both **d** and **h**. NLPCA presents the advantage of preserving proximities and having a unique back-transformation, which is a necessary assumption to study the distribution of the variables in reduced space. To illustrate, Fig. 6 displays the results from the neural network trained using a set of 200 prediction variables (left), 4 of which are highlighted in color. The projections of the four highlighted variables in a two-dimensional reduced space are shown in the middle plot. Finally, a back-transformation using the trained neural network is applied to construct approximations ($\tilde{\mathbf{h}}$) of the four prediction curves from the two-dimensional samples (right). It can be seen that each curve corresponds to one point (unique back-transformation) and that

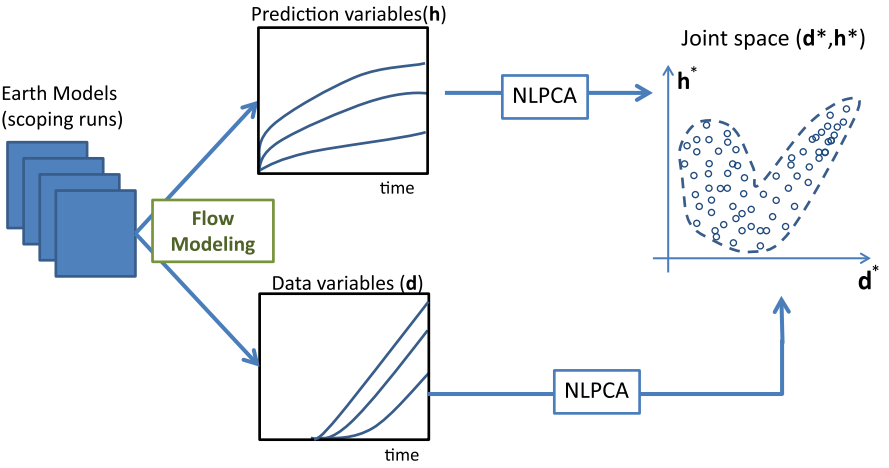

Fig. 7 Schematic representation of the joint space ($\mathbf{d}^*$, $\mathbf{h}^*$) from the data and prediction variables

a good preservation of the proximities is achieved. Finally, a good reconstruction of each curve is obtained.

NLPCA is applied successively on the observation and prediction variables to generate low-dimensional representations $\mathbf{d}^*$ and $\mathbf{h}^*$ of $\mathbf{d}$ and $\mathbf{h}$: $\mathbf{d}^* = f_{\mathrm{NLPCA,d}}(\mathbf{d})$ and $\mathbf{h}^* = f_{\mathrm{NLPCA,h}}(\mathbf{h})$. It then becomes possible to construct a joint space ($\mathbf{d}^*$, $\mathbf{h}^*$) as illustrated in Fig. 7 (right). In this schematic diagram, the x axis represents the compressed simulated data $\mathbf{d}^*$ and the y axis compressed predictions $\mathbf{h}^*$. Each point in this space represents the data-prediction compressed values of a single model, and thus allows analyzing how the data and prediction relate to each other. Only one dimension for $\mathbf{d}^*$ and $\mathbf{h}^*$ is represented for visualization purposes, even though higher dimensions can be used for each set of variables (usually 1 to 3 dimensions for each variable). The joint ($\mathbf{d}^*$, $\mathbf{h}^*$) space represents a global relationship between the simulated data and the prediction variables. If a trend is observed, it indicates that a relationship exists between the simulated data and prediction variables.

In addition, because the transformation from $\mathbf{d}$ to $\mathbf{d}^*$ is known through NLPCA, one can now calculate the compressed field data $\mathbf{d}_{\mathrm{obs}}{}^*$, the low-dimensional representation of the observed data ($\mathbf{d}_{\mathrm{obs}}$): $\mathbf{d}_{\mathrm{obs}}{}^* = f_{\mathrm{NLPCA,d}}(\mathbf{d}_{\mathrm{obs}})$. Since NLPCA preserves proximity distances, compressed fields $\mathbf{d}^*$ close to $\mathbf{d}_{\mathrm{obs}}{}^*$ have similar responses than the observed data $\mathbf{d}_{\mathrm{obs}}$. A local relationship can thus be estimated, the location of $\mathbf{d}_{\mathrm{obs}}{}^*$ in the ($\mathbf{d}^*$, $\mathbf{h}^*$) space being informative on the possible values of $\mathbf{h}^*$, for the given observed data. The distribution of $\mathbf{h}^*$ for the given observed data can thus be analyzed. $\mathbf{d}_{\mathrm{obs}}{}^*$ represents a hyperplane in ($\mathbf{d}^*$, $\mathbf{h}^*$) space. It is represented as a line in ($\mathbf{d}^*$, $\mathbf{h}^*$) space in two-dimensional figures such as shown in Fig. 8. In the schematic diagram presented in Fig. 8 (top, left), $\mathbf{d}_{\mathrm{obs}}{}^*$ lies in a part of the space where only high values of $\mathbf{h}^*$ are observed. This means that the data are informative on the prediction, the uncertainty in $\mathbf{h}^*$ (and therefore $\mathbf{h}$) is reduced for models having similar simulated data variables than $\mathbf{d}_{\mathrm{obs}}{}^*$. A different location of $\mathbf{d}_{\mathrm{obs}}{}^*$ is shown in Fig. 8 (bottom, left), where a much wider range of values of $\mathbf{h}^*$ is found for the given $\mathbf{d}_{\mathrm{obs}}{}^*$. Therefore, a practical way to analyze if the observed data are influential/informative on the prediction is to study

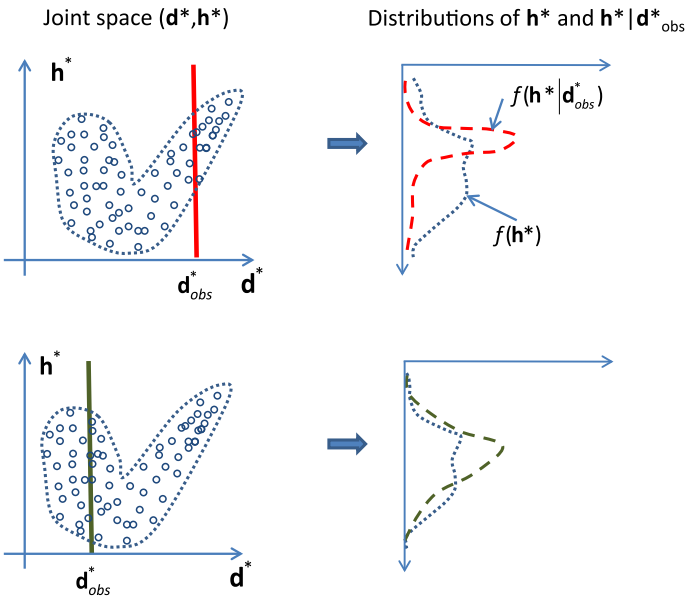

Fig. 8 Schematic diagram of the mapping of the observed data $\mathbf{d}_{\text{obs}}{}^*$ in ($\mathbf{d}^*$, $\mathbf{h}^*$) space. Estimation of the joint distribution and the marginal distribution for $\mathbf{h}^*$

the probability distribution functions (pdf) of $\mathbf{h}^*$ with and without the knowledge of $\mathbf{d}_{\text{obs}}{}^*$, which are denoted $f(\mathbf{h}^*)$ and $f(\mathbf{h}^*|\mathbf{d}_{\text{obs}}{}^*)$, respectively. These distributions are assumed to be approximations of the prior and posterior distributions of $\mathbf{h}$ in reduced space, and thus: $\rho^*(\mathbf{h}^*) = f(\mathbf{h}^*)$ and $\sigma^*(\mathbf{h}^*) = f(\mathbf{h}^*|\mathbf{d}_{\text{obs}}{}^*)$. Comparison of the prior and posterior distribution of $\mathbf{h}$ can now be done in reduced space: large changes in the distributions of $\mathbf{h}^*$ suggest that the data are informative on the prediction.

To generate these distributions, a kernel smoothing algorithm (Bowman and Azzalini 1997; Silverman 1986) is employed in ($\mathbf{d}^*$, $\mathbf{h}^*$) space. Kernel smoothing creates a map of density values at each location, which reflects the concentration of points in the surrounding area. When estimating $\sigma^*(\mathbf{h}^*)$, only points close to $\mathbf{d}_{\text{obs}}{}^*$ in ($\mathbf{d}^*$, $\mathbf{h}^*$) prior space should be accounted for (similarly to the tolerance defined in the likelihood function in physical space). This can be done by estimating first the joint density of ($\mathbf{d}^*$, $\mathbf{h}^*$) using kernel smoothing and by tuning the bandwidth for $\mathbf{d}^*$ to how accurate one want to represent the observed data $\mathbf{d}_{\text{obs}}$. The density $\sigma^*(\mathbf{h}^*)$ is then obtained by taking the joint density value at $\mathbf{d}_{\text{obs}}{}^*$ and normalizing (which is equivalent to dividing by the marginal). A larger bandwidth will allow more deviation from the observed data (more points are accounted for in the calculation of $\sigma^*(\mathbf{h}^*)$, resulting in a wider pdf), whereas a smaller bandwidth is reflective of observed data considered to be accurate (little or no measurement and model error), and thus resulting potentially in a narrower $\sigma^*(\mathbf{h}^*)$. Note that the bandwidth can be estimated based on the tolerance defined in the likelihood. Synthetic concentration curves within the tolerance can be constructed and subsequently projected in the low-dimensional space using $f_{\text{NLPCA,d}}$. The bandwidth can then be estimated based on the interval spanned by the points projected from the synthetic curves.

In the schematic example of Fig. 8, the diagnostic analysis indicates that a significant reduction of uncertainty in $\mathbf{h}^*$ is achieved in the top figure, but very little reduction for the bottom case. In the latter situation, according to the diagnostic, matching the data would not have much effect on the uncertainty assessment of the prediction variables $\mathbf{h}$. Note that even though the distributions are presented in two dimensions for illustrative purposes, in reality the marginal distributions are multi-dimensional. At this point of the study, the need of applying inverse modeling can be assessed based on the difference of distribution between $\mathbf{h}^*$ and $\mathbf{h}^*|\mathbf{d}_{\text{obs}}{}^*$. However, due to the properties of NLPCA, it is additionally possible to generate a rapid estimation of the prediction, in certain cases.

4.3 PFA for Estimation of Uncertainty in Prediction

This section describes how the diagnostic tool can be extended and used for generating new samples from the estimated distribution $\sigma^*(\mathbf{h}^*)$. The new samples indirectly define new prediction variables, which can be evaluated using the same neural network defined by NLPCA ($\tilde{\mathbf{h}} = f^{-1}_{\text{NLPCA,h}}(\tilde{\mathbf{h}}^*)$). New concentration curves can therefore be constructed directly from the sampling, without the need to generate and evaluate new Earth models. An approximation of $\sigma(\mathbf{h})$ is thus obtained, which is denoted $\sigma_{\text{PFA}}(\mathbf{h})$. For the estimation to be reasonable, the NLPCA should reproduce accurately new prediction variables from samples in reduced space that were not used to train the neural network. Figure 9 provides an illustration where values of $\mathbf{h}^*$ were generated by projecting new concentrations curves (not used for the training) in a reduced two-dimensional space and then back transformed using the neural network. A very good reconstruction of the concentration curves is observed. A Metropolis sampling algorithm (Metropolis and Ulam 1974; Metropolis et al. 1953) is used to generate the new samples from $\sigma^*(\mathbf{h}^*)$, denoted as $\tilde{\mathbf{h}}^*$. The corresponding new prediction variables can be constructed. Uncertainty can subsequently be derived from the distribution $\sigma_{\text{PFA}}(\mathbf{h})$ of those new prediction variables. An illustration is provided in Fig. 10.

5 Application of PFA to the Illustrative Case

5.1 PFA as a Diagnostic Tool

PFA is applied to the synthetic example presented in Sect. 3. In this example, 200 prior models $\mathbf{m}$ are generated and their responses $\mathbf{d}$ ($\mathbf{d} = g(\mathbf{m})$) and $\mathbf{h}$ ($\mathbf{h} = r(\mathbf{m})$) are evaluated and serve as inputs for PFA. NLPCA is applied on the 200 simulated data variables $\mathbf{d}$ as well as on the 200 predictions variables $\mathbf{h}$. For both cases, a hyperbolic tangent is used in the hidden layer of the neural network. Figure 11 (left) presents the joint space ($\mathbf{d}^*$, $\mathbf{h}^*$) using one dimension for the data variable ($\mathbf{d}^* = \mathbf{d}_1{}^*$) and one dimension for the prediction variable ($\mathbf{h}^* = \mathbf{h}_1{}^*$). Note that in both cases, the NLPCA achieves considerable dimension reduction, with 96.7 and 99.6 % of fraction of explained variance (FEV) for $\mathbf{d}^*$ and $\mathbf{h}^*$, respectively, for the first dimension, and 98.9 and 99.8 % when including the second dimension. As a consequence, the tracer concentrations

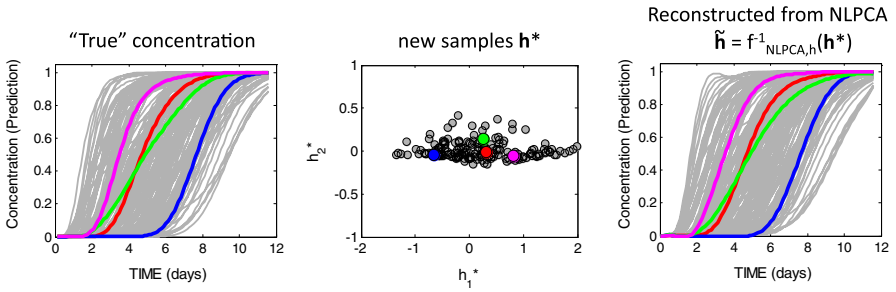

Fig. 9 NLPCA. Concentration curves that were used to train the neural network (*left*). Projection of each curve in 2D space (*middle*). Back-transformation from reduced space to physical space (*right*)

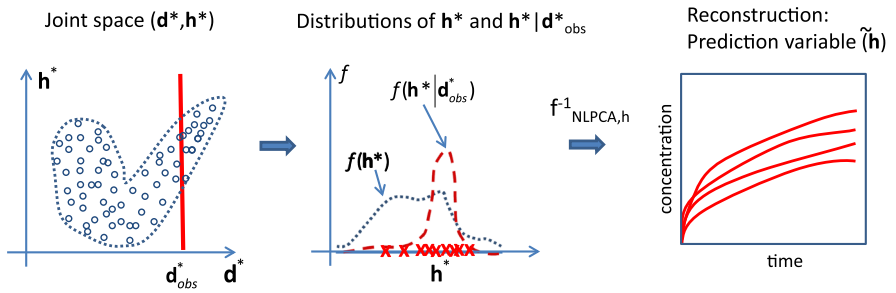

Fig. 10 Sampling from $f(\mathbf{d}^*, \mathbf{h}^* \,|\, \mathbf{d}_{\text{obs}}{}^*)$ and reconstruction of the corresponding variables

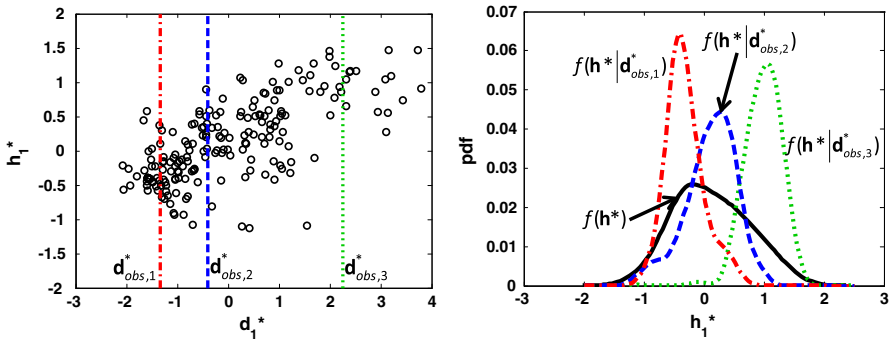

Fig. 11 2D joint space ($\mathbf{d}^*$, $\mathbf{h}^*$) using 200 initial models (*black circles*, *left*). The *lines* represent the observed data for the three different cases. Marginal distribution for the prior models (*black*) and for the posterior models (*right*)

(simulated data and prediction variables) can be explained on the basis of one or two-dimensional variables and an appropriate non-linear function (defined in the neural network).

The configuration of points in the ($\mathbf{d}^*$, $\mathbf{h}^*$) space (shown in Fig. 11) indicates whether there is a relationship between the data and prediction variables. High values of $\mathbf{d}_1{}^*$ (corresponding to early tracer arrival at the observation wells) correspond to high values of $\mathbf{h}_1{}^*$ (corresponding to early tracer arrival at the prediction well) and low

values of $\mathbf{d}_1{}^*$ correspond to low values of $\mathbf{h}_1{}^*$. The low-dimensional representations of the observed data for the three different cases (denoted $\mathbf{d}_{\text{obs},1}{}^*$, $\mathbf{d}_{\text{obs},2}{}^*$ and $\mathbf{d}_{\text{obs},3}{}^*$) are computed using the same NLPCA function ($f_{\text{NLPCA,d}}$) as for the simulated data variables.

Figure 11 (right) shows the distributions for the prior $\rho^*(\mathbf{h}^*)$ and posterior $\sigma^*(\mathbf{h}^*)$ in reduced space for each case. At this stage, PFA shows that the data are informative on the prediction variable for cases 1 and 3 (the prior and posterior distributions are significantly different), but less for case 2. In practice, in situations as illustrated in case 2, one could rely on the prior uncertainty distribution for the prediction without having to run the time-consuming inversion procedure. The accuracy of the diagnostic tool is confirmed by the results of rejection sampling, shown in Fig. 3, where the posterior uncertainty in the forecast almost covers the range of uncertainty of the prior. For cases 1 and 3, the diagnostic made with PFA suggests that there is a need to run the inversion procedure and indeed, Figs. 2 and 4 show an important reduction of uncertainty in the predictions after the integration of the data. So the results of the diagnostic are well in agreement with the rejection sampling.

5.2 Estimation of Uncertainty in Prediction

As stated, the PFA could also be used to obtain an approximate estimation of the posterior uncertainty in $\mathbf{h}$. To test whether this is the case for each of the three data cases, the posterior distribution $\sigma_{\text{PFA}}(\mathbf{h})$ obtained by PFA is compared to the posterior distribution $\sigma(\mathbf{h})$ obtained by rejection sampling. For brevity, results of PFA are only presented by choosing one dimension for $\mathbf{d}^*$ ($\mathbf{d}^* = \mathbf{d}_1{}^*$) and two dimensions for $\mathbf{h}^*$ ($\mathbf{h}^* = (\mathbf{h}_1{}^*, \mathbf{h}_2{}^*)$). However, the method has been applied for higher dimensions (2 dimensions for $\mathbf{d}^*$, 3 dimensions of $\mathbf{h}^*$), with similar results.

Data Case 1 (late arrival of the tracer at the three observation wells) PFA is applied in three dimensions (two dimensions for $\mathbf{h}^*$ and one dimension for $\mathbf{d}^*$). The posterior distribution $\sigma^*(\mathbf{h}^*)$ is therefore a two-dimensional pdf, which is represented by the sliced plane in Fig. 12, and as a contour map in Fig. 13 (left). Metropolis sampling of the kernel smoothed pdf is applied to generate 300 samples of distribution $\sigma^*(\mathbf{h}^*)$. Only the new samples $\tilde{\mathbf{h}}^*$ are shown by red "+" in Fig. 13 (left). Using the neural network that was used to create the compressed predictions $\mathbf{h}^*$, the predicted concentration profiles can be reconstructed directly from the sampled values $\tilde{\mathbf{h}}^* = (\tilde{\mathbf{h}}_1{}^*, \tilde{\mathbf{h}}_2^*)$: $\tilde{\mathbf{h}} = f^{-1}_{\text{NLPCA,h}}(\tilde{\mathbf{h}}^*)$. The estimated prediction variables $\tilde{\mathbf{h}}$ corresponding to each $\tilde{\mathbf{h}}^*$are shown in red in the middle plot (Fig. 13), the resulting uncertainty quantification is shown on the right. Figure 13 (right) shows that the uncertainty in the prediction, represented by the P10–P50–P90 values, obtained by PFA (red dash-dotted line) is significantly reduced compared to the prior uncertainty (shaded in grey). Finally, the uncertainty estimated using PFA (red dash-dotted line) is similar to the one obtained by rejection sampling (solid blue line).

Data Case 2 (late tracer arrival for the observation wells 1 and 2, and early arrival for well 3) The results are displayed in Fig. 14. Sampling was done using the distribution shown in Fig. 14 (left). As expected from the diagnostic analysis, the new prediction

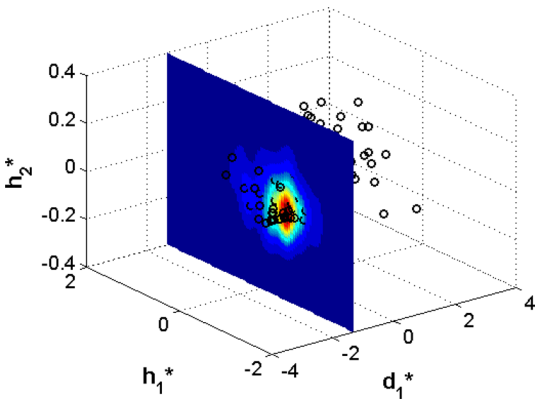

Fig. 12 3D joint space ($\mathbf{d}^*$, $\mathbf{h}^*$). The *slice* corresponds to the probability density $f(\mathbf{h}^* \,|\mathbf{d}_{\text{obs},1}{}^*)$

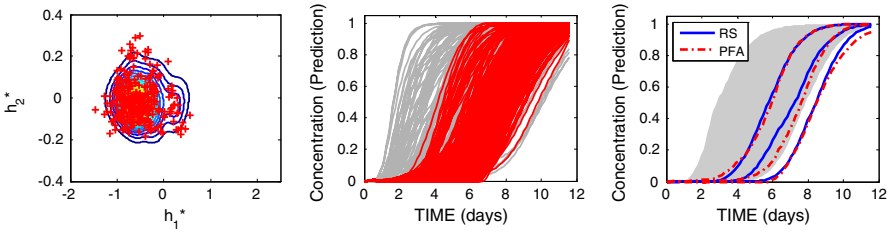

Fig. 13 PFA results for case 1. Marginal distribution $f(\mathbf{h}^* \,|\mathbf{d}_{\text{obs},1}{}^*)$, the *red* "+" represent the new samples (*left*). Corresponding prediction variables (*red lines, middle*). Comparison of uncertainty of the prior models (*shaded in grey*) and posterior models for PFA (*dash-dotted red*) and rejection sampling (RS) (*solid blue*, *right*)

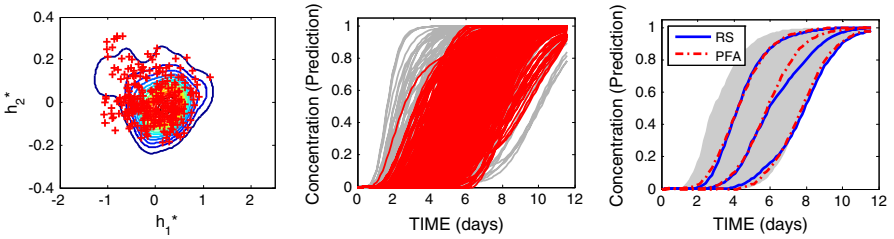

Fig. 14 PFA results for case 2. Marginal distribution $f(\mathbf{h}^* \,|\mathbf{d}_{\text{obs},2}{}^*)$, the *red* "+" represent the new samples (*left*). Corresponding prediction variables (*red lines, middle*). Comparison of uncertainty of the prior models (*shaded in grey*) and posterior models for PFA (*dash-dotted red*) and rejection sampling (RS) (*solid blue*, *right*)

curves (middle) as well as the P10–P50–P90 quantiles (right) obtained by PFA are fairly similar to the one from the prior. In this case, matching the data may not provide any better uncertainty in the prediction. The uncertainty in the prediction is not significantly reduced compared to the prior uncertainty (right). Furthermore, the uncertainty obtained by rejection sampling and PFA is very similar.

Data Case 3 (early tracer arrival at the observation wells 2 and 3 and subsequently at the prediction well) The results are displayed in Fig. 15. Note that only high values of $\mathbf{h}_1{}^*$ are sampled in this case, as shown by the empirical distribution (left). This results

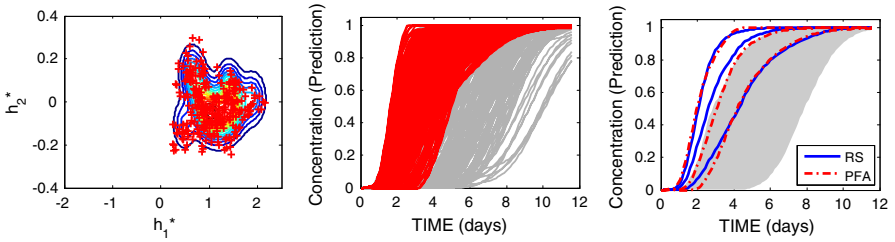

Fig. 15 PFA results for case 3. Marginal distribution $f(\mathbf{h}^* \,|\, \mathbf{d}_{\text{obs},3}{}^*)$, the *red* "+" represent the new samples (*left*). Corresponding prediction variables (*red lines, middle*). Comparison of uncertainty of the prior models (*shaded in grey*) and posterior models for PFA (*dash-dotted red*) and rejection sampling (RS) (*solid blue*, *right*)

in an early tracer arrival time for the prediction (middle). Figure 15 shows that the uncertainty in the prediction for the models matching the data is significantly reduced compared to the uncertainty estimated from the prior models. Again, the uncertainty estimation obtained by PFA is very similar to the uncertainty obtained by rejection sampling.

6 Discussion

The three examples above show that it is possible to establish a reliable diagnostic on how much the observed data are expected to be informative on the prediction variable. It is also possible to obtain a rather accurate estimation of uncertainty in predictions $\mathbf{h}$ without performing inverse modeling. In these examples, the PFA approach provided similar results as the full rejection sampler. This is encouraging from an Earth modeling point of view as the PFA only calls for forward models. No iterative inversion with possible problems of convergence is required; hence all modeling focus can be on the Earth models and the simulators. Finally, the direct estimation of uncertainty in prediction using PFA can only be applied in certain cases. A few models of the prior should already be in the vicinity of $\mathbf{d}_{\text{obs}}{}^*$ to ensure an accurate estimation of the bandwidth and reconstruction of the curves. This requirement is not necessary for the diagnostic tool where only large changes between the prior/posterior distributions of $\mathbf{h}^*$ are of interest. In addition, NLPCA should not be used to generate curves outside of the training bounds (no extrapolation is feasible), hence the importance of a wide prior.

In practice, the application of PFA will be possible and appealing when the following circumstances are found. First, when a very small number of prediction variables are sufficient to describe the problem and are known ahead of time. The PFA approach is not applicable when Earth models need to be created for multiple purposes that are not necessarily known a-priori. Second, when a significant dimension reduction can be obtained for both $\mathbf{d}^*$ and $\mathbf{h}^*$. Kernel smoothing will require more data (hence more forward simulations) as the dimensions of $\mathbf{d}^*$ and $\mathbf{h}^*$ increase and will likely not be successful in very high dimensions (in a recent paper, Park et al. 2013, such smoothing was successfully applied in dimensions as high as ten). If the dimension

reduction is not successful in low dimensions, then the loss can be quantified (as is the case with any dimension reduction method) and should be accounted for, or better, one should use a more appropriate dimension reduction technique. Third, when the prior model is broad enough to cover well the observed data, as would be expected in Bayesian modeling. Finally, when large uncertainty exists in the prior model with many contributing factors. Since only forward modeling is required, the PFA allows for any prior model with an amalgamation of uncertainty, such as uncertainty in the training image or structural model (e.g., faults, Park et al. 2013), or any other non-continuous types of variation. Discrete or scenario uncertainty are difficult to handle and computationally demanding within a classical inversion which tends to focus on within-scenario optimization or sampling.

7 Conclusions

In this paper, some fundamental questions related to flow inverse modeling have been investigated in cases with substantial uncertainty and important influence of prior information on the final inverse solutions. The prediction focused approach (PFA) is introduced to analyze the triangular relationship between data, model and prediction. PFA can be seen as a pre-inverse modeling sensitivity study. Based on NLPCA, PFA offers a diagnostic tool that allows investigating if matching the data, through inverse modeling, provides reduced uncertainty in the prediction. If the data are not informative on the prediction, then there may be no need to apply inverse modeling. If the data are shown to be sensitive, then inverse modeling should be performed. In the cases studied in the paper, a successful dimensionality reduction was obtained, which made an accurate diagnostic in only two dimensions possible. In addition, the cases demonstrate that it is possible to estimate the uncertainty in prediction relatively accurately without performing any flow inverse modeling. The uncertainty estimated by PFA was similar to the one obtained by rejection sampling for the three tested examples. While such results should be tested with more complex models and on a wider range of situations, one can be confident that the proposed approach may have considerable impact on the practice of subsurface modeling, the most direct being computational. Additional research is still needed to account for noise in the data and how such noise propagates in the dimensionality reduction. Finally, another extension of this research could be to investigate to what extent the data should be matched, or, what part of the data should be matched, to provide better predictions and lower computational cost.

Acknowledgments The authors would like to acknowledge the Swiss National Science Foundation for financial support under the contract CRSI22 122249/1: Integrated methods for stochastic ensemble aquifer modeling (ENSEMBLE) project.

References

Alcolea A, Renard P, Mariethoz G, Bertone F (2009) Reducing the impact of a desalination plant using stochastic modeling and optimization techniques. J Hydrol 365(3–4):275–288

Bayer P, Huggenberger P, Renard P (2011) Three-dimensional high resolution fluvio-glacial aquifer analog: part 1: field study. J Hydrol 40(5):1–9

Bishop C (1995) Neural Network for Pattern Recognition. Oxford University Press, New York
Bos CFM (2000) Production forecasting with uncertainty quantification. Final report of EC project, NITG-TNO report NITG 99–255-A
Borg I, Groenen P (1997) Modern multidimensional scaling: theory and applications. Springer, New York
Bowman AW, Azzalini A (1997) Applied smoothing techniques for data analysis. Oxford University Press, Oxford
Caers J, Srinivasan S et al (2002) Statistical pattern recognition and geostatistical data integration. In: Wong P (ed) Soft computing for reservoir characterization and modeling. Springer, New York, pp 355–386
Carrera J, Alcolea A, Medina A et al (2005) Inverse problem in hydrogeology. Hydrogeol J 13:206–222
Cherpeau N, Caumon G, Caers J, Levy B (2012) Method for stochastic inverse modeling of fault geometry and connectivity using flow data. Math Geosci 44(2):147–168
Comunian A, Renard P, Straubhaar J, Bayer P (2011) Three-dimensional high resolution fluvio-glacial aquifer analog—part 2: geostatistical modeling. J Hydrol 40(5):10–23
Ferraille M, Marel A (2012) Prediction under uncertainty on a mature field. Oil Gas Sci Technol 67:193–206
Freeze RA, Massmann J, Smith L, Sperling T, James B (1990) Hydrogeological decision analysis: 1. A framework. Ground Water 28:738–766
Gallagher M, Doherty J (2007) Predictive error analysis for a water resource management model. J Hydrol 334(3–4):513–533
Ginsbourger D, Rosspopoff B, Pirot G, Durrande N, Renard P (2013) Distance-based kriging relying on proxy simulations for inverse conditioning. Adv Water Res 52:275–291
Kramer MA (1991) Nonlinear principal component analysis using auto-associative neural networks. AIChE J 37:233–243
Künze R, Lunati I (2012) Adaptive multiscale simulations of density driven instabilities. J Comput Phys 231:5557–5570
Künze R, Lunati I (2011) MAFLOT: a matlab toolbox to simulate flow through porous media. Technical Report. University of Lausanne, Switzerland. http://www.maflot.com/
Metropolis N, Ulam S (1949) The Monte Carlo method. J Am Stat Assoc 44:335–341
Metropolis N, Rosenbluth AW, Rosenbluth MN et al (1953) Equations of state calculations by fast computing machines. J Chem Phys 21:1087–1091
Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. J Geophys Res 100(B7):12431–12447
Mosegaard K, Tarantola A (2002) Probabilistic approach to inverse problems. In: Lee W, Jennings P, Kisskinger C, Kanamori H (eds) International handbook of earthquake and engineering seismology, Part A. Academic Press, London, pp 237–265
Mugunthan P, Shoemaker CA (2006) Assessing the impacts of parameter uncertainty for computationally expensive groundwater models. Water Resour Res 42(W10428)
Oliver D, Chen Y (2011) Recent progress on reservoir history matching: a review. Comput Geosci 15: 185–221
Park H, Scheidt C, Fenwick DH et al (2013) History matching and uncertainty quantification of facies models with multiple geological interpretations. Comput Geosci 17:609–621
Ripley B (1987) Stochastic simulation. Wiley, New York
Schöelkopf B, Smola A (2002) Learning with kernels. MIT Press, Cambridge
Silverman BW (1986) Density estimation for statistics and data analysis. Chapman and Hall, London
Straubhaar J, Renard P, Mariethoz G et al (2011) An improved parallel multiple-point algorithm using a list approach. Math Geosci 43(3):305–328
Tavassoli Z, Carter JN, King PR (2004) Errors in history matching. SPE J 9(3):352–361
Zhou H, Gómez-Hernández JJ, Li L (2013) Inverse methods in hydrogeology: evolution and recent trends. Adv Water Res. doi:10.1016/j.advwatres.2013.10.014