Technical Commentary

Prediction-focused approaches: an opportunity for hydrology

Author: HERMANS Thomas, Urban and Environmental Engineering, University of Liege, Belgium and F.R.S.-FNRS postdoctoral researcher, Brussel, Belgium.

Quartier Polytech 1, Allée de la Découverte 9, 4000 Liege, Belgium

Tel.: +3243669263

Fax.: +3243669520

thomas.hermans@ulg.ac.be

Conflict of interest: none

Article Impact Statement: *Going directly from data to predictions through prediction-focused approaches: a changing paradigm for solving prediction problems*

Keywords: prediction-focused approach, stochastic inversion, Bayesian formalism, prior

Introduction

The ability of researchers and decision makers to anticipate the consequences of external events or of their actions in complex environments depends on the predictive capacity of models. In hydrology, many predictions can be reduced to a simple curve or map. One of the main paradoxes in hydrological predictions is the complexity of the process necessary to provide such a relatively simple model output.

The most common approach to providing predictions for complex subsurface systems is through physically based numerical modeling. A conceptual model of the aquifer is built on the basis of the available prior knowledge from borehole data, geophysical data, sample descriptions, tests, outcrops, etc. Then, the boundary conditions and the spatial heterogeneity are defined providing the set of parameters that constitute the (generally unique) model. Both conceptual/structural decision and parameter distribution are uncertain, although the former is often neglected in groundwater modeling. Generally, an inverse approach is used to estimate the distribution of the model parameters explaining some available data.

Deterministic approaches simplify the problem to reduce the number of unknowns and add some constraints on the distribution of the parameters to obtain a single solution (e.g., Carrera and Neumann, 1986). However, such solutions are generally not able to describe properly the heterogeneity of the true field. In addition, inverse problems in hydrogeology are ill posed so their solution is non-unique (Zhou et al. 2014). A prediction from a deterministic calibrated model, although physically based, might be far from the true response and, therefore, of limited use.

In contrast to the deterministic point of view, stochastic approaches aim at seeking not only one solution, but multiple realizations explaining the observed data with realistic patterns of heterogeneity (e.g., Oliver et al. 1997). The latter is called the posterior distribution of the

model parameters and allows quantifying uncertainty. Most stochastic inversion methods are based on a Bayesian formalism, where the posterior distribution is expressed as

$$f(\boldsymbol{m}\,|\,\boldsymbol{d}) = kf(\boldsymbol{m})L(\boldsymbol{m}\,|\,\boldsymbol{d}) \qquad (1)$$

where $\boldsymbol{m}$ represents the model parameters and $\boldsymbol{d}$ the data, $f(\boldsymbol{m})$ is the prior distribution of the model parameters, based on the prior knowledge we have about the problem and the study site, $L(\boldsymbol{m}|\boldsymbol{d})$ is a likelihood function, generally measuring the goodness of fit between the observed and the calculated data, and k is a proportionality constant. The set of posterior models can be used to estimate the desired predictions, assessing the uncertainty of these predictions.

Figure 1a sums up the limits of the current paradigm. We are generally more interested in some predictions $\boldsymbol{h}$ related to the subsurface model $\boldsymbol{m}$. But, current approaches first try to identify $\boldsymbol{m}$, a very complex subsurface model, using available data, and then generate the predictions $\boldsymbol{h}$. The full explicit inversion of model parameters is a challenging task. It is computationally difficult, either through the computation of large Jacobian matrix in deterministic calibration (e.g., Tonkin and Doherty 2005) or the thousands of runs required by stochastic inversion techniques (e.g., Mosegaard and Tarantola 1995) so that the conceptual model is often oversimplified.

Prediction-focused approaches

Principle

In many cases, we are not specifically interested in the model itself, only its outcomes. White (2017) discusses the need to take a model's objectives into account from the premise of the modeling approach. Along the same line, the principle of prediction-focused approaches (PFAs) is to replace the time-consuming inversion, centered on the data, by a more straightforward approach focusing on the prediction.

The objective of PFAs is to focus on finding a direct relationship between the data and the forecast. Such a relationship is inevitably dependent on the complexity of the subsurface, and therefore on the model. PFAs assume that this direct relationship can be obtained using an ensemble of surrogate subsurface models sampled from the prior. For each of these models, forward modeling generates both synthetic data and forecasts. Those outcomes are analyzed to see if a direct relationship exists (Figure 1b). Such an approach does not require computationally expensive inversions or thousand of runs, but only a limited number of forward simulations.

PFAs actually reformulate the prediction problem and the likelihood directly in terms of the forecast $\boldsymbol{h}$, not in terms of the model $\boldsymbol{m}$

$$f(\boldsymbol{h}|\boldsymbol{d}) = kf(\boldsymbol{h})L(\boldsymbol{h}|\boldsymbol{d}) . \qquad (2)$$

PFAs in practice

The key point of PFAs is to find a direct relationship between data and forecast to solve equation 2. Forward operators in hydrology are solving non-linear, partial-derivative, and coupled equations. Data and forecast variables, although less complex than models, remain generally high dimensional. Solving equation 2 remains challenging.

To overcome this limitation, PFAs propose to reduce the dimensionality of data and forecast variables, obtained through the surrogate models, to a few dimensions. Dimensionality reduction can be done based on physically meaningful variables, through principal-component-analysis-related techniques (PCA) or functional analysis. If the dimensionality of the reduced sets is sufficiently low, it might be possible to estimate directly the joint distribution of reduced data and forecast variables (Scheidt et al. 2015). An alternative is to

linearize the data-forecast relationship in the reduced dimension space using canonical correlation analysis (CCA) (Satija and Caers 2015; Hermans et al. 2016).

Knowing the relationship between data and forecast from our surrogate models (i.e., a way to compute the likelihood), it is now possible to derive the posterior distribution of the prediction in the reduced dimension space given the field observed data (equation 2). If the joint distribution is known, it can be sampled using standard techniques such as the Metropolis sampler. If CCA is used, the posterior distribution of the forecast can be obtained through linear regression. The last step of the procedure is to back transform the posterior into the original high-dimensional space.

Illustrative example

Figure 2 illustrates how a short-term heat tracing experiment can be used to predict the long-term heat recovery curve during a heat storage experiment forecasted using PFA:

(1) definition of the prior (geometry of the aquifer, boundary conditions, spatial heterogeneity, range of hydraulic conductivity, etc.) and generation of 500 surrogate subsurface models,

(2) simulation of both data (heat tracing) and prediction (heat storage) for the 500 models,

(3) reduction of the dimensionality of data and forecast variables, through PCA

(4) analysis and linearization of the direct relationship in the joint data/prediction space, through CCA,

(5) sampling of the posterior distribution given the observed data (black line in Data/Forecast relationship), through linear regression,

(6) back-transformation into the original high dimension space.

Recent works in hydrology (Scheidt et al. 2015; Satija and Caers 2015; Hermans et al. 2016) have revealed the potential high gain of PFAs to solve forecasting problems and quantify uncertainty. These approaches make it possible to integrate all types of uncertainty (including structural uncertainty) simultaneously in the prior. At the same time, computational effort is reduced because PFAs focus on predictions and do not attempt to solve for complex model parameter distributions. Finally, PFAs offer further computational efficiency because they require a limited number of forward simulations and can be fully parallelized.

Discussion

PFAs recognize that it is very difficult, if not impossible, to generate realistic subsurface models explaining the data and generating reliable predictions. Indeed, we generally observe a discrepancy between our predictions and actual observations. Therefore, PFAs suggest changing the current paradigm used to solve prediction problems. Instead of spending many efforts in the inversion/calibration process, PFAs focus on the generation of a realistic prior distribution and the analysis of the relationship between data and forecast through physically-based simulations. Using PFAs means admitting that *a realistic prediction can be obtained without calibrating a model, if we understand and are able to model the direct link between the data and the prediction.*

When can we expect PFAs to work ?

The focus of PFAs is prediction problems rather than identification of model parameters. Two conditions must be met for a successful application of the method.

1. The data has to be informative regarding the prediction of interest. A first measure is to check the sensitivity of both data and forecast on all uncertain model parameters

(including conceptual/structural uncertainty). A global sensitivity analysis (e.g., Park et al. 2016) carried out on the prior sets of data and forecast variables (already available) can reveal if data and forecast variables are driven by the same model parameters. The analysis of the direct relationship between data and forecast will reveal the risk of failure of the method. If CCA is used, a poor correlation indicates that it is difficult or impossible to estimate the forecast from the data. This might mean that the data are simply not informative for the prediction or that a more complex relationship (non-linear) has to be sought

2. It is of uttermost importance to define realistically the prior distribution and to check its consistency with data. In any Bayesian method (equations 1 and 2), the posterior distribution depends on the prior distribution.

Checking for prior-data consistency

The prior distribution expresses our prior knowledge about the problem and the study site and encompasses many components such as structural uncertainty (e.g., number and geometry of layers, physical limits of the models), geological scenarios and spatial heterogeneity (choice of one or several training images or variogram models, range of variations of hydrological parameters), and hydrological conditions (e.g., boundary conditions). Some physical choices, such as neglecting the effect of temperature, or assuming that the flow is not density-dependent, influence forward model outcomes and can be seen as part of the prior in PFAs.

Prior-data consistency relies on the comparison of the prior model outcomes ("synthetic data") with the real observed data (e.g., Hermans et al. 2015). This falsification process may reveal that some scenarios, some parameter values or combination of parameters considered in the prior are not likely given the observed data. It may also reveal that the prior is simply not consistent with the data, i.e. that the range of outcomes from the prior does not encompass

the observed data. Observing field data outside of the distribution of surrogate models is an indicator of an inadequacy between the proposed prior and the observed data. One should therefore reconsider its prior. Standard calibration methods may help to identify the problem.

To build the prior, Tarantola (2006) suggests that one should start with a very wide prior and then falsify models with the data (Popper-Bayes theory). Indeed, a restricted prior, although consistent with the data, might overlook the importance of non-considered parameters or scenarios, and reduce artificially uncertainty. Other models, maybe less likely could still be consistent with the data and lead to totally different predictions. Considering large prior will reduce the risk of seeking model surrogates that capture the bulk of model predictions, but that may miss « outliers » that are particularly important.

Research perspectives in PFAs

Working on the data/prediction relationship opens new challenges and opportunities to further develop PFAs. Ferré (2017) suggests an alternative view in decision-making problems through the use of advocacy and rival models representing specific stakeholders' concerns. One drawback of this approach is that model sampling is subjective and bears the risk to skew prediction ensembles. If the data is informative about the forecast, PFAs offer a way to map the direct relationship between both. If this relationship is good across the whole data/prediction space, including extremes, it would allow generating meaningful prediction ensembles for the advocacy models proposed by Ferré (2017). However, this would require sampling the limits of the plausible model space and therefore identify the kind of models leading to outcomes particularly important for stakeholders. The way those two approaches can complete each other should be investigated.

First PFA studies have investigated relatively simple prediction curves. Hermans et al. (2016) have shown that it was also possible to predict more complex variables (spatially distributed).

However, complex forecasts cannot be reduced easily to a few dimensions. The prediction of complex variables will require the use of advanced dimension reduction techniques, such as non-linear and kernel PCA or the generalization of functional analysis to higher dimensions (Gong et al. 2015).

Nowadays, advanced hydrological characterization is based on the acquisition of classical hydrological data but also of indirect data such as geophysical surveys or remote sensing. PFA techniques should therefore be further developed and improved to efficiently combine different data sources with various spatial coverage and time resolution. This may require the definition of new dimension reduction schemes, such as two-step PCA, to identify the redundancy between the different data sets, weight efficiently their importance in the prediction, and consider their different resolution and sensitivity to measurement error (Kikuchi et al. 2015).

Once the data/prediction relationship is known, PFAs allow fast estimation of the posterior distribution. This is precisely what is required to evaluate the worth of future prospective data, for experimental or monitoring network design (Kikuchi 2017).

The data/prior consistency is also an exciting research topic. The analysis of the data space can reveal important characteristics about the model itself. Surrogate models simulating data similar to the field observations might also share specific characteristics with the "true" subsurface model, such as a spatial heterogeneity model (spatial correlation or training image; e.g. Pirot 2017), a specific range of hydraulic conductivity or porosity. The analysis of the data-prior consistency can therefore be useful to update the prior distribution according to the observed data or to resample the prior in the region where the field data lies in order to improve the prediction and history-matching capacity (Caers et al. 2017).

The combination of PFAs with the large amount of computational power available on the cloud (Hayley 2017) makes the broad application of the Popper-Bayes theory more feasible. It is now possible to launch thousand of independent simulations simultaneously. This should allow the common investigation of all uncertainty components, including structural and spatial uncertainty.

The definition of the prior model is also the biggest challenge when working with field applications. Working in uncontrolled environments makes the definition of the prior difficult, but this is also true for other techniques. Nevertheless, applications of PFAs for field case studies in order to validate the methodology are mandatory.

Conclusion

Prediction problems in hydrology have generally been using an inversion approach, where a set of parameters is first inferred from data and subsequently used to forecast a system under varying scenarios. Instead of focusing on recovering a very complex distribution of model parameters, prediction-focused approaches assume that, it is possible to find a direct relationship between the data and the forecast, and to generate the full posterior distribution of the forecast without a fully explicit inversion. In PFA, this process is achieved through the simulations of many data and prediction sets, using subsurface surrogate models sampled from the prior distribution.

PFAs provide solutions that can drastically reduce the computational costs and enable us to realistically address the inherent uncertainty quantification associated with forecasting, a significant gain in the risk analysis and decision-making problem for complex systems.

PFAs are a very recent set of techniques under ongoing development. Future research should investigate more deeply the possibilities, advantages and limitations of PFAs and help to showcase PFAs as sound techniques for solving real-world prediction problems in hydrology.

Acknowledgement

I thank J. Caers for his supervision and mentoring during my postdoctoral stay at Stanford University and for involving me in the development of prediction-focused approaches. I am also grateful to T. Ferre for the opportunity to write a contribution within this collection of articles and to K. Hayley for his careful reviewing of the manuscript.

References

Caers, J., C. Scheidt and L. Li. 2017. *Quantifying Uncertainty in Subsurface Systems*. Wiley-Blackwell, Wichester, UK.

Carrera, J. and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research* 22, n°2: 199–210.

Ferre, T. 2017. Revisiting the relationship between data, models, and decision-making. *Groundwater* 55, n°5: xx-xx.

Gong, M., C. Miller and M. Scott. 2015. Functional PCA for Remotely Sensed Lake Surface Water Temperature Data. *Procedia Environmental Sciences* 26 : 127–130.

Hayley, K. 2017. The present state and future application of cloud computing for numerical groundwater modeling. *Groundwater* 55, n°5:.

Hermans, T., F. Nguyen and J. Caers. 2015. Uncertainty in training image-based inversion of hydraulic head data constrained to ERT data: Workflow and case study. *Water Resources Research* 51, n°7: 5332–5352.

Hermans, T., E.K. Oware and J. Caers. 2016. Direct prediction of spatially and temporally

varying physical properties from time-lapse electrical resistance data. *Water Resources Research* 52, n°9: 7262-7283.

Kikuchi, C., T.P.A. Ferré, and J.A. Vrugt. 2015. On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models. *Water Resources Research* 51, n°6: 4454–4481.

Kikuchi, C. 2017. Towards increased use of data worth analyses in groundwater studies. *Groundwater* 55, n°5:.

Mosegaard, K. and A. Tarantola. 1995. Monte Carlo sampling of solutions to inverse problems. *Journal of Geophysical Research* 100, n°7: 12431–12447.

Oliver, D.S., L.B. Cunha and A.C. Reynolds. 1997. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology* 29, n°1: 61–91.

Park, J., G. Yang, A. Satija, C. Scheidt and J. Caers. 2016. DGSA: A Matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments. *Computers & Geosciences* 97: 15–29.

Pirot, G. 2017. Training images as advocacy models to describe structural variability of spatio-temporal parameters. *Groundwater*, 55, n°5.

Satija, A. and J. Caers. 2015. Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources* 77: 69–81.

Scheidt, C., P. Renard and J. Caers. 2015. Prediction-Focused Subsurface Modeling: Investigating the Need for Accuracy in Flow-Based Inverse Modeling. *Mathematical Geosciences* 47, n°2: 173–191.

Tarantola, A., 2006. Poper, Bayes, and the inverse problem. *Nature Physics* 2: 492–494.

Tonkin, M.J. and J. Doherty. 2005. A hybrid regularized inversion methodology for highly

parameterized environmental models. *Water Resources Research* 41, n°10: W10412.

White, J. 2017. Forecast First : an argument for groundwater modeling in reverse. *Groundwater* 55, n°5:.

Zhou, H., J.J. Gómez-Hernández and L. Li. 2014. Inverse methods in hydrogeology: Evolution and recent trends. *Advances in Water Resources* 63: 22–37.

Figure captions

Figure 1. A) The current paradigm is based on the inversion of available data to generate "calibrated" model(s). Those models are then used to generate the desired prediction. B) In prediction-focused approaches, models are sampled from the prior distribution to compute "synthetic" data and predictions. A direct relationship is sought between data and predictions, later used for direct forecasting of the posterior distribution of the prediction for observed field data.

Figure 2 : Schematic overview of the different steps involved in PFAs.

Figures 1

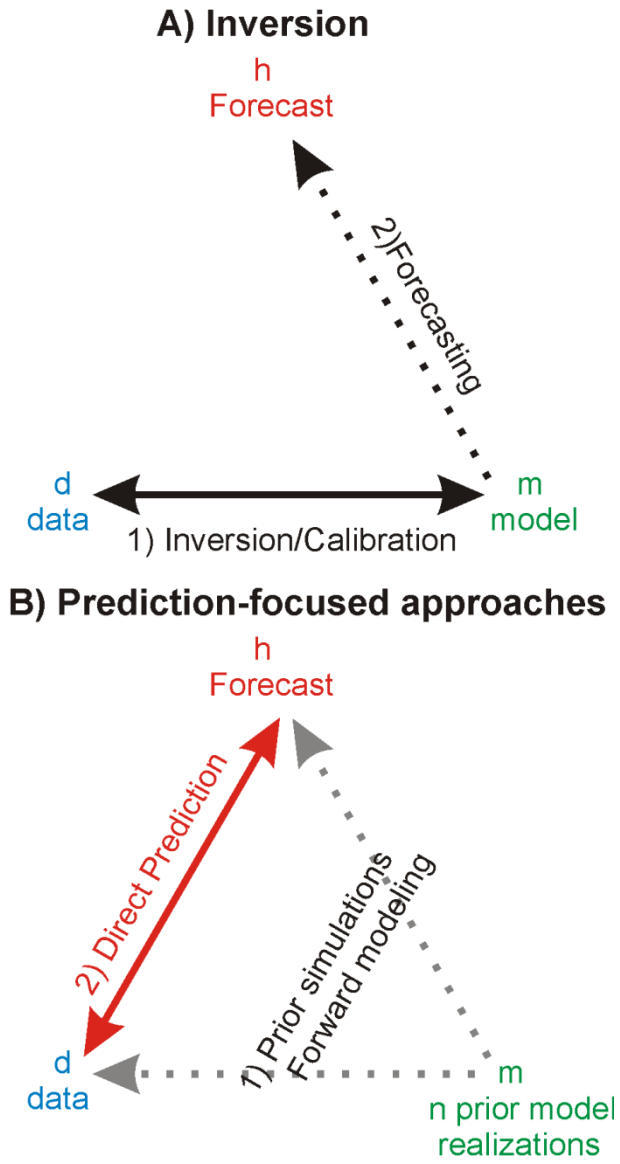
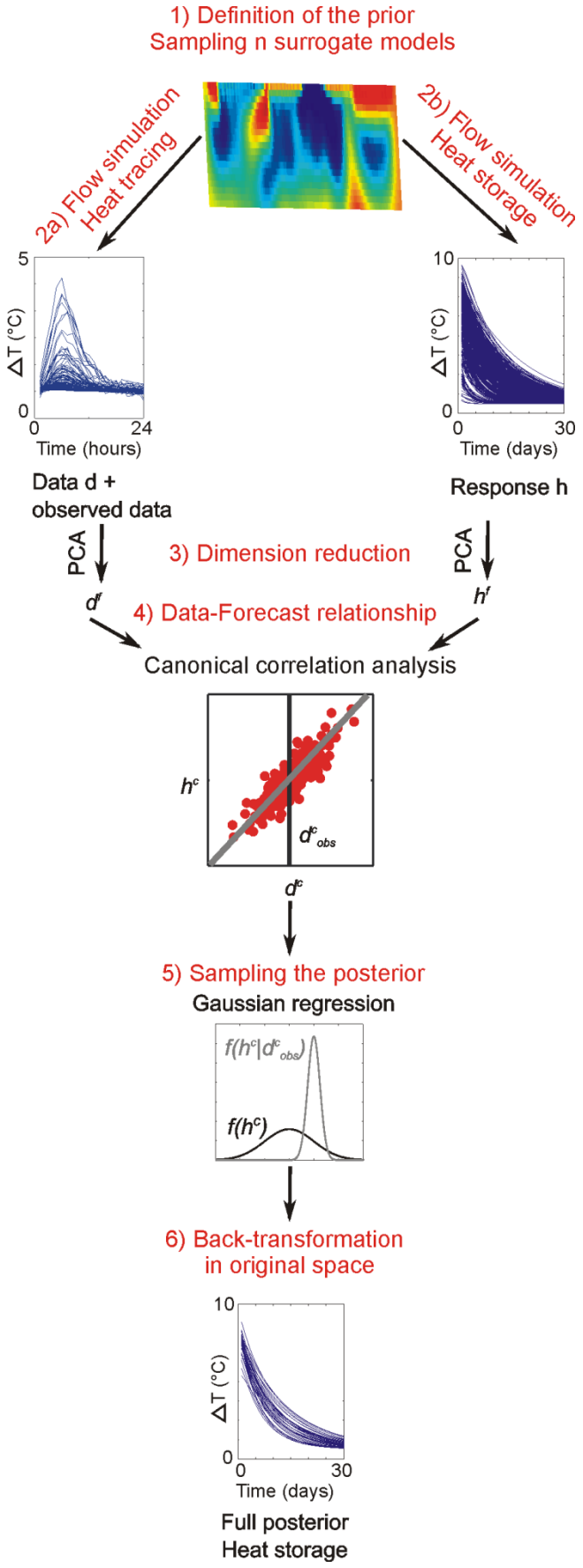

A) Inversion
h
Forecast
2)Forecasting
d
data
m
model
1) Inversion/Calibration
B) Prediction-focused approaches
h
Forecast
2) Direct Prediction
1) Prior simulations
Forward modeling
d
data
m
n prior model
realizations

Figure 2


1) Definition of the prior
Sampling n surrogate models
2a) Flow simulation
Heat tracing
2b) Flow simulation
Heat storage
5
0
ΔT (°C)
0
24
Time (hours)
10
0
ΔT (°C)
0
30
Time (days)
Data d +
observed data
Response h
PCA
3) Dimension reduction
PCA
d^f
4) Data-Forecast relationship
h^f
Canonical correlation analysis
h^c
d^c_{obs}
d^c
5) Sampling the posterior
Gaussian regression
$f(h^c|d^c_{obs})$
$f(h^c)$
6) Back-transformation
in original space
10
0
ΔT (°C)
0
30
Time (days)
Full posterior
Heat storage