# Efficient probabilistic hydrogeophysical prediction using maximum covariance analysis based metamodel of the forward problem

Erasmus K. Oware[1],
Jef Caers[2]
Thomas Hermans[3]

[1]*Department of Earth Sciences, SUNY at Buffalo, 126 Cooke Hall, Buffalo, NY 14260.*
[2]*Department of Geological Sciences, Stanford University, California, USA.*
[3]*Department of Geology, Ghent University, Belgium.*

**Key points**

- Improving computational efficiency of Bayesian inversion and sensitivity analysis problems in hydrogeology and hydrogeophysics.
- Finding coupled patterns between two space-time domains using maximum covariance analysis (MCA).
- First MCA-based metamodeling in hydrogeology and hydrogeophysics.

## Abstract

Bayesian inverse modeling (BIM) provides a framework for uncertainty assessment in estimation of earth models from multiple sources of limited, noisy, and incomplete data. BIM, however, requires repeated solutions of computationally expensive full-physics forward (FPF) runs, rendering BIM overwhelmingly computationally prohibitive and intractable in high-dimensional problems. We present a novel maximum covariance analysis (MCA)-based metamodel to approximate the FPF problem by capturing coupled patterns from small number of mutual observations between the parameter and data fields. We implement the strategy as a difference MCA prediction-focused approach (DMCA-PFA) and test it on electrical resistivity tomography (ERT) field data acquired during a heat-tracer experiment. We also invert the data using the conventional smoothness-constrained inversion (SCI). The DMCA-PFA outperformed SCI in estimation of realistic plume morphologies and direct temperature validations. We show an excellent MCA-based approximations of the FPF simulated resistances, with the potential to reduce CPU-time of BIM by over 99%.

## Plain Language Summary

Mathematical modeling is essential for efficient management of groundwater and energy resources. Such mathematical procedures require input parameters of some earth properties. Since the earth cannot be sampled exhaustively to obtain the input parameters at all locations, inverse modeling is performed to derive the input parameters at unsampled locations from observations at limited sampled locations. The inverse modeling procedure also requires forward modeling, which involves mathematical modeling of a process response given the input earth properties. The severally repeated forward modeling task, however, can become overwhelming computationally expensive, especially when the modeling involves several spatially distributed

parameters, which limits our inverse modeling capabilities. We present a novel maximum covariance analysis (MCA)-based strategy to approximate the forward problem in order to circumvent the tremendous computational costs of computing several forward solutions. We test the strategy on electrical resistivity tomography (ERT) field data acquired during a heat-tracer experiment. We also invert the data using the conventional smoothness-constrained inversion (SCI). Our strategy outperformed SCI in estimation of realistic plume morphologies and direct temperature validations. We show an excellent MCA-based approximations of the forward problem, with the potential to reduce CPU-time of inverse modeling by over 99%.

## 1. Introduction

Inverse modeling in hydrogeology [e.g., *Zhou et al.*, 2014] and hydrogeophysics [e.g., *Binley et al.*, 2015] involves estimating some earth physical properties and processes from observational data. Such observations are typically noisy and limited in number and resolution coupled with incomplete understanding of coupled processes occurring at multiple spatio-temporal scales. Bayesian inversion [*Tarantola*, 2005] provides a framework to infer earth models from multiple sources of information, including noisy and incomplete data while allowing incorporation of prior information. It also allows uncertainty quantification, thereby enabling comprehensive interrogation of the estimates. Due to the usually high-dimensional and complex non-linear models, the solution necessitates Markov chain Monte Carlo (McMC) sampling [*Hansen et al.*, 2016]. Bayesian McMC, however, requires repeated solutions of computationally expensive full-physics forward runs, which renders them computationally prohibitive and intractable in high-dimensional problems.

There is growing interest in the use of proxy-models of the forward solution in applications that demand several computationally expensive forward runs, such as Bayesian McMC and sensitivity analysis. There are two broad categories of proxy-models, the lower-fidelity models (LF-models) and metamodels [*Linde et al*., 2017]. The LF-models are usually physics-based and rely on model simplifications to speed-up the forward runs. The simplification can be achieved by approximating some aspects of the physics or by ignoring them completely [e.g., *Josset et al.*, 2015], or by performing the forward runs on a coarser grid [*Arridge et al*., 2015]. While LF-models speed-up the forward runs, they are less accurate. The metamodels, in contrast, are data-driven proxies involving Monte Carlo simulations of relatively small number of ensembles obtained from the full-physics (high-fidelity) forward simulations. There are several methods for developing the proxy-models from the small number of high-fidelity ensembles [e.g., *Khu and Werner*, 2003; *Myers et al*., 2016]. In the context of metamodeling in hydrogeology, the polynomial chaos expansion (PCE) [*Beck et al*., 2014] and reduced-order models (ROM) [e.g., *Liu et al.*, 2013] are commonly applied. In the PCE metamodeling, polynomial approximations of the forward problem are constructed over the support of the prior distributions [e.g., *Marzouk and Xiu*, 2009]. PCE, however, suffers in high-dimensional parameter spaces since the number of PCE terms increases dramatically with increased number of input parameters. PCE also underperforms when the input random field is highly heterogeneous. The ROM approach constructs orthogonal bases from the small number of high-fidelity ensembles (snapshots) and then employs the orthogonal bases as projection matrices to map the high-dimensional target system into a low-dimensional subspace. Unlike the current ROM techniques that consider orthogonal bases of a single state space, such as hydraulic heads [.e.g., *Liu et al*., 2013] or solute concentrations [e.g., *Oware et al.*, 2013; *Oware and Moysey*, 2014], our metamodeling relies on coupled (joint) patterns between the model parameter and data domains, making the strategy particularly well suited for metamodeling in model-data integration and sensitivity analysis applications. *Linde et al.* [2017] provides an excellent review of proxy-modeling in hydrogeology and hydrogeophysics.

Furthermore, to avoid the overwhelmingly computationally expensive Bayesian inversion, *Scheidt et al*. [2015] proposed the prediction-focused approach (PFA) to predict hydrologic variables directly from geophysical measurements without the full inversion of the data nor post-inversion petrophysical transform. The PFA uses surrogate models to derive a linearized, statistical relationship between the data and the target parameters [e.g., *Hermans et al*., 2016b]. While PFA shows a lot of promise, it predicts the target hydrologic models directly from the geophysical data without iteratively fitting the data, which might limit its robustness to reconstruct complex hydrologic features that are not well represented in the prior samples [*Oware et al.*, in-press]. The current PFA framework [e.g., *Satija and Caers*, 2015], moreover, uses canonical correlation analysis (CCA) to capture the coupled relationship between the two (model-data) space-time domains. The use of multivariate statistical tools, such as maximum covariance analysis (MCA) to capture coupled patterns between two space-time parameter fields for the purpose of forecasting is widely used in climate science [e.g., *von Storch and Zwiers*, 1999]. MCA is simply singular value decomposition (SVD) of the cross-covariance between the two space-time domains. *Bretherton et al.* [1992], for instance, found CCA to be uncompetitive compared to MCA due to high sampling variability unless the coupled signal was highly localized.

We propose here a difference MCA-PFA (DMCA-PFA) scheme to advance the PFA framework with two key contributions: 1) implement PFA in the MCA coupled space by actually fitting the data in a Bayesian sense, and 2) develop an MCA-based metamodel of the geophysical forward problem. We also outline a strategy to calibrate and account for metamodel-discrepancy in the proxy-approximation. We illustrate the performance of the DMCA-PFA on a field-scale geoelectrical data acquired during a heat-tracer experiment. We intend to show that the MCA-based metamodel presents a key contribution toward improving the computational efficiency of stochastic inversion and sensitivity analysis procedures in hydrogeology and hydrogeophysics.

## 2. Methods

### 2.1 Overview of Maximum Covariance Regression

Consider two parameter fields, a geophysical data field, $\mathbf{d} \in \mathbf{R}^{q}$ and a hydrologic model parameter space, $\mathbf{h} \in \mathbf{R}^{p}$, where $q$ and $p$ are the number of geophysical data points and hydrologic model parameters, respectively. A linear multivariate regression between the two domains can be expressed as:

$$\mathbf{d} = \mathbf{A}\mathbf{h}, \tag{1}$$

where **A** is a regression matrix. The mapping in Equation 1 usually involves complex, non-linear relationships including spatially dependent petrophysical transformation between the **d** and **h**. To linearize such complex relationships, we propose to use maximum covariance analysis (MCA). A good treatment of MCA is provided by *von Storch and Zwiers* [1999]. To accomplish this, we construct the data matrix $\mathbf{D} \in \mathbf{R}^{q\ x\ n}$ and the model parameter matrix $\mathbf{H} \in \mathbf{R}^{p\ x\ n}$ from Monte Carlo simulations of $n$ number of mutual observations (snapshots) between **d** and **h**. If $\mathbf{H}\mathbf{H}^{\boldsymbol{T}}$ is invertible, then **A** in Equation 1 can be factorized in terms of MCA projections [e.g., *Tippett et al.*, 2008]:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\boldsymbol{T}}(\mathbf{H}\mathbf{H}^{\boldsymbol{T}})^{-\mathbf{1}}, \tag{2}$$

where $\boldsymbol{T}$ denotes transpose, $\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\boldsymbol{T}}$ is the SVD of $\mathbf{D}\mathbf{H}^{\boldsymbol{T}}$, $\boldsymbol{\Lambda} \in \mathbf{R}^{pxp}$ is a diagonal matrix of singular values, and $\mathbf{U} \in \mathbf{R}^{q\ x\ p}$ and $\mathbf{V} \in \mathbf{R}^{p\ x\ p}$ are the left (data) and right (hydrologic model) coupled patterns, respectively. From Equations 1 and 2, we recast **d** as:

$$\mathbf{d} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\boldsymbol{T}}(\mathbf{H}\mathbf{H}^{\boldsymbol{T}})^{-\mathbf{1}}\mathbf{h} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\boldsymbol{\varepsilon}$ is the metamodel-discrepancy that accounts for the inexactness of the MCA-based approximation of the high-fidelity $\mathbf{d}$. A consequence of Equation 3 is that if we learn the mutual behavior between $\mathbf{d}$ and $\mathbf{h}$ and the metamodel-discrepancy structure from training surrogates $\mathbf{D}$ and $\mathbf{H}$, then we can directly predict $\mathbf{d}$ associated with any given $\mathbf{h}$ without the need for the typically computationally expensive geophysical forward simulations. We only need to run the high-fidelity forward simulations only *n* number of times.

In the event that $\mathbf{HH}^{\boldsymbol{T}}$ is not invertible directly, an inverted version can be approximated via SVD [e.g., *Castleman*, 1996]:

$$(\mathbf{HH}^{\boldsymbol{T}})^{-\mathbf{1}} = \left(\mathbf{U}_h\boldsymbol{\Lambda}_h\mathbf{V}_h{}^{\boldsymbol{T}}\right)^{-\mathbf{1}} \approx \mathbf{V}_h\boldsymbol{\Lambda}_h{}^{-\mathbf{1}}\mathbf{U}_h{}^{\boldsymbol{T}} \tag{4}$$

where $\mathbf{U}_h\boldsymbol{\Lambda}_h\mathbf{V}_h{}^{\boldsymbol{T}}$ is SVD of $\mathbf{HH}^{\boldsymbol{T}}$, $\boldsymbol{\Lambda}_h{}^{-\mathbf{1}}$ is a diagonal matrix with its diagonal elements equal to $1/\Lambda_{ij}$ and $\Lambda_{ij}$ are the diagonal elements of $\boldsymbol{\Lambda}_h$.

## 2.2 Difference Maximum Covariance Analysis Prediction-Focused Approach

Difference inversion [*LeBrecque and Yang*, 2001] has become increasingly appealing for geophysical monitoring of hydrogeological processes because inverting on the background differenced data results in rapid convergence, ability to detect small changes, eliminate systematic errors, and reduce inversion artifacts. Hence, we test the strategy as a difference maximum covariance analysis prediction-focused approach (DMCA-PFA). We apply Bayes' rule for the problem of estimating the posterior distribution of $\mathbf{h}$ from observed data, $\mathbf{d}_{obs}$. Specifically,

$$\mathbf{h}_{post} = \mathbf{h}_{prior}L(\mathbf{h}|\mathbf{d}_{obs}) \tag{5}$$

where $\mathbf{h}_{post}$ and $\mathbf{h}_{prior}$ are the posterior and prior distributions of $\mathbf{h}$, respectively, and $L(\cdot)$ is the likelihood, which evaluates the probability of a proposed $\mathbf{h}$ given $\mathbf{d}_{obs}$. We compute the likelihood as a multivariate Gaussian error distribution, i.e.,

$$L(\mathbf{h}|\mathbf{d}_{obs}, \mathbf{W}_d) = \exp\left[-\frac{1}{2}(\mathbf{e}^T * \mathbf{W}_d * \mathbf{e})\right], \quad (6)$$

where $\mathbf{W}_d$ is the data weight matrix and $\mathbf{e}$ is the data misfit. To implement Equation 6 in a difference inversion framework, we express $\mathbf{e}$ as:

$$\mathbf{e} = [\mathbf{d}_t - \mathbf{d}_0] - [f(\mathbf{h}_t) - f(\mathbf{h}_0)], \quad (7)$$

where $\mathbf{d}_t$ and $\mathbf{d}_0$ represent the data at the time-step of interest and background, respectively. The terms $f(\mathbf{h}_t)$ and $f(\mathbf{h}_0)$ are, respectively, the predicted data associated with a proposed model and the model obtained from the classical inversion of the background data. To advance the PFA framework by actually fitting the observed data in a computationally efficient manner, we circumvent the high-fidelity geophysical forward runs in Equation 7 by directly predicting the data for any given $\mathbf{h}$ according to Equation 3. Hence,

$$\mathbf{e} = [\mathbf{d}_t - \mathbf{d}_0] - [\mathbf{U\Lambda V}^{\boldsymbol{T}}\left(\mathbf{HH}^{\boldsymbol{T}}\right)^{-\mathbf{1}}\mathbf{h}_t - \mathbf{U\Lambda V}^{\boldsymbol{T}}\left(\mathbf{HH}^{\boldsymbol{T}}\right)^{-\mathbf{1}}\mathbf{h}_0 + (\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_0)]. \quad (8)$$

Similar to *Hermans et al*. [2016b], Equation 8 predicts the hydrologic model directly from the geophysical data, thereby avoiding post-inversion petrophysical transformation. There is growing popularity in estimation algorithms that proceed in the reduced-dimensional space due to their computational stability and efficiency [e.g., *Banks et al.*, 2000]. Hence, we express the target parameters, $\mathbf{h}_t$, as a linear combination of its basis vectors, $\mathbf{B}$, and expansion coefficients, $\mathbf{c}$, i.e., $\mathbf{h}_t = \mathbf{Bc}_t$ [e.g., *Oware et al.*, 2013]. Therefore,

$$\mathbf{e} = [\mathbf{d}_t - \mathbf{d}_0] - \mathbf{U\Lambda V}^{\boldsymbol{T}}(\mathbf{HH}^{\boldsymbol{T}})^{-\mathbf{1}}[\mathbf{Bc}_t - \mathbf{h}_0] - [\boldsymbol{\varepsilon}_t - \boldsymbol{\varepsilon}_0]. \quad (9)$$

It should be emphasized that the basis, $\mathbf{B}$, is constructed from SVD of $\mathbf{H}$ and, therefore, $\mathbf{B}$ is not the same as $\mathbf{V}$. While $\mathbf{V}$ captures the MCA coupled patterns between $\mathbf{H}$ and $\mathbf{D}$, $\mathbf{B}$ represents orthogonal bases of $\mathbf{H}$ only.

To obtain prior distributions for the metamodel-discrepancy, $\boldsymbol{\varepsilon}_t$, in Equation 9, we perform MCA-based approximations (Equation 3) of the geophysical data associated with the *n* number of training samples ($\mathbf{H}$), to construct the data matrix $\mathbf{D}_{mca} \in \mathbf{R}^{q \times n}$. We then compute the prior error, $\boldsymbol{\varepsilon}_{prior} \in \mathbf{R}^{q \times n}$, as the discrepancies between the predicted data of the high-fidelity geophysical forward runs and those of the MCA-based approximations, i.e.,

$$\boldsymbol{\varepsilon}_{prior} = \mathbf{D} - \mathbf{D}_{mca}. \tag{10}$$

This implies that there are *n* realizations of the approximation errors for each geophysical data point, which defines prior distributions for sampling the metamodel-discrepancy, $\boldsymbol{\varepsilon}_t$, for each geophysical data point. We compute $\boldsymbol{\varepsilon}_0$ in Equation 9 as the residuals between the predicted data of the high-fidelity geophysical forward run and that of the MCA-based approximation obtained from the inverted background geophysical model.

### 2.3 Complete Overview of DMCA-PFA

We now present the full workflow of the sampling of the posterior distribution of **h** (Figure 1):

1) Perform Monte Carlo simulations to generate training set (TS) of multiple realizations of the physics of the target hydrologic process (e.g., captures outcomes of multiple rates of advection and multiple scales of dispersion and plume morphologies). Collect all simulated time-lapse models into a single library of hydrologic (space-time) TS (**H**).

2) To obtain mutual observations between **h** and **d**, perform petrophysical transformation of each **h** into geophysical properties and run geophysical forward simulations to predict **d** associated with each **h**. Collect all simulated **d** into a matrix of geophysical TS, **D**. Now, **D** and **H** constitute mutual observations between the two parameter fields, **h** and **d**.

3) Perform MCA of **D** and **H** to obtain coupled patterns between the two fields. Also, SVD of **H** produces orthogonal bases, **B**.

4) To obtain prior distributions for sampling the expansion coefficients, **c**, project **H** onto **B**, i.e., $\mathbf{c}_{prior} = \mathbf{B}^{\boldsymbol{T}}\mathbf{H}$. Obtaining the prior coefficients from the physics-based TS imposes physics-based parameter bounds for the coefficients in an attempt to produce physically realistic plume morphologies [*Oware et al*., 2018].

5) Propose coefficients from $\mathbf{c}_{prior}$. We accept or reject the proposed coefficients based on the classical Metropolis-Hastings acceptance rule [*Metropolis et al*., 1953; *Hastings*, 1970]. The posterior coefficients are then mapped onto **B** to obtain multiple realizations of the target, i.e., $\mathbf{h}_{post} = \mathbf{B}\mathbf{c}_{post}$. Note, Step 5 is simply the standard McMC sampling parameterized in the reduced-dimensional space. It also uses the MCA-based metamodel without performing the typically computationally expensive geophysical forward runs.

## 3. Application to Field Data

We demonstrate the efficacy and efficiency of the DMCA-PFA algorithm on a field-scale heat-tracer experiment conducted in an alluvial aquifer and monitored with cross-well ERT (XBh-ERT). Details of the heat-tracer and XBh-ERT surveys are outlined in *Hermans et al*. [2015]. To summarize, water was continuously pumped to induce GW flow toward the pumping well. Hot water was then injected continuously in an injection well for 24 hours. Changes in electrical conductivity were monitored in a XBh-ERT panel across the GW flow direction. We invert the first six time-lapse resistances acquired at 6 h, 12 h, 18 h, 21.5 h, 25 h, and 30 h after the commencement of the heat injection. Each time-step inversion involves only 410 quadrupoles. Direct temperatures were also monitored in two piezometers, pz14 and pz15, for validation of the ERT predicted temperatures.

We first performed Monte Carlo simulations to obtain a training set (TS) tuned to the physics of the presupposed heat-tracer test. We used the same 3,000 (500 hydrologic models x 6 time-steps) temperature TS (**H**) employed by *Hermans at al*. [2018]. The TS was obtained via Monte Carlo simulations of the heat tracing experiment for 500 different GW models, considering uncertainties in the underlying hydrogeologic and transport properties. Through petrophysical transformations, we converted each temperature distribution, **h**, into resistivity models and ran resistivity forward simulations to obtain resistance data, **d**, associated with each **h**. A collection of all **d** comprise the geophysical training data (**D**). We then performed MCA of **D** and **H** to construct the coupled patterns between the two fields. Figure 2 shows the first 5 dominant MCA coupled patterns between the hydrological (log(**h**)) and geophysical data (**d**) spaces constructed from the 3,000 mutual observations between the two fields. Figure 2, essentially, depicts how the two fields covary such that given any resistance measurements, **d**, we should be able to leverage the prior coupled behavior to predict its associated temperature distributions, **h**. For comparison, we also inverted all the datasets using the classical smoothness-constrained inversion (SCI). We applied the 2.5D ERT inversion code CRTomo [*Kemna*, 2000] for all resistivity forward simulations and the SCI. We utilized the petrophysical relationship presented by *Hermans et al*. [2015] and the parameters presented therein for all conversions of ERT into thermograms.

## 4. Results and Discussion

### *4.1 MCA-based metamodeling*

For the inversion of $\mathbf{HH}^{\boldsymbol{T}}$ in Equation 3, we used Equation 4 since $\mathbf{HH}^{\boldsymbol{T}}$ was not invertible in the case study presented here. Histogram analyses (not shown) of the metamodel-discrepancies of the individual data points (Equation 10) reveal that the errors are not normally distributed. Hence, we assumed no knowledge of the prior error distributions and sampled uniformly over the interval of the prior errors of each data point for **ε** in Equation 3. To assess the performance of the MCA-based metamodel, we applied the high-fidelity resistivity forward simulation and MCA-based approximation to estimate resistances from resistivity tomograms obtained from smoothness-constrained inverison of the observed resistance data at three time steps, 12 hours (t2), 21.5 hours (t4), and 30 hours (t6).

Figure 3 shows the scatter plots of the full resistivity forward simulated resistances against those of the MCA-based metamodel for the three time steps. The coefficients of determination ($R^2$) for the MCA-based metamodel for t2, t4, and t6 are 0.9996, 0.9990, and 0.9988, respectively. The $R^2$s indicate marginal deterioration of the MCA metamodel with increasing time-steps. Nevertheless, there is almost a perfect one-to-one MCA proxy-approximations of all the high-fidelity forward simulated resistances, indicating high approximation accuracy of the MCA metamodeling in the examples considered here. It takes ~2.64 seconds of CPU-time to complete each high-fidelity resistivity forward simulation. This implies that Bayesian inversion involving about 300,000 iterations, for instance, will require ~13,200 minutes of CPU-time. The Bayesian inversion with the MCA-based metamodel (DMCA-PFA) presented in the next section takes ~35 minutes to complete 300,000 iterations. This represents a significant reduction in the computional time of ~99%, considering the 3,000 high-fidelity forward runs needed to calibrate the MCA-based metamodel and the ~35 minutes needed to complete the inversion.

### *4.2 Posterior Prediction*

We ran the algorithm for 300,000 iterations for all of the six time-lapse profiles. We applied 20 bases, **B**, to reconstruct the 1092 full-dimensional space, resulting in over 98% truncation in the dimensionality of the problem. The 20 selected basis vectors represented 99.8% of the total variance in the TIs. The difference thermograms recovered for the 12h (t2), 21.5h (t4), and 30h (t6) time-steps based on the SCI and the DMCA-PFA are presented in Figure 3. While both strategies captured similar evolution (locations and spatial extents) of the heat plume (Figure 3 Columns 1-4), smoothing of the heat plume is less severe in our approach in contrast to smoothing in the SC thermograms. The estimation of physically realistic plume morphologies without excessive smoothing in our approach (Figure 3 Columns 2-4) is attributable to the use of physics-based prior constraints as compared to the use of a generalized smoothness spatial filter in the SCI. The standard deviation panels (Figure 3 Column 5) reveal the spatial variabilities of uncertainty in the estimates. While uncertainty is expected to be low near the borehole locations (extreme vertical ends) due to high cross-borehole resistivity data sensitivity near the ERT wells, there appears to be generally high uncertainty in the recovered temperatures around 8-9 m depths, especially near the left borehole corresponding to high amplitudes. This trend in the estimated uncertainty reveal increasing difficulty of the algorithm to estimate high temperature deviations from the background values. The ability of our strategy to accurately capture the migration of the heat plume (different locations and morphologies) using the same set of basis-constraints demonstrates the flexibility of the algorithm to recombine the bases in a manner that honors each time-step ERT measurements.

*4.3 Model Validation with Direct Temperature Measurements*

Figure 5 outlines the validation of estimated temperature breakthrough curves at the two piezometers, pz14 and pz15, respectively, located at (1.125 m, 9 m) and (2.25 m, 8.5 m) from the left borehole. The temporal behavior of the validation breakthrough curves were accurately captured by both methods. Comparisons of the estimated temperatures with the direct temperatures, however, indicate that DMCA-PFA outperformed SCI on almost all the direct temperature measurements. The 90% confidence interval (CI) of the DMCA-PFA estimates captured all the true temperatures with the true values seemingly well centered within the range of the 90% CI. *Hermans et al*. [2018] concluded that a change of 1$^{o}$C produced only 2% change in electrical conductivity for the data presented here. Such small changes are undetectable in deterministic inversions [e.g., *Doetsch et al.*, 2012). *Hermans et al*. [2015] estimated the limit of detection of ERT of this experiment at ~1.5 $^{o}$C given the estimated noise level. Accounting for the physics of the target process seems to improve the limit of detection in the DMCA-PFA. Particularly, the direct temperature measurements at both pz14 and pz15 (Figures 5A and 5B) show that 6 hours (t1) of heat injection resulted in a change of ~0.5 $^{o}$C, which was undetected by the SCI since it is well below the ~1.5 $^{o}$C ERT detection limit. Our approach, in contrast, accurately estimated the small temperature change and captured the true values within 90% CI.

## 5. Conclusion

Inverse modeling is the foremost strategy for inferring earth properties and processes from observational data in hydrogeology and hydrogeophysics. In spite of the numerous benefits of stochastic inversion, deterministic inverse methods remain widely used due to their simplicity and computational efficiency. We propose here a novel maximum covariance analysis (MCA)-based metamodel to reduce the overwhelming computational costs of repeatedly computing the full-physics of the forward problem in stochastic inversions. We construct the MCA-based metamodel from coupled patterns captured from small number of ensembles of the joint evolution of the parameter and data fields. Hence, the strategy accounts for the physics of the target system on two fronts, the physics of the parameter and data acquisition systems. We conclude that incorporating the physics of the target process improves estimation and produces physically realistic target plumes without excessive smoothing in contrast to results obtained from the conventional smoothness-constrained inversion. We found an excellent MCA-based proxy-approximations of the full-physics forward simulated data, with the potential to reduce CPU-time of Bayesian inverse procedures by over 99%. The MCA-based metamodel presents a promising general framework to speed-up the computational efforts of hydrogeological and geophysical applications that necessitate repeated computations of the full-physics of the forward problem, such as high-dimensional Bayesian inversion and sensitivity analysis problems.

**Acknowledgements.**

The data and code required to reproduce the results are available from the first author upon request.

## References

Arridge, S., J. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, M. Vauhkonen (2006), Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Probblems*, 22 (1), 175–195. http://dx.doi.org/10.1088/0266-5611/22/1/010.

Banks, H.T., M.L. Joyner, B. Winchesky, W.P. Winfree (2000), Nondestructive evaluation using a reduced order computational methodology, *Inverse Problems,* 16, 1–17.

Beck, J., F. Nobile, L. Tamellini, R. Tempone (2014), A quasi-optimal sparse grids procedure for groundwater flows. Spectral and High Order Methods for Partial Differential Equations - ICOSAHOM 2012. *Lecture Notes in Computational Science and Engineering 95*. Springer, pp. 1–16. http://dx.doi.org/10.1007/978-3-319-01601-6_1.

Bretherton, C.S., C. Smith, and J.M. Wallace (1992), An intercomparison of methods for finding coupled patterns in climate data, *J. Climate*, 5, 541-560.

Binley, A., S. S. Hubbard, J. A. Huisman, A. Revil, D. A. Robinson, K. Singha, and L. D. Slater (2015), The emergence of hydrogeophysics for improved understanding of subsurface processes over multiple scales, *Water Resour. Res*., 51, doi:10.1002/2015WR017016.

Castleman, K.R. (1996), Digital Image Processing: Prentice Hall, Inc., Upper Saddle River, NJ.

Doetsch, J., Linde, N., Vogt, T., Binley, A., & Green, A. G. (2012), Imaging and quantifying salt-tracer transport in a riparian groundwater system by means of 3D ERT monitoring, *Geophysics*, 77, B207‑B218. https://doi.org/10.1190/geo2012-0046.1

Hansen, T.M., K.S. Cordua, A. Zunino, K. Mosegaard, (2016), Probabilistic integration of geo-information. In: Moorkamp, N.L.M., Leliévre, P.G., Khan, A. (Eds.), Integrated Imaging of the Earth: Theory and Applications, John Wiley & Sons, Inc, pp. 93–116.

Hastings, W. (1970), Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, 57, no. 1, 97.

Hermans, T, E. K. Oware, and J.K. Caers, 2016b, Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data: Water Resources Research, 52, no. 9, 7262-7283, doi.org/10.1002/2016WR019126.

Hermans, T., F. Nguyen, M. Klepikova, A. Dassargues, & J. Caers (2018), Uncertainty quantification of medium-term heat storage from short-term geophysical experiments using Bayesian evidential learning, *Water Resources Research*, 54.

Hermans, T., S. Wildemeersch, P. Jamin, P. Orban, S. Brouyere, A. Dassargues, and F. Nguyen (2015), Quantitative temperature monitoring of a heat tracing experiment using cross borehole ERT: *Geothermics*, 53, 14–26, doi:10.1016/j.geothermics.2014.03.013.

Josset, L., V. Demyanov, A.H. Elsheikh, I. Lunati (2015), Accelerating Monte Carlo Markov chains with proxy and error models, *Computer Geoscience*, 85, 38–48.

Kemna, A. (2000), Tomographic inversion of complex resistivity: theory and application: PhD Thesis, Bochum Ruhr University, Germany.

Khu, S.T., and M. Werner (2003), Reduction of Monte-Carlo simulation runs for uncertainty estimation in hydrological modelling, *Hydrol. Earth Syst. Sci.* 7 (5), 680–692.

LaBrecque, D.J., and X. Yang (2001), Difference inversion of ERT data: a fast inversion method for 3-D in-situ monitoring: Journal of Environmental and Engineering Geophysics, 6, no. 2, 83 – 89.

Linde, N, D. Ginsbourgerb, J. Irving, F. Nobiled, and A. Doucet (2017), On uncertainty quantification in hydrogeology and hydrogeophysics, *Advances in Water Resources*, 110, 166–181: http://dx.doi.org/10.1016/j.advwatres.2017.10.014.

Liu, X., Q. Zhou, J. Birkholzer, W.A. Illman (2013), Geostatistical reduced-order models in underdetermined inverse problems, *Water Resources Research*, 49 (10), 6587–6600.

Marzouk, Y., and D. Xiu (2009), A stochastic collocation approach to Bayesian inference in inverse problems. *Commun. Comput. Phys.* 6, 826–847. http://dx.doi.org/10.4208/cicp.2009.v6.p826.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953), Equation of state calculations by fast computing machines, Journal of Chemical Physics, 21, 1087–1092.

Myers, R., D. Montgomery, C. Anderson-Cook (2016), Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley.

Oware, E. K., and S. M. J. Moysey (2014), Geophysical evaluation of solute plume spatial moments using an adaptive POD algorithm for electrical resistivity imaging: Journal of Hydrology, 517, 471-480. http://dx.doi.org/10.1016/j.jhydrol.2014.05.054

Oware, E. K., S. M. J. Moysey, and T. Khan (2013), Physically based regularization of hydrogeophysical inverse problems for improved imaging of process-driven systems, *Water Resources Research*, 49**(**10), 6238-6247. http://dx.doi.org/10.1002/wrcr.20462.

Oware, E. K., M. Awatey, T. Hermans and J. Irving (2018), Basis-Constrained Bayesian-McMC: Hydrologic Process Parameterization of Stochastic Geoelectrical Imaging of Solute Plumes: *SEG Technical Program Extended Abstract*, October 14 - 18, Anaheim, CA, Proceedings, 5472–5476. 10.1190/segam2018-w12-01.1.

Oware, E. K., J. Irving, and T. Hermans (in-press), Basis-constrained Bayesian McMC difference inversion for geoeletrical monitoring of hydrogeological processes, *Geophysics Letters*.

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 Research*, 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, *Math Geoscience*, 47, 73‑91. http://dx.doi.org/10.1007/s11004-014-9521-6.

Tarantola, A. (2005), Inverse problem theory and methods for model parameter estimation, *Society of Industrial and Applied Mathematics*.

Tippett, M. K., T. Delsole, S. J. Mason, and A. G. Barnstone (2008), Regression-Based Methods for Finding Coupled, *J. Climate*, 21, 4384 - 4398.

von Storch, H. and F.W. Zwiers (1998), Statistical Analysis in Climate Research, *Cambridge University Press*, *Cambridge*.

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

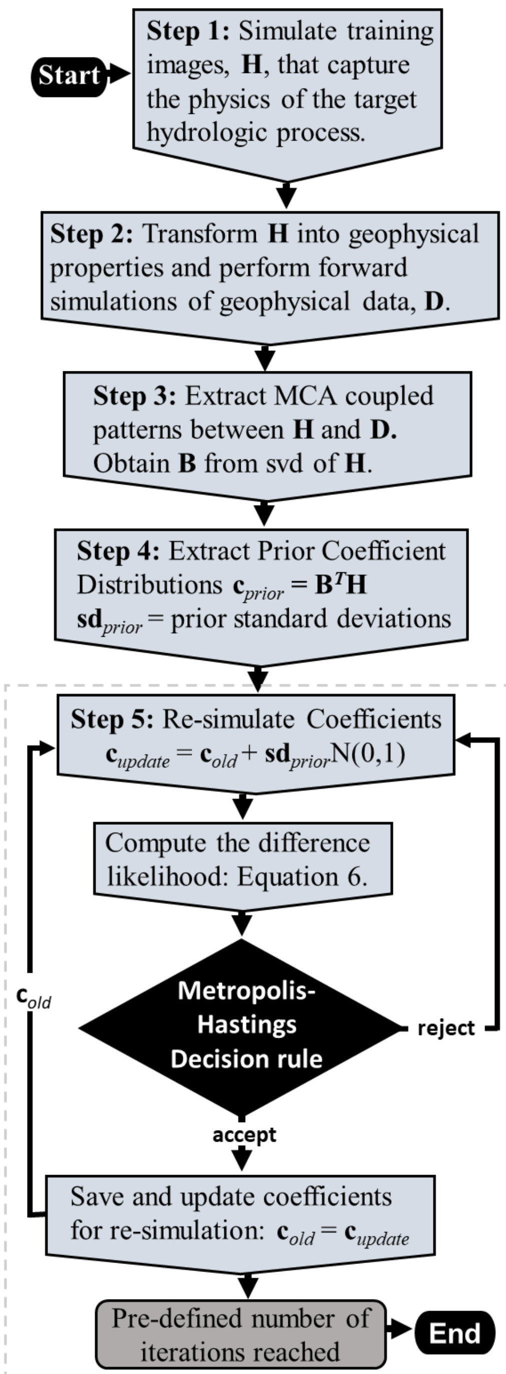


Figure 1: Flowchart for posterior sampling of the difference maximum covariance analysis prediction-focused approach (DMCA-PFA).

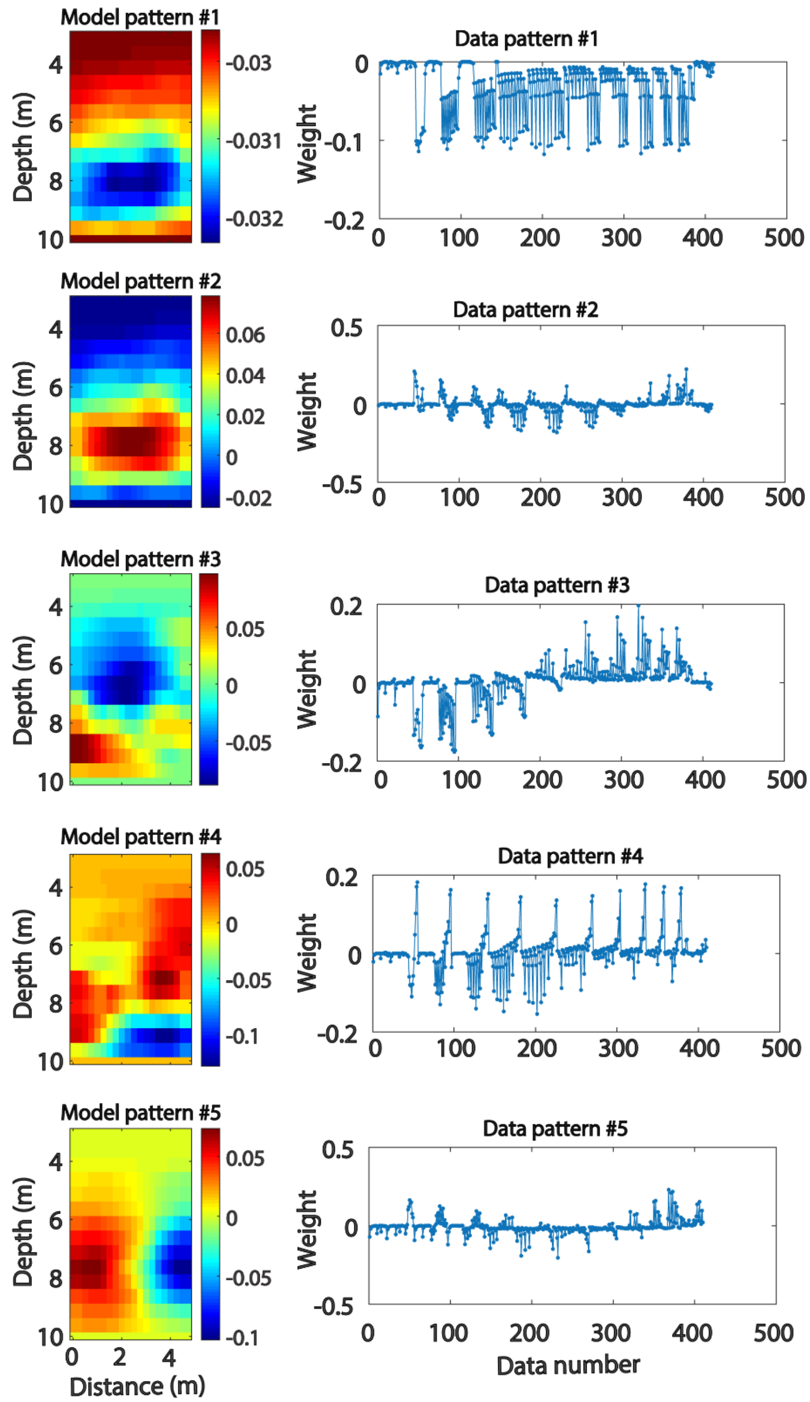


Figure 2: First 5 dominant maximum covariance analysis (MCA) coupled patterns between the hydrological parameter (log(temperature); column 1) and geophysical data (resistivity; column 2) fields constructed from training images of 3000 mutual observations between the two fields. The rows represent corresponding coupled patterns.

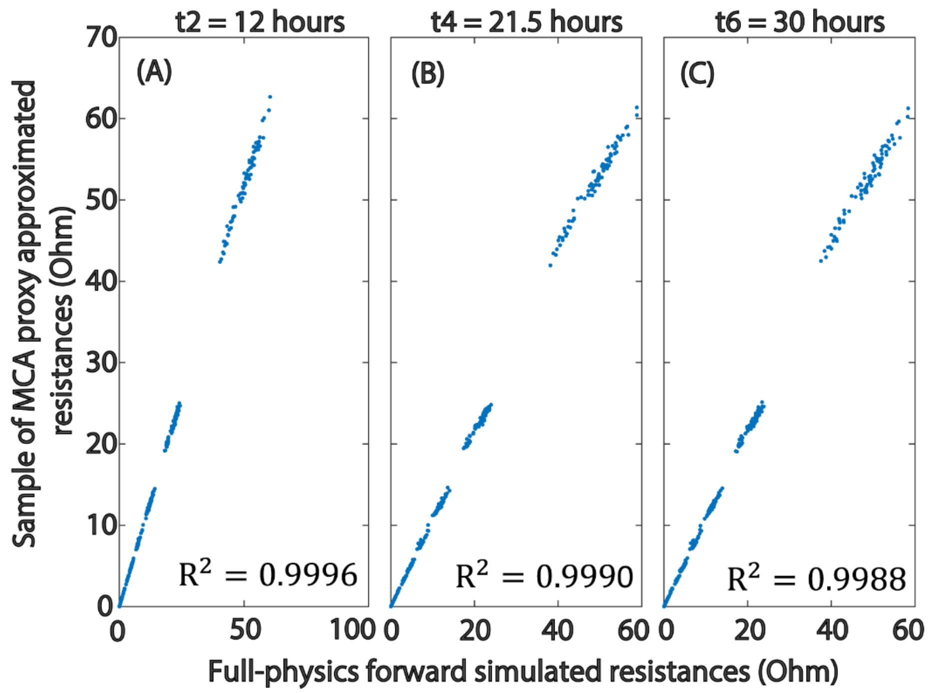


Figure 3: Scatter plots of high-fidelity resistivity forward simulated resistances vs sample of maximum covariance analysis (MCA) approximated resistances obtained from smoothness-constrained resistivity tomograms for time-step data at: (A) t2 (12 hours), (B) t4 (21.5 hours), and (C) t6 (30 hours). The coefficients of determination ($R^2$) indicate almost a perfect one-to-one MCA proxy-approximation of the high-fidelity forward simulated resistances.

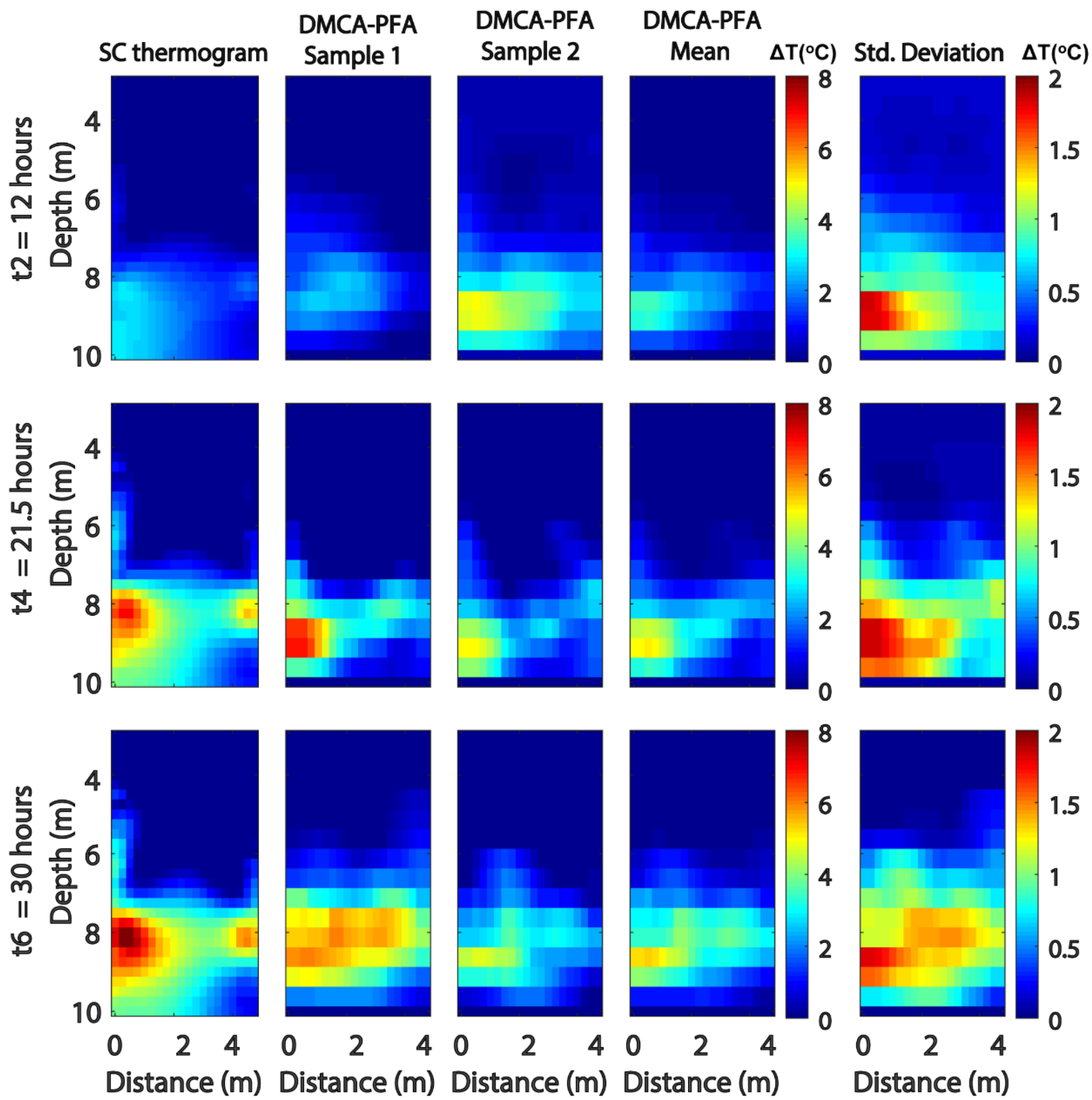


Figure 4: Difference thermograms recovered directly from the ERT measurements at three different time steps: (row 1) 12h, (row 2) 21.5h, and (row 3) 30h. Column 1 shows thermograms from the classical smoothness-constrained (SC) inversion, columns 2 , 3, 4, and 5 show, respectively, two realizations, posterior mean and standard deviations estimated from the difference maximum covariance analysis prediction-focused approach (DMCA-PFA). Piezometers pz14 and pz15 are, respectively, located at (1.125 m, 9 m) and (2.25 m, 8.5 m).

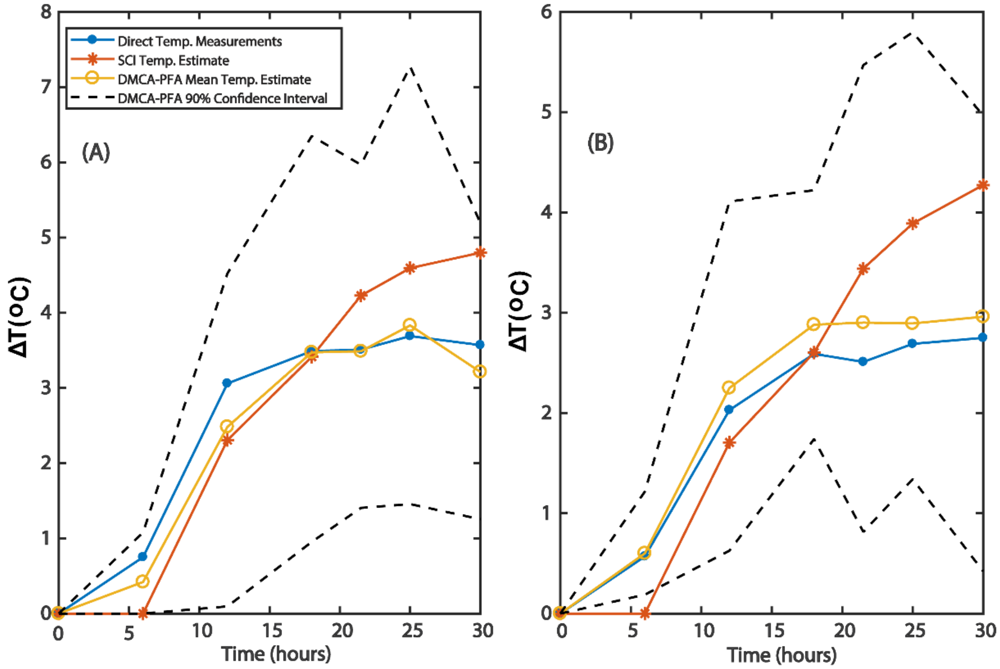


Figure 5: Validation of estimated temperature break-through curves at two validation locations: (A) pz14 and (B) pz15. (Blue lines) direct temperature measurements, and estimated temperature break-through curves from the: (orange lines) classical smoothness-constrained inversion (SCI), (yellow lines) posterior mean of the difference maximum covariance analysis prediction-focused approach (DMCA-PFA) estimates. The two black dashed lines define the 90% confidence interval of the DMCA-PFA predictions.