Data-worth Analysis through Probabilistic Collocation-based Ensemble Kalman Filter

Cheng Dai[1, 2, 3], Liang Xue[2,6*] Dongxiao Zhang,[3] and Alberto Guadagnini[4,5]

[1] State Key Laboratory of Shale Oil and Gas Enrichment Mechanisms and Effective Development, SINOPEC Group, Beijing, China

[2] Department of Oil-Gas Field Development Engineering, College of Petroleum Engineering, China University of Petroleum, Beijing, China.

[3] Department of Energy and Resources Engineering, College of Engineering, Peking University, Beijing, China.

[4] Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy.

[5] Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.

[6] State Key Laboratory of Petroleum Resources and Prospecting, China University of Petroleum, Beijing, China.

[*] Corresponding author, Department of Oil-Gas Field Development Engineering, College of Petroleum Engineering, China University of Petroleum, Beijing 102249, China. (xueliang@pku.edu.cn)

Abstract

We propose a new and computationally efficient data-worth analysis and quantification framework keyed to the characterization of target state variables in groundwater systems. We focus on dynamically evolving plumes of dissolved chemicals migrating in randomly heterogeneous aquifers. An accurate prediction of the detailed features of solute plumes requires collecting a substantial amount of data. Otherwise, constraints dictated by the availability of financial resources and ease of access to the aquifer system suggest the importance of assessing the expected value of data before these are actually collected. Data-worth analysis is targeted to the quantification of the impact of new potential measurements on the expected reduction of predictive uncertainty based on a given process model. Integration of the Ensemble Kalman Filter method within a data-worth analysis framework enables us to assess data worth sequentially, which is a key desirable feature for monitoring scheme design in a contaminant transport scenario. However, it is remarkably challenging because of the (typically) high computational cost involved, considering that repeated solutions of the inverse problem are required. As a computationally efficient scheme, we embed in the data-worth analysis framework a modified version of the Probabilistic Collocation Method-based Ensemble Kalman Filter proposed by Zeng et al. (2011) so that we take advantage of the ability to assimilate data sequentially in time through a surrogate model constructed via the polynomial chaos expansion. We illustrate our approach on a set of synthetic scenarios involving solute migrating in a two-dimensional random permeability field. Our results demonstrate the computational efficiency of our approach and its ability to quantify the

37 impact of the design of the monitoring network on the reduction of uncertainty
38 associated with the characterization of a migrating contaminant plume.

39 Keywords: Data worth; Ensemble Kalman Filter; Probabilistic Collocation;
40 Contaminant migration;

1. Introduction

Groundwater resources constitute a remarkable reserve of multipurpose (including potable, agricultural, and industrial) water and are critical for ecosystem and society sustainability and development (Foster and Chiltion, 2003). The severe challenges posed by the need to obtain accurate predictions of contaminant dynamics in natural aquifers hinge on the diverse sources of uncertainty associated with a selected predictive model. These uncertainties may originate from the (typically) unknown spatial distribution of subsurface geomaterials and the intrinsic heterogeneity of their hydrogeological properties, as well as from the insufficient level of knowledge of the key processes governing contaminant transport at the scale of interest. Stochastic inverse-modeling methods and data assimilation approaches have been developed to characterize uncertainties through model calibration against available measurements (e.g., Zimmerman et al., 1998; Alcolea et al., 2006; Fu and Gomez-Hernandez, 2009; Hendricks-Franssen et al., 2009; Riva et al., 2009; Rubin et al., 2010; Zhang et al., 2013; Zhou et al., 2014). In this context, the ensemble Kalman filter (EnKF) has gained significant popularity for the purpose of sequential (in time) data assimilation due to its relatively straightforward operational implementation and capability of quantification of predictive uncertainty (Anonsen et al., 2009; Chen and Zhang, 2006; Oliver et al., 2008; Xie and Zhang., 2010; Liu et al., 2012). Applications of the EnKF to contaminant transport settings can be found in (Liu et al., 2008; Li et al., 2012).

All inverse-modeling and data assimilation methods require the collection of suitable types of data for system characterization. While some inverse approaches, such

63 as those analyzed by Zimmerman et al. (1998) and Hendricks-Franssen et al. (2009)
64 and recently reviewed by Zhou et al. (2014), use all available data simultaneously, the
65 fast growing ability of setting monitoring networks with the capability of delivering
66 high-resolution real-time measurements has somehow increased the interest towards
67 assimilation methods capable of sequentially updating models as soon as data become
68 available. Even as sequential data assimilation tends to be computationally more
69 efficient than traditional batch inverse approaches, in practice, the high cost associated
70 with the extension of an existing monitoring setting suggests the need to develop robust
71 methodologies conducive to the identification of optimal strategies for the collection of
72 future data which are potentially valuable for a specific environmental goal considered.
73 Following Back (2007), such goal oriented data sets can assist to improve one's ability
74 to understand the system behavior and minimize uncertainty while considering budget
75 constraints. Applications of data-worth analysis in the context of groundwater-related
76 problems include the works of James and Freeze (1993), Abbaspour et al. (1996), Rada
77 and Schultz (1998), Russell and Rabideau (2006), and Dausman et al. (2010). Nowak
78 et al. (2010) and Leube et al. (2012) introduced the preposterior data impact assessor
79 (PreDIA) in the data-worth analysis framework and applied it to a (late time) steady-
80 state solute transport scenario taking place in a random hydraulic conductivity field.
81 The use of PreDIA to assist the optimal design of a monitoring scheme by minimizing
82 the probability of making incorrect decisions during Bayesian hypothesis testing is then
83 introduced by Nowak et al. (2012) and illustrated with reference to the prediction of
84 contaminant arrival times at a sensitive location in an aquifer, under a variety of

85 uncertain system parameters.

86 The aforementioned rapid development of *in-situ* monitoring technologies
87 motivates us to explore methodologies to quantify data-worth sequentially, to allow for
88 an optimal and flexible dynamic (in space and time) adjustment of the monitoring
89 scheme. To the best of our knowledge, data-worth analysis in the context of a
90 dynamically evolving subsurface transport setting has not been studied in little
91 literatures, Kollat et al. (2011) propose a framework for the design of a monitoring
92 network upon combining the EnKF and multi-objective evolutionary optimization. In
93 this context, Zhang et al. (2015) employ the relative entropy, or the Kullback-Leibler
94 divergence, as a global metric according to which they study the issue of proposing a
95 monitoring network for contaminant source identification. Here, we employ in our data-
96 worth analysis framework an approach which relies on the EnKF concept because: (*a*)
97 it is a non-intrusive method, which can be integrated in a straightforward manner in
98 available computational systems, (*b*) it allows the flexibility of providing the
99 uncertainty associated with the estimated system states at each assimilation step, thus
100 facilitating the data-worth analysis; and (*c*) it can be extended to enable us to handle
101 the challenges posed by the typically large number of parameters that are required to
102 characterize hydrogeological systems under uncertainty.

103 A critical element that might hamper the efficiency of the approach is the large
104 number of system replicates that are required to ensure the accuracy of the EnKF-based
105 results and minimize filter inbreeding (see, e.g., Panzeri et al., 2014 and references
106 therein). Therefore, considering that the planned additional measurements are not yet

107 collected and (at best) only estimates of these can be obtained from prior information,
108 the EnKF can still be computational demanding when incorporated in a data-worth
109 analysis. Panzeri et al. (2013, 2014) proposed to embed stochastic groundwater flow
110 moment equations (MEs) in the EnKF in a way that obviates the need for Monte Carlo
111 simulation. These authors demonstrated the computational feasibility and accuracy of
112 the methodology on a moderate size problem and showed that the approach mitigates
113 issues of filter inbreeding and spurious covariances often plaguing standard Monte
114 Carlo based EnKF. While theoretically and operationally elegant and effective for small
115 to medium size problems, the ME-based EnKF approach can be classified as an
116 intrusive method at the current stage of development, because it requires solving
117 equations satisfied by (conditional) statistical moments (ensemble means and
118 covariances) of hydraulic heads and fluxes in randomly heterogeneous media, the
119 structure of these equations being typically different from that of the equations
120 governing the dynamics of the actual (random) state variables. Here, we rely on a non-
121 intrusive approach which aims at improving computational efficiency through the
122 construction of a surrogate model (or proxy) to replace the original system model in the
123 standard Monte Carlo based EnKF. The use of surrogate models within a Bayesian
124 inference (data assimilation) framework has been explored in petroleum engineering by
125 Amudo et al. (2006) and Schaaf (2006). Carrer et al. (2007) relied on a variety of
126 algorithms, such as polynomial regression and Kriging, to construct proxies of the
127 target system response. Li et al. (2011) considered model proxies constructed by
128 polynomial regression for subsurface flow related problems and showed that Kriging

and neural network approaches may not yield accurate statistical moments of target state variables. As a consequence, results of an EnKF approach based on these types of proxies may be inaccurate.

In recent years, there has been increasing interest in the use of techniques based on Polynomial Chaos Expansion (PCE) for the construction of surrogate models of subsurface flow and transport processes. The method was first introduced by Ghanem and Spanos (2003) and has then been employed in a variety of areas and for diverse purposes, including optimization and global sensitivity analysis in the context of uncertainty quantification of selected model outputs (e.g., Ghanem, 1998; Reagan et al., 2005; Sudret, 2008; Fajraoui et al., 2011; Hays et al., 2011; Oladyshkin and Nowak, 2012; Oladyshkin et al., 2012; Ciriello et al., 2013a, b; Formaggia et al., 2013; Dai et al., 2014; Wu et al., 2014, and references therein). PCE-based surrogate models have also been combined with a variety of Bayesian updating methods, including, e.g., Markov chain Monte Carlo method (MCMC) (Marzouk, 2007; Jin, 2008), the EnKF method (Saad and Ghanem, 2009; He et al., 2011) and Bootstrap filter (Oladyshkin et al. 2012a, b) to alleviate computational burden. In essence, the PCE method relies on: (*a*) constructing a representation of the system state of interest in terms of a polynomial expansion expressed as a function of a set of uncertain system parameters; and (*b*) deriving appropriate discretized equations for the (deterministic) coefficients of the expansion through a Galerkin technique. The solution of these (typically coupled) equations can be computationally demanding in the presence of a large number of uncertain parameters and/or high-order terms in the polynomial approximation. In this

151 context, the Probabilistic Collocation Method (PCM) is an efficient non-intrusive
152 approach that can be employed to construct PCE-based proxies of groundwater flow
153 models. Li and Zhang (2007, 2009) explored the ability and efficiency of the PCM to
154 quantify uncertainty for single- and two-phase flow in randomly heterogeneous porous
155 media by combining the approach based on the Karhunen-Loeve (KL) expansion with
156 the PCE representation. Their results suggest that accurate estimates of key statistics of
157 variables of interest, such as pressure heads or saturations, can be obtained with a
158 limited number of runs of the original full system model. A PCM-based EnKF (PCKF)
159 has been employed by Li and Xiu (2009), Zeng (2010), Zeng et al. (2011), and Li et al.
160 (2014) for parameter estimation in flow settings typical of groundwater hydrology and
161 petroleum engineering applications.

162 In this work, we embed the PCKF into the data-worth analysis framework to assess
163 the worth of dynamically monitored data in a contaminant transport setting taking place
164 in a randomly heterogeneous aquifer. Doing so is consistent with the way we approach
165 the data worth challenge from a theoretical standpoint, according to which only
166 expected values (i.e., ensemble moments) of quantities of interest are required (see
167 Section 2). Due to the need for repeated solutions of the inverse problem at the
168 preposterior stage which can lead to high computational cost, we propose to increase
169 computational efficiency in the data-worth analysis context by modifying the way
170 PCKF is employed. In essence, unlike the PCKF introduced by Li and Xiu (2009), Zeng
171 (2010), Zeng et al. (2011) and Li et al. (2014), in which the purpose of the PCKF was
172 limited to improve the computational efficiency of the EnKF-based data assimilation in

173 the context of system parameter estimation and the statistics of uncertain geological
174 system parameters are estimated on the basis of the updated PCE coefficients, our
175 proposed modification to the PCKF method yields marked increase of computational
176 efficiency in the repeated EnKF process required during the preposterior data worth
177 analysis. To achieve this objective, we follow Zeng et al. (2012) and Marzouk and Xiu
178 (2014) and update the system parameters by adjusting the random quantities in terms
179 of which the PCE is constructed. As compared with the PCKF method originally
180 illustrated by Zeng et al. (2011), our modified and adapted PCKF requires prior
181 knowledge on the geostatistical descriptors characterizing the random parameter field,
182 such as mean, covariance and integral scale, so that these do not need to be updated
183 during the inverse modeling process. Through this modification, the entire simulation
184 process relies directly on the model proxy which has been constructed prior to actual
185 data collection, thus resulting in an alleviation of the computational burden associated
186 with performing repeated inversing modeling, as required at the preposterior stage.

187 Section 2 illustrates the theoretical bases and the workflow of our PCKF-based
188 data-worth analysis. Section 3 is devoted to the presentation of the set of examples that
189 we employ to demonstrate the ability of the proposed framework to quantify the impact
190 of the design of the monitoring network for solute concentrations (in terms of spatial
191 location, temporal sampling frequency, and prior data content) on the reduction of
192 uncertainty associated with the characterization of a contaminant plume migrating in a
193 randomly heterogeneous aquifer. Conclusions are presented in Section 4.

194 ## 2. Methodology

2.1 Data-worth analysis

We rely on the data-worth analysis framework proposed by Neuman et al. (2012) and Xue et al. (2014). The analysis is performed according to three stages. The first (or prior) stage relies on the prior data, i.e, the data which are available at the outset of the analysis. The second (or preposterior) stage relies on statistics of potential additional data conditional on the prior data. In this stage, the worth of the potential additional data from a given monitoring scheme is estimated. The third (or posterior) stage utilizes joint statistics of prior data and new data made available following the preposterior stage. The posterior statistics enable us to assess the quality of their associated preposterior estimates. Although we do not explore it in this work, our analysis framework can be extended to consider conceptual model uncertainty with the aid of the newly developed multimodel EnKF method (Xue and Zhang, 2014) which embeds uncertainty quantification associated with the use of diverse conceptual models of the system behavior.

Consider a set of discrete values of a target variable, representing, e.g., solute concentrations at a number of points distributed in space and time in an aquifer, and collected as the entries of a random vector, $\boldsymbol{\Delta}$. Here, we start by considering the mean, $E(\boldsymbol{\Delta}|\mathbf{D})$, and covariance, $Cov(\boldsymbol{\Delta}|\mathbf{D})$, of $\boldsymbol{\Delta}$ conditioned on a discrete set of prior available data forming the entries of vector **D** and predicted through a model characterized by a set of random variables grouped in vector $\boldsymbol{\xi}$. In the preposterior stage, we assume that the prior data set **D** is augmented by a set of data collected in vector **C'**. The entries of **C'** are yet to be observed and are planned to be collected in a

future monitoring campaign. At the preposterior stage, random estimates **C** of **C'** can be obtained on the basis of the results of the prior data-worth analysis, according, e.g., to the strategy illustrated in the following. Predictive statistics of $\boldsymbol{\Delta}$, i.e., mean, $E(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C})$, and covariance, $Cov(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C})$, are then calculated by joint conditioning on {**D**, **C**}. Note that the predictive statistics in the prior and preposterior data-worth analyses are theoretically related to each other by:

$$E(\Delta|\mathbf{D}) = E_{\mathbf{C}|\mathbf{D}}E(\Delta|\mathbf{D},\mathbf{C}) \tag{1}$$

$$Cov(\Delta|\mathbf{D}) = E_{\mathbf{C}|\mathbf{D}}Cov(\Delta|\mathbf{D},\mathbf{C}) + Cov_{\mathbf{C}|\mathbf{D}}E(\Delta|\mathbf{D},\mathbf{C}) \tag{2}$$

Here, $E_{\mathbf{C}|\mathbf{D}}E(\Delta|\mathbf{D},\mathbf{C})$ and $E_{\mathbf{C}|\mathbf{D}}Cov(\Delta|\mathbf{D},\mathbf{C})$, respectively, are the expectation of $E(\Delta|\mathbf{D},\mathbf{C})$ and $Cov(\Delta|\mathbf{D},\mathbf{C})$ over all **C** vectors generated, conditional on **D**; and $Cov_{\mathbf{C}|\mathbf{D}}E(\Delta|\mathbf{D},\mathbf{C})$ is the covariance of $E(\Delta|\mathbf{D},\mathbf{C})$ calculated over all **C** vectors generated, conditional on **D**. Following Neuman et al. (2012) and Xue et al. (2014), a scalar measure quantifying the data-worth can be introduced through the trace operator:

$$Tr\left[Cov(\boldsymbol{\Delta}|\mathbf{D})\right] = Tr\left[E_{\mathbf{C}|\mathbf{D}}Cov(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C})\right] + Tr\left[Cov_{\mathbf{C}|\mathbf{D}}E(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C})\right] \tag{3}$$

where *Tr* indicates the trace (sum of diagonal entries) of a matrix. Note that $Tr\left[E_{\mathbf{C}|\mathbf{D}}Cov(\Delta|\mathbf{D},\mathbf{C})\right]$ is given by the difference between the scalar measure of prior predictive uncertainty, $Tr\left[Cov(\Delta|\mathbf{D})\right]$, and the scalar measure of the estimated posterior predictive uncertainty, $Tr\left[E_{\mathbf{C}|\mathbf{D}}Cov(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C})\right]$. This quantity is to be compared against the reference data-worth, i.e., $Tr\left[Cov(\boldsymbol{\Delta}|\mathbf{D})\right] - Tr\left[Cov(\boldsymbol{\Delta}|\mathbf{D},\mathbf{C}')\right]$, which can be calculated at the posterior data-worth analysis stage, after **C'** has been observed.

The conditional moments introduced above need to be obtained through an

appropriate inverse-modeling method. For example, Xue et al. (2014) employed the geostatistical inversion method of Hernandez et al. (2003, 2006) based on equations satisfied by (ensemble) moments of the target state variable in a steady-state groundwater flow setting. The dynamic nature of the process of contaminant transport suggests evaluating the data-worth sequentially, in a data assimilation framework. To this end, our approach takes advantage of the flexibility of the Monte Carlo based EnKF method (Chen and Zhang, 2006; Liu et al., 2008) to allow for sequential (in time) and simultaneous updating of a collection of state vector realizations.

For completeness, we recount here the key theoretical elements underlying the typical EnKF approach. We start by considering a collection of N_e realizations of the state vector **S**:

$$\mathbf{S} = \left\{ \mathbf{s}^1, \mathbf{s}^2, ..., \mathbf{s}^{N_e} \right\} \tag{4}$$

Superscripts in (4) refer to the number identifying the realization associated with each vector **s** whose entries are given by the random quantities that characterize the model, ξ, the dynamic state variables, $\mathbf{u}$, and the observation data, $\mathbf{d}^{obs}$. Observations at time *t* and their true values are related by:

$$\mathbf{d}_t^{obs} = \mathbf{H}\mathbf{s}_t^{true} + \boldsymbol{\varepsilon}_t \tag{5}$$

Here, the superscripts *obs* and *true* respectively stand for the observation data and the true (usually unknown) system state; measurement errors collected in vector $\boldsymbol{\varepsilon}_t$ are assumed to be zero-mean Gaussian with covariance matrix $\mathbf{R}_t$; matrix **H** is the observation operator, which relates the state and observation vectors. The EnKF entails two stages, i.e., the forecast and assimilation stage. In the forecast step, each state vector

in the collection (4) is projected from time step ($t-1$) to time t via:

$$\mathbf{s}_t^{f,i} = F\left(\mathbf{s}_{t-1}^{a,i}\right); \qquad i = 1,2,...,N_e \tag{6}$$

The operator $F(\bullet)$ in (6) represents the forward numerical/analytical model of choice; superscripts f and a indicate the forecast and assimilation stage, respectively. In the assimilation stage, the Kalman gain, $\mathbf{G}_t$, is calculated as:

$$\mathbf{G}_t = \mathbf{C}_t^f \mathbf{H}_t^T \left(\mathbf{H}_t \mathbf{C}_t^f \mathbf{H}^T + \mathbf{R}_t\right)^{-1} \tag{7}$$

$\mathbf{C}_t^f$ being the covariance matrix of the system state. This matrix is approximated through the N_e model realizations as:

$$\mathbf{C}_t^f \approx \frac{1}{N_e} \sum_{n=1}^{N_e} \left\{ \left[\mathbf{s}_t^{f,n} - \left\langle \mathbf{s}_t^f \right\rangle\right] \left[\mathbf{s}_t^{f,n} - \left\langle \mathbf{s}_t^f \right\rangle\right]^T \right\} \tag{8}$$

Each state vector in the collection is then updated as

$$\mathbf{s}_t^{a,i} = \mathbf{s}_t^{f,i} + \mathbf{G}_t \left(\mathbf{d}_t^{obs,i} - \mathbf{H}\mathbf{s}_t^{f,i}\right) \tag{9}$$

It is worth noting that the EnKF method is characterized by a linear updating step due to the first-order-second-moment approximation. As such, the results based on this approach are optimal in the presence of moderate non-linearity of the processes governing the system dynamics.

The updated ensemble mean and covariance respectively are:

$$E\left(\mathbf{s}_t^a \mid \mathbf{d}_{1:t}^{obs}\right) = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{s}_t^a \tag{10}$$

$$Cov\left(\mathbf{s}_t^a \mid \mathbf{d}_{1:t}^{obs}\right) = \frac{1}{N_e - 1} \sum_{i=1}^{N_e} \left\{ \left[\mathbf{s}_{i,t}^a - E\left(\mathbf{s}_t^a\right)\right] \left[\mathbf{s}_{i,t}^a - E\left(\mathbf{s}_t^a\right)\right]^T \right\} \tag{11}$$

where $\mathbf{d}_{1:t}^{obs}$ are is the vector of observations collected up to time t, $d_{1:t}^{obs} = \left[d_1^{obs},...,d_t^{obs}\right]^T$.

Incorporating the EnKF into the data-worth analysis framework enables one to

quantify the worth of additional data in a sequential way, after assessing the uncertainty associated with data currently available. The workflow that we propose for assessing data-worth is depicted in Figure 1 and a synopsis of the various steps involved is detailed in the following.

1. At the prior stage, the EnKF is implemented on the basis of the observations available; a set of N_{e1} system realizations of state vectors is updated via (9) to obtain the ensemble mean $E(\mathbf{\Delta}|\mathbf{D})$ and covariance $Cov(\mathbf{\Delta}|\mathbf{D})$, conditioned on the prior data set **D**.
2. A number N_{e1} of random realizations of additional hypothetical data **C** are synthetically generated by assuming these are described through a multivariate Normal distribution, i.e., $\mathbf{C} \sim N\left[E(\mathbf{C}|\mathbf{D}), Cov(\mathbf{C}|\mathbf{D})\right]$. Note that **C** is a subset of $\mathbf{\Delta}$ obtained in step 1.
3. At the preposterior stage, EnKF is implemented through a set of N_{e2} Monte Carlo realizations for each of the N_{e1} hypothetical data vectors, **C**, obtained at step 2; this allows calculating updated ensemble mean $E(\mathbf{\Delta}|\mathbf{D},\mathbf{C})$ and covariance $Cov(\Delta|\mathbf{D},\mathbf{C})$, jointly conditioned on $\{\mathbf{D},\mathbf{C}\}$. Quantities $E_{\mathrm{C|D}}Cov(\Delta|\mathbf{D},\mathbf{C})$ and $Cov_{\mathrm{C|D}}E(\Delta|\mathbf{D},\mathbf{C})$ appearing in (3) are then calculated by averaging over the collection of $N_{e1}\times N_{e2}$ realizations. The corresponding scalar measure of the worth of additional data, $Tr\left[Cov_{\mathrm{C|D}}E(\Delta|\mathbf{D},\mathbf{C})\right]$, can be calculated as the trace of the resulting covariance matrix.
4. At the posterior stage, the actual measurement vector, $\mathbf{C}'$, becomes available. The ensemble mean $E(\mathbf{\Delta}|\mathbf{D},\mathbf{C}')$ and covariance $Cov(\mathbf{\Delta}|\mathbf{D},\mathbf{C}')$, jointly conditioned

on $\mathbf{D}$ and $\mathbf{C}'$, can be obtained via EnKF in a way similar to the procedure outlined in step 1. The scalar measure of the reference data-worth in light of the additional measurements is then calculated as $Tr\left[Cov\left(\Delta \mid \mathbf{D}\right)\right] - Tr\left[Cov\left(\Delta \mid \mathbf{D}, \mathbf{C}'\right)\right]$. This step (*a*) is employed to assess the quality of the preposterior estimates obtained at step 2, and (*b*) can be regarded as the prior analysis stage for a subsequent round of sequential data-worth assessment.

The above procedure shows that, for a given monitoring strategy, a total number of $N = N_{e1} \times N_{e2}$ numerical solutions of the system model are required to assess the data-worth in the preposterior analysis step. Since N_{e1} and N_{e2} need to be sufficiently large to ensure the accuracy of the estimated covariance matrix (*Li et al.*, 2014; *Zeng et al.*, 2010), we introduce the PCKF to address the challenge posed by the computational efficiency of the analysis.

2.2 Data-worth analysis via PCKF

The PCKF has been recently developed as an alternative to the traditional Monte Carlo based EnKF. It combines the polynomial chaos expansion (PCE) introduced by Wiener (1938) with EnKF.

According to Ghanem (1998), a target random variable can be expressed by a PCE with a set of deterministic coefficients. These coefficients are typically calculated through the Galerkin method (Ghanem, 1998; Mathelin et al., 2005). The latter entails computing the PCE coefficients upon solving a set of coupled equations. This somehow hampers a routine application of the method to groundwater flow and contaminant transport problems in which the relationship between the uncertain system parameters

326 and state variables is typically nonlinear. The probabilistic collocation method (PCM)
327 enables one to compute the PCE coefficients by solving a set of uncoupled equations at
328 the so-called collocation points in the parameter space (Tatang et al., 1997). A few key
329 details of PCE and PCM are reported in Appendix A for completeness. In a previous
330 study on the performance of PCKF (Zeng and Zhang 2010; Zeng et al., 2011), both the
331 parameters and state variables collected in the state vector (3) are approximated by the
332 PCE, and the coefficients of the PCE need to be subsequently updated for each system
333 realization at the selected collocation points as new data become available and are
334 assimilated in the model. When translated to preposterior data-worth analysis, this
335 approach would imply multiple reconstructions/updates of the PCE, consistent with
336 each realization of **C**, and would lead to unsustainable computational cost.

337 Here, we introduce a modified version of the PCKF to improve the computational
338 efficiency of the proposed data-worth analysis scheme. As we illustrate in the following,
339 our approach to PCKF relies on constructing the PCE-based model proxy only once by
340 (a) evaluating the PCE coefficients and (b) considering the (uncertain) random
341 quantities upon which the PCE-based model proxy relies as the quantities to be updated
342 during the data assimilation process (rather than the PCE coefficients). This approach
343 to PCKF enables us to employ the proxy model directly in the context of the required
344 repeated model inversions. Thus, it obviates the need for multiple evaluations of the
345 original (full system) model, which is critical to improve computational efficiency in
346 the preposterior data worth stage. The procedure requires having at our disposal as prior
347 information the key geostatistical descriptors of the parameter field, such as mean,

covariance structure and integral scale. If the prior information content is insufficient to this end, one can in principle resort to an iterative PCE-based bootstrap filter, along the lines proposed by Oladyshkin et al. (2013) to enhance the reliability of the proxy model. Given the demonstration-oriented nature of our test cases described in Section 3.2, we do not pursue an intensive study of the latter strategy.

At each time step, the random vector $\mathbf{\Delta}$ is characterized by a set of N_p (statistically independent) random variables (collected in vector). The following PCE approximation of $\mathbf{\Delta}$ can be constructed:

$$\hat{\mathbf{\Delta}} = \sum_{n \le d} a_n \Gamma_n \left(\xi_1, \xi_2, ..., \xi_{N_p} \right) \tag{12}$$

Here, $\Gamma_n \left(\xi_1, \xi_2, ..., \xi_{N_p} \right)$ is an *n*-th order multivariate orthogonal polynomial of the variables $\xi_1, \xi_2, ..., \xi_{N_p}$; a_n is a deterministic PCE coefficient; and *d* is the highest order of the expansion. At each time step, the coefficients in (12) can be obtained via the PCM. Before data are collected (that is, in the absence of conditioning on the future data), we compute the coefficients in (12) by running the system model for a simulation time encompassing all future time steps of interest and constructing a proxy for the whole simulation period. To do so, the PCM requires one to perform only a number:

$$M = \frac{\left(N_p + d \right)!}{N_p ! d !} \tag{13}$$

of runs of the full system model, which corresponds to the number of selected collocation points for ($\xi_1, \xi_2, ..., \xi_{N_p}$).

After the construction of the PCE surrogate model for the whole target simulation period, data assimilation is then applied to sequentially update on the constructed proxy the uncertain model parameters collected in vector $\boldsymbol{\xi}$. For each updated $\boldsymbol{\xi}$, the

corresponding entries of $\mathbf{\Lambda}$ can be obtained by sampling (12) rather than running the full model simulation, which renders data assimilation performed on the PCE proxy much more efficient than relying on the full model. The statistics of $\mathbf{\Lambda}$, i.e., mean and variance at various space and time points can then be evaluated by sampling a collection of realizations of ξ.

The selection of the order of PCE is a core step in the application of PCM. Note that: (*a*) increasing the order of the PCE through the data assimilation process might improve the accuracy of the surrogate model but can also require a considerably larger number of collocation points, at the expense of computational efficiency; and (*b*) the accuracy of the statistics obtained at a given PCE order cannot be assessed until different orders are actually implemented. As suggested by Dai et al. (2014), we assess the quality of the constructed proxy through a blind test. The latter is based on the comparison between the results obtained through the proxy and the original (full) system model with the same parameter sets, the latter being randomly selected in the parameter space at locations which differ from those of the collocation points.

The data-worth analysis based on the PCKF is essentially developed according to the workflow illustrated in Section 2.1. The only difference lies in the appearance of an additional step at the prior data-worth analysis stage, in which the coefficients of the PCE surrogate model are calculated through PCM, in addition to the evaluation of hypothetical additional data on the N_{e1} system realizations. The preposterior stage can then be implemented efficiently with the constructed surrogate model. As a result, the total number of full model runs to be performed in the data-worth analysis via the PCKF

is:

$$N = N_{e1} + M = N_{e1} + \frac{\left(N_p + d\right)!}{N_p!\,d!} \tag{14}$$

3. Illustrative Examples

3.1 Governing equations

A two-dimensional contaminant transport scenario taking place in a synthetic heterogeneous two-dimensional aquifer is considered to: (*a*) illustrate the feasibility of the proposed PCKF-based data-worth analysis; and (*b*) analyze key elements of alternative sampling strategies and conditions. We assume that conservative solute transport can be described by:

$$\frac{\partial\left(\theta C\right)}{\partial t} = \nabla\left(\theta \mathbf{D}_d \nabla C\right) - \nabla\left(\theta \mathbf{v} C\right) + q_s C_s \delta\left(\boldsymbol{x} - \boldsymbol{x}_0\right) \tag{15}$$

with appropriate initial and boundary conditions. In (15), θ is porosity; C is solute concentration (ML^{-3}); t is time (T); $\boldsymbol{x} = (x_1, x_2)$ is vector location in two-dimensional space (L); $\mathbf{v} = (v_1, v_2)$ (LT^{-1}) is seepage velocity vector; q_s (L^3T^{-1}) and C_s (ML^{-3}) respectively are volumetric flux and solute concentration of a point source positioned at $\boldsymbol{x}_0$; δ is the Dirac function; and $\mathbf{D}_d$ (L^2T^{-1}) is the dispersion tensor, which is defined as:

$$\begin{cases} \mathbf{D}_{d11} = \left(\alpha_L v_1^{\,2} + \alpha_T v_2^{\,2}\right)/\left|\boldsymbol{v}\right| + D_m \\ \mathbf{D}_{d11} = \left(\alpha_L v_1^{\,2} + \alpha_T v_2^{\,2}\right)/\left|\boldsymbol{v}\right| + D_m \\ \mathbf{D}_{d12} = \mathbf{D}_{d21} = \left(\alpha_L - \alpha_T\right) v_1 v_2 /\left|\boldsymbol{v}\right| \end{cases} \tag{16}$$

where D_m (L^2T^{-1}) is the diffusion coefficient; $\left|\boldsymbol{v}\right|$ (LT^{-1}) is magnitude of velocity; and α_L and α_T (L) are the longitudinal and transverse dispersivity, respectively.

In our study, we consider that solute migrates within a steady-state saturated flow,

described as:

$$\nabla\left(\mathbf{k}\frac{\rho}{\mu}\nabla h\right)=0 \tag{17}$$

$$\mathbf{v}=-\mathbf{k}\frac{\rho}{\theta\mu}\nabla h \tag{18}$$

subject to boundary conditions:

$$h(\boldsymbol{x})=H(\boldsymbol{x}),\quad \boldsymbol{x}\in\Gamma_D;\qquad -\mathbf{q}(\boldsymbol{x})\cdot\boldsymbol{n}(\boldsymbol{x})=Q(\boldsymbol{x}),\quad \boldsymbol{x}\in\Gamma_N \tag{19}$$

Here, $\mathbf{k}$ (L^2) is permeability tensor, which we assume to be isotropic; ρ and μ are fluid density and viscosity, respectively; $\mathbf{q}(\boldsymbol{x})=\mathbf{v}(\boldsymbol{x})/\theta$ is the Darcy flux vector at $\boldsymbol{x}$; $H(\boldsymbol{x})$ is prescribed head on Dirichlet boundary segment Γ_D; $Q(\boldsymbol{x})$ is prescribed flux across Neumann boundary segment Γ_N; and $\boldsymbol{n}(\boldsymbol{x})$ is the outward unit normal to the boundary $\Gamma_D\cup\Gamma_N$. Numerical solution of (15)-(19) is performed through a finite difference method.

3.2 Model setup

The key parameters of the problem are listed in Table 1, where all quantities are given in consistent length and time units. We consider an aquifer with a constant porosity $\theta = 0.15$, which is discretized along the horizontal plane into 40×40 grid cells of uniform size. Flow conditions are designed upon setting prescribed head boundary values of 12 and 7, respectively, on the left and right sides of the domain, while keeping impervious upper and bottom boundaries. The (natural) logarithm of permeability is considered to be a (second-order stationary) Gaussian (correlated) random field. In this study, we assume that the key statistical attributes of the random log permeability field are known. We take the covariance function of log permeability (Zhang, 2002) as:

$$Cov(\boldsymbol{x};\boldsymbol{y}) = \sigma^2 \exp\left[-\frac{|x_1 - y_1|}{\eta_1} - \frac{|x_2 - y_2|}{\eta_2}\right] \quad (20)$$

Here, $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$ are two spatial locations; σ^2 is log permeability variance, which is set to unity in our demonstration study (see Table 1); and η_1 and η_2 are the correlation lengths parallel to x_1 and x_2 directions, respectively, which we set as $\eta_1 = \eta_2 = 80$ (Table 1), thus resulting in a computational domain which spans 5 correlation scales along the x_1 and x_2 directions. We parameterize the random permeability field through the Karhunen-Loeve (KL) expansion:

$$\ln\left(k(\boldsymbol{x})\right) = \left\langle \ln(k) \right\rangle + \sum_{n=1}^{\infty} \sqrt{\lambda_n}\, f_n(\boldsymbol{x})\, \xi_n \quad (21)$$

where $\left\langle \ln(k) \right\rangle$ is the mean of log permeability, which is set to 5 in this study (Table 1); λ_n and $f_n(\boldsymbol{x})$ respectively are deterministic eigenvalues and eigenfunctions of the covariance function; and $\boldsymbol{\xi}$ is the vector of (zero-mean and unit variance) Gaussian random variables, ξ_n, which constitutes the set of uncertain model parameters that we consider. In practice, expansion (21) is often truncated up to the first N_p dominant terms.

The true (or reference) log permeability field is an unconditional random realization obtained via the geostatistical software library (GSLIB) (Deutsch and Journel, 1992) and is depicted in Figure 2. A point-wise contaminant source with a uniform (and constant in time) concentration $C_s = 250$ and $q_s = 100$ is placed at $\boldsymbol{x}_0 = (20, 200)$. Diverse arrays of monitoring wells are introduced in the system corresponding to the different scenarios that we investigate, as illustrated in the following. A uniform time step $\Delta T = 30$ is used for the transport simulations. Concentrations in the reference field are sampled at monitoring wells at equally spaced time intervals of duration ΔT. We set

454 the final observation time at T_{60} = 60 ΔT. The goal of the alternative
455 monitoring/sampling strategies that we analyze in our tests is to accurately predict the
456 spatial distribution of the contaminant plume at this final observation time.

457 We illustrate our approach on a set of test cases, listed in Table 2. Test case TC1 is
458 designed to validate the accuracy of the proposed PCKF through a comparison against
459 the results obtained with the standard EnKF method. Concentration values are sampled
460 at all sixteen monitoring wells depicted in Figure 2 and are used for sequential updates
461 of the uncertain model parameters. The observation time T_{60} is subdivided into 60
462 temporal intervals of uniform unit length ΔT.

463 Assimilation of concentration values sampled at the monitoring wells is performed
464 at the end of each time interval ΔT, starting at time T_0 = 0 and up to T_{60} = 60 ΔT. Prior
465 concentration data in all of our test cases are assumed to be associated with errors as
466 large as 10% of their observed values.

467 Test cases TC1-TC8 are designed to investigate the effect of diverse factors on
468 data-worth analysis. TC1 is set as the base case. It relies on prior data collected at all
469 wells for the first 24 assimilation times (i.e., up to time T_{24} = 24 ΔT). The worth of
470 additional data which are collected at the same wells with uniform (unit) frequency for
471 the remaining assimilation intervals (up to T_{48} = 48 ΔT) is then quantified both at the
472 preposterior and posterior steps.

473 Test cases TC2 and TC3 differ from TC1 in the sampling frequencies employed
474 for the data-worth analysis. In these cases, concentration samples are planned to be
475 collected at all 16 wells up to time T_{60} = 60 ΔT and every 3 ΔT (TC3) or 6 ΔT (TC4).

476 Test cases TC4, TC5, and TC6 consider augmenting the prior data set through
477 observations taken at diverse sets of wells pertaining to the monitoring network of TC1
478 (i.e., Wells in Column 1 and 2, Column 3 and 4, or Column 2 and 3, respectively for
479 TC4, TC5 and TC6). Similar to TC1, the prior dataset is formed by the concentrations
480 collected during the first 24 assimilation times. For each of these test cases, the worth
481 of additional measurements observed at the corresponding selected well network with
482 uniform (unit) frequency for the remaining assimilation intervals (up to $T_{60} = 60\ \Delta T$) is
483 then quantified both at the preposterior and posterior steps. Test cases TC7 and TC8
484 differ from TC1 in that they are designed to investigate the influence of the size of the
485 prior database on the worth of additional data collected at the same well locations up to
486 the target final time.

487 **3.3 Results and Discussion**

488 **3.3.1 Validation of the PCKF (Test case TC1)**

489 We start by considering test case TC1, which is devoted to the assessment of the
490 performance of the proposed PCKF by comparison against the standard EnKF. The
491 random log permeability is parameterized via the KL expansion (21). To obtain a
492 balance between computational accuracy and cost, we only retain the leading terms in
493 the KL expansion, i.e., these terms with the largest eigenvalues. Figure 3 depicts the
494 magnitude of the eigenvalues associated with the covariance matrix as a function of the
495 number of terms retained in (21). These results indicate a rapid decay of the magnitude
496 of the eigenvalues with increasing number of expansions terms. In this study, the first

30 terms are retained to parameterize the random log permeability field. This modeling choice enables us to account approximately for 70% of the energy of the target spatial random field and to achieve an appropriate balance between computational efficiency and accuracy of the overall procedure in the settings we analyze. A discussion on how to select the number of the retained terms can be found in Chang and Zhang (2009). We then select a second-order PCKF, rendering a number of $M = (30 + 2)! / (30!\ 2!) = 496$ collocation points at which the full system model is required to be evaluated for the construction of the PCE of the concentration, which is considered as the target system state. We select second-order PCKF for two reasons: (*a*) *Li and Zhang* (2007) conclude that PCEs of even order have a superior performance when compared against PCEs of the subsequent (odd) order when the unknown system parameters are Gaussian; and (*b*) employing a fourth-order PCKF would require a total of 46,376 full model evaluations to construct the PCE approximation, thus rendering this option unfeasible to pursue. Concentration measurements are generated upon solving the flow and transport problems on the reference permeability field and sampling at all the wells from time $T_0 = 0$ to $T_{60} = 60\ \Delta T$ with uniform sampling spacing equal to ΔT.

For the purpose of our comparison, both the PCKF and the EnKF are implemented with a number of realizations equal to 300. Figure 4 depicts the true concentration field, and the (ensemble) mean plume obtained via the PCKF and the EnKF at time T_{60}. These results suggest that the PCKF and the EnKF render mutually consistent predictions of the spatial distribution of the reference concentrations at the target time in our setting, even as some small differences can be noted between estimated and true distributions,

519 which can also be related to the lack of information associated with the region within
520 which the contaminant has not been observed. Figure 5 depicts the spatial distribution
521 of the predictive variance of concentrations obtained via the PCKF and the EnKF at
522 time T_{60}.

523 We note that the overall quality of the comparison might be influenced by issues
524 related to filter inbreeding, which can be associated with the limited number of
525 realizations that we employ. However, employing a considerably larger collection of
526 realizations would render unfeasible the application of the standard EnKF from a
527 computational standpoint. For example, the estimated CPU time of the implementation
528 of EnKF with 3,000 realizations is 78,300 s, i.e., 21.75 h. With this in mind, and noting
529 that repeated runs of the full system model are required at the preposterior stage, we
530 consider that the result that we obtain can imbue us with a relatively high confidence in
531 our selection of the reduced PCKF approach.

532 **3.3.2 Effect of the sampling frequency of planned future measurements (Test**
533 **cases TC1, TC2, and TC3)**

534 Sampling frequency is a critical factor in a monitoring strategy design. While high
535 sampling frequency can provide an increased flux of data that may help to reduce
536 predictive uncertainty, it will also yield an increased workload and financial demands.
537 The purpose of the comparison of TC1, TC2, and TC3 is to identify the relative worth
538 of diverse sampling frequencies. In these cases, the prior data set **D** comprises the
539 measurements taken at the monitoring wells in Column 1 to 4 during the first 24

assimilation times (i.e., from time $T_1 = \Delta T$ up to $T_{24} = 24\ \Delta T$). Three hundred realizations of the additional data set are used at the prior stage of the analysis, while the preposterior analysis is performed through the PCKF due to its notable efficiency.

As stated above, Figure 5a depicts the spatial distribution at time T_{60} of the predictive variance of the contaminant concentration, conditional only on **D**, i.e., $Var(\Delta \mid \mathbf{D})$. We note that the predictive uncertainty is larger at the front of the plume than at locations in the vicinity of the source. This result is consistent with the high spatial concentration gradients in the proximity of the plume front (*Rubin*, 1991).

Figure 6 depicts comparisons between concentration values observed at four wells located at (5, 5), (15, 15), (25, 15) and (35, 25) at times $T_i = i\ \Delta T$ (i = 25, 26, 27, …, 48) and corresponding estimates of mean concentrations computed on the basis of the prior **D** values. Intervals of width corresponding to ± 1 standard deviations about the mean are depicted to complete the picture.

These results show that, even as in some cases the true concentration values fall within the limits of the considered uncertainty bounds, predicted values tend to systematically deviate from their true counterparts (albeit with different degrees of discrepancy, depending on the monitoring well location) with increasing time, as expected. We recall that estimates of mean concentration together with the associated predictive variance constitute the bases for the construction of the additional data sets **C** to be employed in the preposterior analysis step for each of these test cases. A posterior analysis is then performed after $\mathbf{C}'$ becomes available.

561 Data-worth analysis via the PCKF is applied to each test case to assess the worth
562 of the additional data sets **C** which are built as illustrated in Section 3.2. Figures 7a, c,
563 and e depict the spatial distribution of the expected concentration variance reduction
564 $Var_{\mathbf{C}|\mathbf{D}}E\left(\Delta|\mathbf{D},\mathbf{C}\right)$, respectively for TC1, TC2, and TC3, calculated at the preposterior
565 stage. Figures 7b, d, and f depict the spatial distribution of concentration variance
566 reductions, $Var\left(\Delta|\mathbf{D}\right)-Var\left(\Delta|\mathbf{D},\mathbf{C}^{'}\right)$, calculated at the posterior stage. It can be
567 noted that the spatial distributions of the expected variance reductions resemble closely
568 their reference (posterior) counterparts. The scalar measures of the expected and
569 reference data-worth are compared in Figure 8. For ease of reference, these values are
570 also listed in Table 3 together with the corresponding results associated with all test
571 cases examined. From these results, we conclude that the expected data-worth can
572 estimate its reference counterpart with high accuracy. The preposterior data-worth
573 metric $Tr\left[Cov_{\mathbf{C}|\mathbf{D}}E\left(\Delta|\mathbf{D},\mathbf{C}\right)\right]$ of TC1 is 215,742 and is larger than that associated
574 with TC2 and TC3, these being equal to 125,292 and 200,672, respectively. These
575 results support the idea that an increased sampling frequency should lead to uncertainty
576 reduction. However, our results also reveal that the rate of increase of the preposterior
577 data-worth somehow decreases with increasing sampling frequency. This observation,
578 in turn, implies that the effect of an increased sampling frequency becomes less
579 significant when the frequency has reached a certain threshold level. The latter should
580 depend on the spatial location of the planned monitoring well, as well as on the
581 underlying aquifer heterogeneity and transport setting.

582 A remarkable advantage of our PCKF-based data-worth analysis is the

583 computational efficiency. Each of the test cases here analyzed requires only 796
584 forward runs of the full system model. These include: (*a*) 300 simulations for the
585 estimation of the statistics (mean and covariance) of the additional data set to be
586 included in **C**; and (*b*) 496 simulations to construct the surrogate model. The CPU time
587 for the data-worth analysis with PCKF for TC1, TC2, and TC3 respectively are 9,884
588 s, 3,884 s, and 5,084 s. Otherwise, employing EnKF based on a collection of $N_{e1} = N_{e2}$
589 = 300 MC realizations for data-worth analysis requires a number of full model solutions
590 equal to 90,000 for each assimilation step, with estimated CPU times of 1,879,200 s,
591 1,096,200 s, and 1,252,800 s, respectively for TC1, TC2, and TC3.

592 **3.3.3 Effect of the monitoring network location (Test cases TC4, TC5, and TC6)**

593 Test cases TC4, TC5, and TC6 are identical to TC1 except that additional sampling
594 after collection of the prior data set (i.e., the data set composed by observations
595 collected for the first 24 time intervals) is performed at locations listed in Table 2 and
596 identified in Figure 2. The additional data set **C** is then formed by the data observed at
597 the wells in Column 1 and 2, Column 3 and 4, and Column 2 and 3 for TC4, TC5 and
598 TC6, respectively, from assimilation times $T_{25} = 25\ \Delta T$ to $T_{48} = 48\ \Delta T$ with sampling
599 frequency equal to ΔT. Figure 9 depicts the spatial distributions of the expected
600 preposterior concentration variance reductions (Figure 9a, c, and e) and the reference
601 posterior concentration variance reduction (Figures 9b, d, and f) at the latest time T_{60}.
602 Only minute deviations are observed between the patterns of preposterior estimations
603 and their posterior counterparts. Scalar measures of preposterior and posterior data-
604 worth are also depicted in Figure 10. As compared to TC1 (for which

605 $Tr\left[Cov_{\mathbf{C}|\mathbf{D}}E\left(\Delta\,|\,\mathbf{D},\mathbf{C}\right)\right]$ = 215,742), reducing the monitoring well network brings in an
606 overall decrease of the predictive concentration variance reduction, as expected, the
607 scalar measures of preposterior data-worth for TC4, TC5, and TC6 respectively being
608 126,698, 190,141, and 207,585. These results suggest that acquisition of additional data
609 from monitoring wells in Column 2 and 3 has the largest potential for providing
610 valuable information in the setting we analyze. This is consistent with the observation
611 that wells in Columns 2 and 3, and particularly those in Column 3, are located in a
612 region where the prior predictive uncertainty is considerably high, as opposed to wells
613 in Column 1 and 4, which are placed at locations with smaller predictive uncertainties.
614 Even as a generalization of these results is hard to propose and the specific choices
615 associated with the selection of new monitoring wells can be associated with additional
616 constraints of diverse nature (including, e.g., political and socio-economical elements),
617 our results support the idea that an effective choice would be to acquire new data at
618 wells located within a region where the prior predictive uncertainty is relatively large.

619 **3.3.4 Effect of prior data content (Test cases TC1, TC7, and TC8)**

620 Prior data content is a critical element that may impact the results of data-worth
621 analysis. In this sense, TC7 and TC8 are identical to TC1 except that the prior data
622 vector **D** is formed by diverse amounts of sampled data. In TC7 we collect prior data
623 from time T_1 to time $T_{12} = 12\,\Delta T$, while TC8 considers prior observations sampled from
624 time T_1 to time $T_{18} = 18\,\Delta T$. The spatial distributions of prior predictive concentration
625 variance at time T_{60} are depicted in Figure 11a and b, respectively for TC7 and TC8. As
626 expected, the prior predictive concentration variances of TC7 and TC8 are generally

627 larger than that of TC1 because of the diminished prior information content. Figures
628 12a and b, and Figures 12c and d depict the comparison of spatial patterns of the
629 preposterior and reference posterior predictive uncertainty reductions for TC7 and TC8,
630 respectively. The preposterior patterns approximate their posterior counterparts with a
631 sufficiently high quality in both cases. Figure 13 depicts the scalar measures of
632 (preposterior and posterior/reference) data-worth for these test cases. The scalar
633 measures of preposterior data-worth are 530,745 and 330,082, respectively for TC7 and
634 TC8, to be compared against the value of 215,742 for TC1.

635 As expected, a reduced number of prior data yields an increased benefit arising
636 from collecting additional data. Due to the limited amount of prior information,
637 additional data appear to be associated with a relatively high value.

638 ## 4. Conclusions

639 Our work leads to the following major conclusions.

640 1. Integrating a Probabilistic Collocation Method within EnKF allows assessing data-
641 worth through a surrogate of the full system model. This leads to a remarkable
642 improvement of the computational efficiency of the data-worth procedure. This is
643 particularly relevant considering that, even as incorporating the EnKF into data-
644 worth analysis frameworks enables one to quantify the worth of additional data in
645 a sequential way, routine applications are still challenging due to large
646 computational costs involved.
647 2. We illustrate our approach on a suite of synthetic scenarios involving conservative
648 solute migration in a two-dimensional random permeability field. Our test cases are

designed to: (a) test the accuracy of our approach when compared to an EnKF data-worth analysis based on Monte Carlo solutions of the full system model (TC1); and (b) investigate the effect of diverse factors on data-worth analysis (TC1-TC8), including design components of an envisioned monitoring scheme, such as spatial location of monitoring wells, temporal sampling frequency, and prior data content for solute concentrations. When compared against results obtained via the full system model, our Probabilistic Collocation EnKF (PCKF) renders significantly accurate results at a considerably reduced computational time in the setting considered, in terms of predicted concentration values and associated uncertainty quantification.

3. In our examples (TC1, TC2, and TC3), augmenting the temporal sampling frequency associated with planned acquisitions of additional concentration data leads to increased benefits in terms of uncertainty reduction. The effects of the increased sampling frequency tend to become less significant when the assimilation frequency has reached a certain threshold level.
4. Foreseeing the design of the location of a monitoring well within a region where prior predictive uncertainty is highest is more likely to provide valuable data (TC4, TC5, and TC6). Our examples (TC1, TC7, and TC8) suggest that the value of additional data is highest when these are supplemented to the smallest sets of prior data. The scalar measure of data-worth shows a sharp decrease with the increase of the number of measurements forming the prior dataset.

Appendix A

The PCE approximation $\hat{\mathbf{\Delta}}$ of the random vector $\mathbf{\Delta}$ that is characterized by a set of N_p (statistically independent) random variables (collected in vector $\boldsymbol{\xi}$) relies on the spectral expansion:

$$\hat{\mathbf{\Delta}} = \sum_{n \le d} a_n \Gamma_n \left(\xi_1, \xi_2, ..., \xi_{N_p} \right) \quad \text{(A. 1)}$$

Here, d is the highest order of the expansion; a_n are deterministic coefficients; and $\Gamma_n \left(\xi_1, \xi_2, ..., \xi_{N_p} \right)$ are the n-th order multi-dimensional orthogonal polynomials in terms of random variables $\xi_1, \xi_2, ..., \xi_{N_p}$ (collected in vector $\boldsymbol{\xi}$), which constitute the input to the system model. If these uncertain parameters are Gaussian, the Hermite polynomials form the appropriate orthogonal basis for Gaussian random variables (Xiu and Karniadakis, 2002).

The number of unknown coefficients in (A.1) is N_c:

$$N_c = \frac{\left(N_p + d \right)!}{N_p ! d !} \quad \text{(A. 2)}$$

Hence, N_c linear independent equations are required to solve the coefficients.

Calculation of the PCE coefficients can be accomplished through the Probabilistic collocation method (PCM). Let us define the residual R between the real output $\mathbf{\Delta}$ and its approximation $\hat{\mathbf{\Delta}}$ as:

$$R\left(\{a_i\}, \boldsymbol{\xi} \right) = \hat{\mathbf{\Delta}} - \mathbf{\Delta} \quad \text{(A. 3)}$$

where $\{a_i\}$ is the set of the PCE coefficients.

The residual (A. 3) should satisfy the following integral equation:

691 $$\int R\left(\{a_i\},\xi\right)\delta\left(\xi-\xi_j\right)P\left(\xi\right)d\xi = 0 \qquad \text{(A. 4)}$$

692 where δ is the Dirac delta function; and ξ_j is a sub-set of ξ whose entries are the
693 collocation points. Equation (A.4) results in a set of independent equations whose
694 solution yields the coefficients of the PCE.

695 Selection of the collocation points is the key issue of PCM. *Li and Zhang* (2007)
696 suggest that the collocation points can be selected from the roots of the next higher
697 order orthogonal polynomial for each uncertain parameter. Once the coefficients of the
698 polynomial chaos expansions are obtained, (A.1) can be employed as a proxy for the
699 original full system model. The statistics of Δ can be evaluated by the PCE
700 coefficients (Li and Zhang, 2007; Li et al., 2009) or by sampling on the surrogate model
701 (A.1) (Li et al., 2011 and Dai et al., 2014) with a group realizations of random variables.
702

References

Aanonsen, S. I., Naevdal , G., Oliver, D. S. and Reynolds, A. C., 2009, The ensemble Kalman filter in reservoir engineering--a review, SPE J., 14(3), 393-412, doi: 10.2118/117274-PA.

Abbaspour, K. C., Schulin, R., Schläppi, E. and Flühler, H. 1996, A bayesian approach for incorporating uncertainty and data worth in environmental projects, Environ. Model Assess., 1(3), 151-158, doi:10.1007/BF01874902.

Alcolea, A., Carrera, J. and Median, A. 2006, Pilot points method incorporating prior information for solving the groundwater flow inverse problem, Adv. Water Resour., 29, 1678-1689, doi:10.1016/j.advwatres.2005.12.009.

Amudo, C., Graf, T., Harris, N. R., Dandekar, R., Amor, F. B. and May, R. S., 2006, Experimental design and response surface models as a basis for stochastic history match - a Niger delta experience, paper presented at International Petroleum Technology Conference, Kuala Lumpur, doi:10.2523/12665-MS.

Back, P. E., 2007, A model for estimation the value of sampling programs and the optimal number of samples of contaminated soil, Environ. Geol., 52, 573-85, doi:10.1007/s00254-0488-6.

Carrero, E, Queipo, N. V., Pintos, S. and Zerpa, L. E., 2007, Global sensitivity analysis of Alkali–Surfactant–Polymer enhanced oil recovery processes, J. Petrol Sci. & Eng., 58(1), 30-42, doi:10.1016/j.petrol.2006.11.007.

Chang, H., and Zhang, D., 2009, A comparative study of stochastic collocation methods

724 for flow in spatially correlated random fields. Commun. Comput. Phys., 6(3), 509-
725 535.

726 Chang, H. and Zhang, D., 2014, History matching of statistically anisotropic fields
727 using the Karhunen-Loeve expansion-based global parameterization technique,
728 Computat. Geosci., 18(2), 265-282, doi:10.1007/s10596-014-9409-z.

729 Chen, Y. and Zhang, D., 2006, Data assimilation for transient flow in geologic
730 formations via ensemble Kalman filter, Adv. Water Resour., 29(8), 1107-1122,
731 doi:10.1016/j.advwatres.2005.09.007.

732 Ciriello, V., Federico Di V., Riva, M., Cadini, F., Sanctis De J., Zio, E. and Guadagnini,
733 A., 2013, Polynomial chaos expansion for global sensitivity analysis applied to a
734 model of radionuclide migration in randomly heterogeneous aquifer, Stoch. Env.
735 Res. Risk A., 27(4), 945-954, doi: 10.1007/s00477-012-0616-7.

736 Ciriello, V., Guadagnini, A., Federico, Di V., Edery, Y. and Berkowitz, B., 2013,
737 Comparative analysis of formulations for conservative transport in porous media
738 through sensitivity-based parameter calibration, Water Resour. Res., 49(9), 5206-
739 5220, doi:10.1002/wrcr.20395.

740 Dai, C., Li, H. and Zhang, D., 2013, Efficient and accurate global sensitivity analysis
741 for reservoir simulation by use of probabilistic collocation method, SPE J., 19(4),
742 621-635, doi:10.2118/167609-PA.

743 Dausman, A. M., Doherty, J., Langevin, C.D., and Sukop, M. C., 2010, Quantifying
744 data worth toward reducing predictive uncertainty, Ground Water, 48(5),729-740 ,

745 doi: 10.1111/j.1745-6584.2010.00679.x.

746 Deutsch, C.V. and Journal, A.G., 1992, GSLIB: Geostatistical software library and
747 user's guide, 340 pp., Oxford Univ. Press, New York.

748 Formaggia, L., Guadagnini, A., Imperiali, I., Lever, V., Porta, G., Riva, M. and
749 Tamellini, L., 2013, Global sensitivity analysis through polynomial chaos
750 expansion of a basin-scale geochemical compaction model, Comp. Geosci., 17(1),
751 25-42, doi: 10.1007/s10596-012-9311-5.

752 Foster, S. S. D. and Chiltion, P. J., 2003, Groundwater: The Processes and Global
753 Significance of Aquifer Degradation, Philosophical Transactions of the Royal
754 Society of London Series B: Biological Sciences, 358(1440), 1957-1972,
755 doi:10.1098/rstb.2003.1380.

756 Fu, J. and Gomez-Hernandez, J. J., 2009, A blocking Markov Chain Monte Carlo
757 method for inverse stochastic hydrogeological modeling, Math. Geosci., 41, 105-
758 128, doi: 10.1007/s11004-008-9206-0.

759 Ghanem, R., 1998, Scales of fluctuation and the propagation of uncertainty in random
760 porous media, Water Resour. Res., 34(9): 2123-2136, doi:10.1029/98WR01573.

761 Ghanem, R. and Spanos, P. D., 2003, Stochastic finite elements: A spectral approach.
762 Dover Publications, New York.

763 Hays, J., Sandu, A. C. Sandu, and Hong, D., 2011, Parametric design optimization of
764 uncertain ordinary differential equation systems. J Mech Design, 134,181003,

765 doi:10.1115/1.4006950.

766 He J., Sarma P. and Durlofsky L., 2009, Use of Reduced-order Models for Improved
767 Data Assimilation within an EnKF Contex, Spe Reservoir Simulation Symposium,
768 2011, 50(50):1762-1763.

769 Hendricks-Franssen, H. J., Alcolea, A., Riva, M., Bakr, M., Van der Wiel, Stauffer, N.
770 and Guadagnini, A., 2009, A comparison of seven methods for the inverse
771 modelling of groundwater flow: Application to the characterization of well
772 catchments, Adv. Water Resour., 32(6), 851-872, doi:
773 10.1016/j.advwatres.2009.02.011.

774 Hernandez, A. F., Neuman, S. P., Guadagnini, A. and Carrera, J., 2003, Conditioning
775 mean steady state flow on hydraulic head and conductivity through geostatistical
776 inversion, Stoch. Environ. Res. Risk A., 17(5), 329-338, doi:10.1007/s00477-003-
777 0154-4.

778 Hernandez, A. F., Neuman, S. P., Guadagnini, A. and Carrera, J., 2006, Inverse
779 stochastic moment analysis of steady state flow in randomly heterogeneous media,
780 Water Resour. Res., 42, W05425, doi:10.1029/2005WR004449.

781 James, B. R. and Freeze, R. A., 1993, The worth of data in predicting aquitard
782 continuity in hydrogeologic design, Water Resour. Res., 29(7), 2049-2065,
783 doi:10.1029/93WR00547.

784 Jin B.,2009, Fast Bayesian approach for parameter estimation. International Journal for
785 Numerical Methods in Engineering, 76(2):230–252.

786 Kollat, J. B., Reed, P. M. and Maxwell, R. M., 2011, Many-objective groundwater
787 monitoring network design using bias-aware ensemble Kalman filtering,
788 evolutionary optimization, and visual analytics, Water Resour. Res,
789 47(2),W02539, doi:10.1029/2010WR009194.

790 Leube, P. C., Geiges, A. and Nowak, W., 2012, Bayesian assessment of the expected
791 data impact on prediction confidence in optimal sampling design, Water Resour.
792 Res., 48(2), doi:10.1029/2010WR010137.

793 Li, H., Sarma, P. and Zhang, D., 2011, A comparative study of the probabilistic-
794 collocation and experimental-design methods for petroleum-reservoir uncertainty
795 quantification, SPE J., 16(2), 429-439, doi:10.2118/140738-PA.

796 Li, H. and Zhang, D., 2007, Probabilistic collocation method for flow in porous media:
797 Comparisons with other stochastic methods. Water Resour. Res., 43(9), W09409 ,
798 doi:10.1029/2006WR005673.

799 Li, H. and Zhang, D., 2009, Efficient and accurate quantification of uncertainty for
800 multiphase flow with the probabilistic collocation method. SPE J., 14(4), 665-679,
801 doi:10.2118/114802-PA.

802 Li, J. and Xiu, D., 2009, A generalized polynomial chaos based ensemble Kalman filter
803 with high accuracy. J. Comput. Phys., 228(15), 5454-5469,
804 doi:10.1016/j.jcp.2009.04.029.

805 Li, L., Zhou, H., Gomez-Hernandez, J. J. and Franssen, H. H., 2012, Jointly mapping
806 hydraulic conductivity and porosity by assimilating concentration data via

807 ensemble Kalman filter, J. Hydrology, 428-429, 152-169,
808 doi:10.1016/j.jhydrol.2012.01.037.

809 Li, W., Lu, Z. and Zhang, D., 2009, Stochastic analysis of unsaturated flow with
810 probabilistic collocation method, Water Resour. Res., 45, W08425,
811 doi:10.1029/2008WR007530.

812 Li, W., Lin, G. and Zhang, D., 2014, An adaptive ANOVA-based PCKF for high-
813 dimensional nonlinear inverse modeling, J. Comput. Phys., 258, 752-772,
814 doi:10.1016/j.jcp.2013.11.019.

815 Li, Y., Weerts, M. Clark, Hendricks-Franssen, H. J., Kumar, S., Moradkhani, H. and
816 Restrepo, P., 2012, Advancing data assimilation in operational hydrologic
817 forecasting: Progresses, challenges, and emerging opportunities, Hydrol. Earth
818 Syst. Sci., 16, 3863-3887, doi:10.5194/hess-16-3863-2012.

819 Liu, G., Chen, Y. and Zhang, D., 2008, Investigation of flow and transport process at
820 the MADE site Using ensemble Kalman filter, Adv. Water Resour., 31, 975-986,
821 doi:10.1016/j.advwatres.2008.03.006.

822 Marzouk, Y., Najm, N. and Rahn, L. A., 2007, Stochastic spectral methods for efficient
823 Bayesian solution of inverse problems. J. Comp. Phy., 224(2):560-586. doi:
824 10.1016/j.jcp.2006.10.010

825 Marzouk, Y. and Xiu, D., 2009, A stochastic collocation approach to Bayesian
826 inference in inverse problems. Commu. Comput. Phys., 6(4), 826-847.

827 Mathelin, L., Hussaini, M. Y. and Zang, T. A., 2005, Stochastic approaches to
828 uncertainty quantification in CFD simulations. Numer. Algorithms, 38 (1-3), 209-
829 236, doi:10.1007/BF02810624.

830 Nowak, W., Barros de F. P. J. and Rubin, Y., 2010, Bayesian geostatistical design:
831 Task-driven site investigation when the geostatistical model is uncertain, Water
832 Resour. Res., 46, W035535, doi:10.1029/2009WR008312.

833 Nowak, W., Rubin, Y. and Barros de F. P. J., 2012, A hypothesis-driven approach to
834 optimize field campaigns, Water Resour. Res., 48, W06509,
835 doi:10.1002/wrcr.20113.

836 Neuman, S. P., Xue, L., Ye, M. and Lu, D., 2011, Bayesian analysis of data-worth
837 considering model and parameter uncertainties, Adv. Water Resour., 36:75-85,
838 doi:10.1016/j.advwatres.2011.02.007.

839 Oladyshkin, S., Barros de F. P. J., Nowak, W., 2012a, Global sensitivity analysis: A
840 flexible and efficient framework with an example from stochastic hydrogeology.
841 Adv. Water Resour., 37, 10-22, doi:10.1016/j.advwatres.2011.11.001.

842 Oladyshkin, S. and Nowak, W., 2012b, Data-driven uncertainty quantification using
843 the arbitrary polynomial chaos expansion, Reliab. Eng. Syst. Safe, 106, 179-190,
844 doi:10.1016/j.ress.2012.05.002.

845 Oladyshkin, S., Class, H., and Nowak, W., 2013, Bayesian updating via bootstrap
846 filtering combined with data-driven polynomial chaos expansions: methodology
847 and application to history matching for carbon dioxide storage in geological

formations. Comput. Geosci., 17(4), 671-687.

Oliver, D. S., Reynolds, A.C. and Liu, N., 2008, Inverse theory for petroleum reservoir characterization and history matching, 380 pp., Cambridge University Press, Cambridge.

Panzeri, M, Riva, M. and Guadagnini, A., 2013, Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow, Water Resour. Res., 49(3), 1334-1344, doi:10.1002/wrcr.20113.

Panzeri, M., Riva, M., Guadagnini, A. and Neuman, S. P., 2014, Comparison of ensemble Kalman filter groundwater-data assimilation methods based on stochastic moment equations and Monte Carlo simulation, Adv. Water Resour., 66, 8-18, doi:10.1016/j.advwatres.2014.01.007.

Rada E. S. and Schultz, G. A., 1998, Worth of hydrological data in water resources projects. Application: Bolivians Amazones zone. Water Management of the Amazon Basin.

Reagan M. T., Najm, H., Pebay, P., Knio, O. and Ghanem, R., 2005, Quantifying uncertainty in chemical systems modeling. Int. J. Chem. Kinet., 37 (6), 368-382, doi:10.1002/kin.20081.

Riva, M., Guadagnini, A., Neuman, S. P., Janetti, E. B. and Malama, B., 2009, Inverse analysis of stochastic moment equations for transient flow in randomly heterogeneous media, Adv. Water Resour., 32(10), 1495-1507,

869 doi:10.1016/j.advwatres.2009.07.003.

870 Riva, M., Panzeri, M., Guadagnini, A. and Neuman, S. P., 2011, Role of model
871 selection criteria in geostatistical inverse estimation of statistical data- and model-
872 parameters, Water Resour. Res., 47(7), doi:10.1029/2011WR010480.

873 Rubin, Y., 1991, Transport in heterogeneous porous media: Prediction and uncertainty,
874 Water Resour. Res., 27(7), 1723-1738.

875 Rubin, Y., Chen, X., Murakami, H. and Hahn, M., 2010, A Bayesian approach for
876 inverse modeling, data assimilation, and conditional simulation of spatial random
877 fields, Water Resour. Res., 25(3), 351-62, doi:10.1029/2009WR008799.

878 Russell, K.T. and Rabideau, A., 2006, Decision analysis for Pump‐and‐Treat Design.
879 Ground Water Monit. R., 20 (3), 159-168, doi:10.1111/j.1745-
880 6592.2000.tb00281.x.

881 Saad G. and Ghanem R.,2009, Characterization of reservoir simulation models using a
882 polynomial chaos-based ensemble Kalman filter, Water Resour. Res., 45(45):546-
883 550, doi:10.1029/2008WR007148

884 Schaaf, T., Coureaud, B. and Labat, N., 2006, Using experimental designs, assisted
885 history matching tools and Bayesian framework to get probabilistic production
886 forecasts, present at Europec/EAGE Conference and Exhibition, Rome,
887 doi:10.2118/113498-MS.

888 Sudret, B., 2008, Global sensitivity analysis using polynomial chaos expansions. Reliab.

889 Eng. Syst. Safe, 93 (7), 964-979, doi:10.1016/j.ress.2007.04.002.

890 Tatang, M. A., Pan, W., Prinn, R. G. and Mcrae, G. J., 1997, An efficient method for
891 parametric uncertainty analysis of numerical geophysical models, J. Geophys.
892 Res., 102 (D8), 21925-21932, doi:10.1029/97JD01654.

893 Wiener, N., 1938, The homogeneous chaos. Am J Math 60 (4): 897-936.

894 Wu, B., Zheng, Y., Tian, Y., Wu, X., Yao, Y., Han, F. and Zheng, C., 2014, Systematic
895 assessment of the uncertainty in integrated surface water‐groundwater modeling
896 based on the probabilistic collocation method. Water Resour. Res., 50(7), 5848-
897 5865, doi:10.1002/2014WR015366.

898 Xue, L., Zhang, D., Guadagnini, A. and Neuman, S. P., 2014, Multimodel bayesian
899 analysis of groundwater data worth, Water Resour. Res.,
900 doi:10.1002/2014WR015503.

901 Xue, L. and Zhang, D., 2014, A multimodel data assimilation framework via the
902 ensemble Kalman filter, Water Resour Res., 50(5), 4197-4219,
903 doi:10.1002/2013WR014525.

904 Xie, X. and Zhang, D., 2010, Data assimilation for distributed hydrological catchment
905 modeling via ensemble Kalman filter, Adv. Water Resour.,22(6), 678-690,
906 doi:10.1016/j.advwatres.2010.03.012.

907 Xiu, D. and Karniadakis, G., 2002, Modeling uncertainty in steady state diffusion
908 problems via generalized polynomial chaos, Comput. Method. Appl. M., 191 (43),

909 4927-4948, doi:10.1016/S0045-7825(02)00421-8.

910 Zeng, L. and Zhang, D., 2010, A stochastic collocation based Kalman filter for data
911 assimilation, Comp. Geosci., 14 (4), 721-744, doi:10.1007/s10596-010-9183-5.

912 Zeng, L., Chang, H. and Zhang, D., 2011, A probabilistic collocation-based Kalman
913 filter for history matching. SPE J., 16 (2), 294-306, doi:10.2118/140737-PA.

914 Zeng, L., Shi, L., Zhang, D. and Wu, L., 2012, A sparse grid based Bayesian method
915 for contaminant source identification, Adv. Water Resour., 37: 1-9,
916 doi:10.1016/j.advwatres.2011.09.011.

917 Zhang, D., 2002, Stochastic Methods for Flow in Porous Media: Coping with
918 Uncertainties, 350pp, Academic Press, San Diego, Califorina

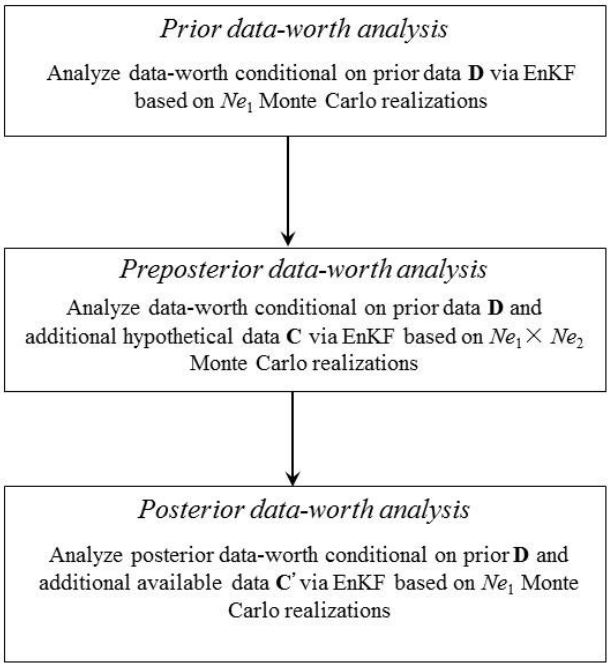
919 Zhang, G., Lu, D., Ye, M., Gunzburger, M. and Webster, C., 2013, An adaptive sparse-
920 grid high-order stochastic collocation method for Bayesian inference in
921 groundwater reactive transport modeling. Water Resour. Res., 49(10), 6871-6892,
922 doi:10.1002/wrcr.20467.

923 Zhang, J., Zeng, L., Chen, C., Chen, D. and Wu, L. 2015, Efficient Bayesian
924 experimental design for contaminant source identification, Water Resour. Res.,
925 51(1), 576-598, doi: 10.1002/2014WR015740

926 Zhou, H., Gomez-Hernadez, J. J. and Li, L., 2014, Inverse methods in hydrogeology:
927 Evolution and recent trends. Adv. Water. Resour., 63, 22-37,
928 doi:10.1016/j.advwatres.2013.10.014.

929 Zimmerman, D. A., Marsily, G. D., Gotway, C. A., Marietta, M. G., Axness, C. L.,
930 Beauheim, R. L. and Rubin, Y., 1998, A comparison of seven geostatistically
931 based inverse approaches to estimate transmissivities for modeling advective
932 transport by groundwater flow, Water Resour. Res., 34(6), 1373-1413,
933 doi:10.1029/98WR00003.
934

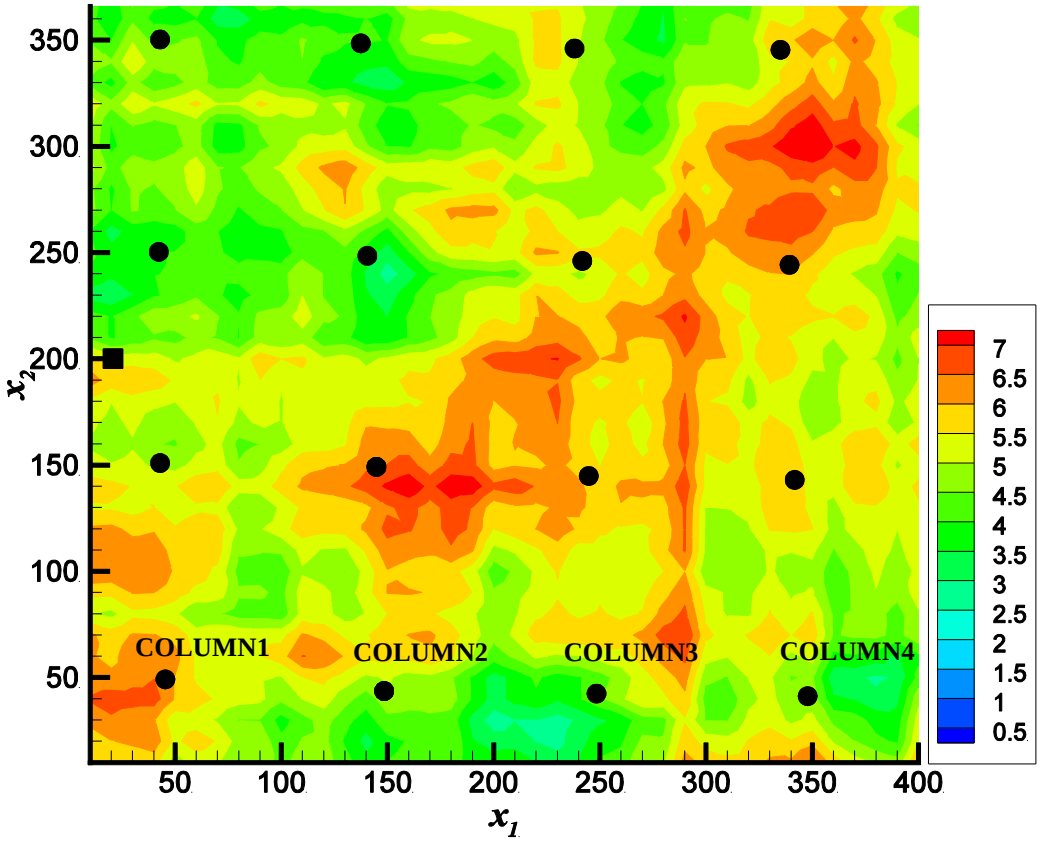
935



936

937 Figure 1. Workflow of data-worth analysis for a dynamic system.

938



939

940 Figure 2. The true (or reference) log permeability field. Solid black square indicates
941 the contaminant source; solid circles denote monitoring wells employed in the test
942 cases examined.

943

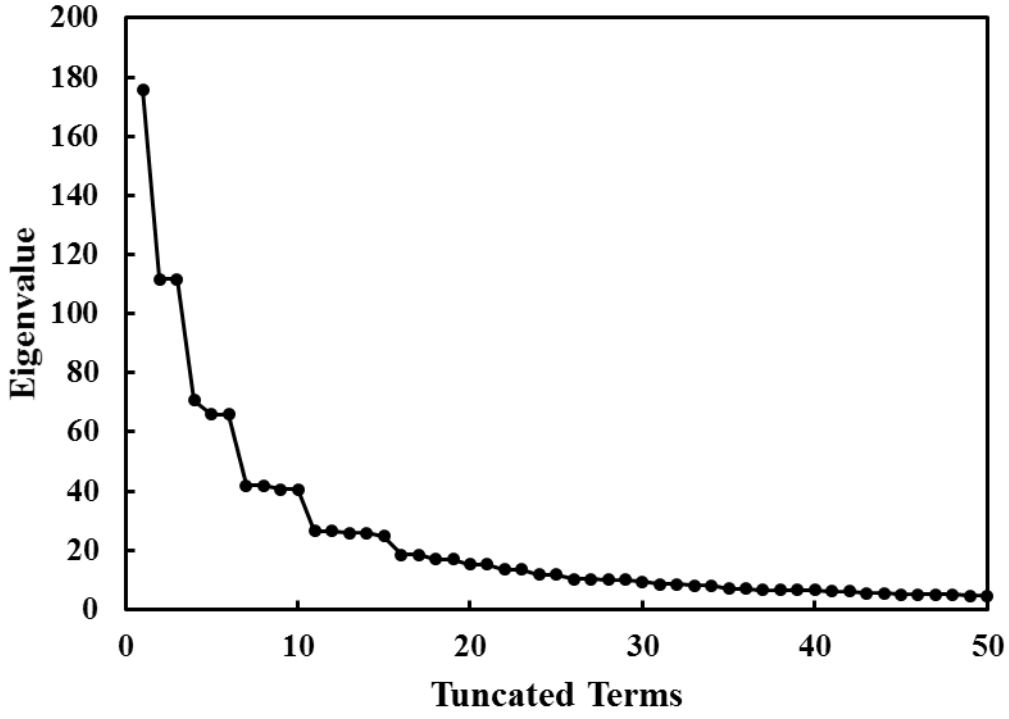
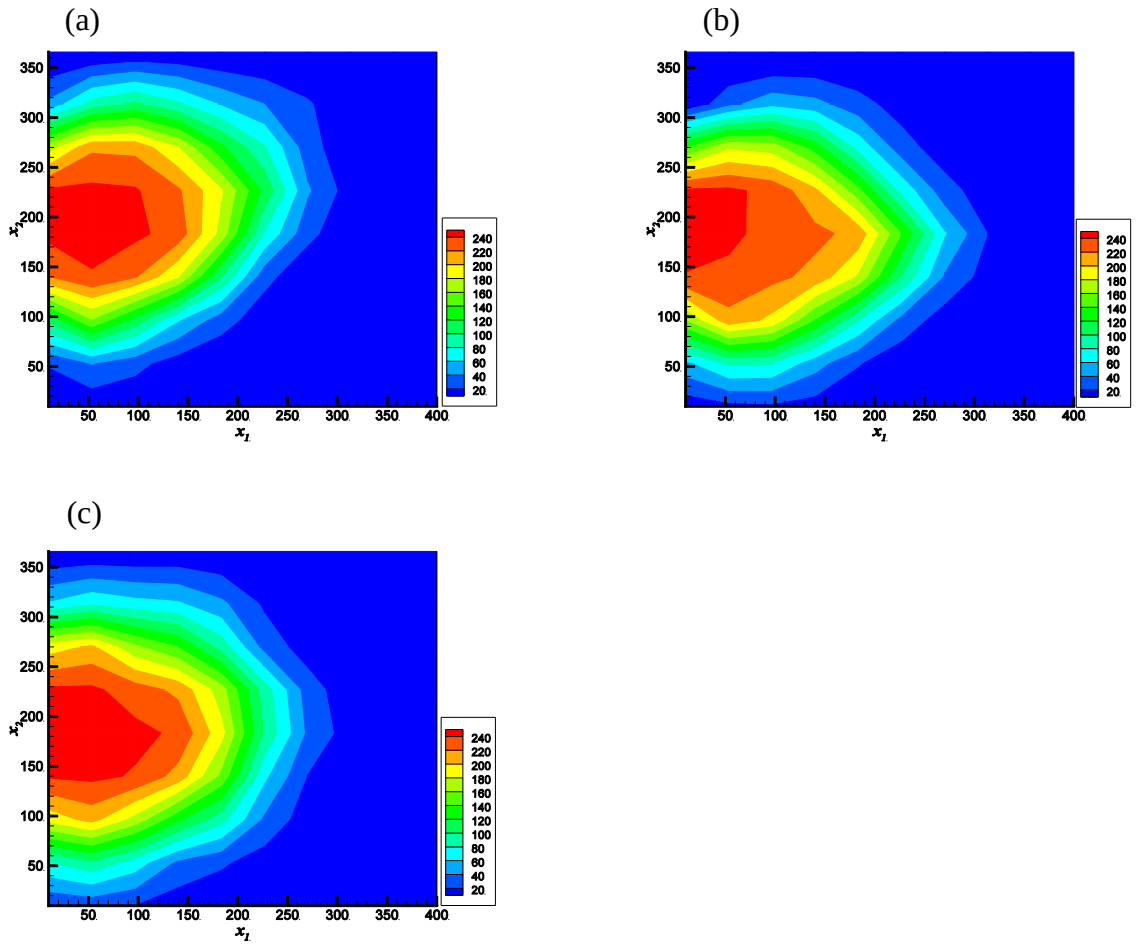
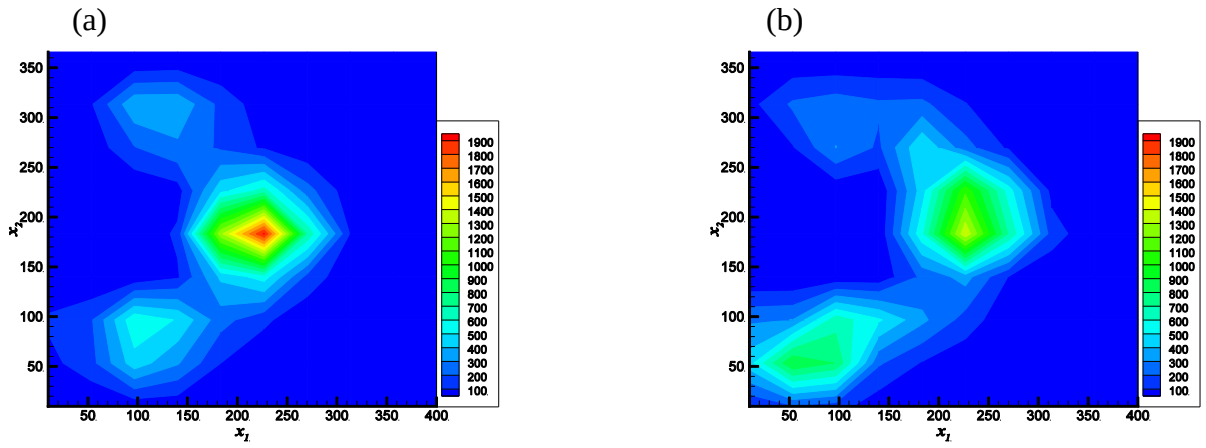


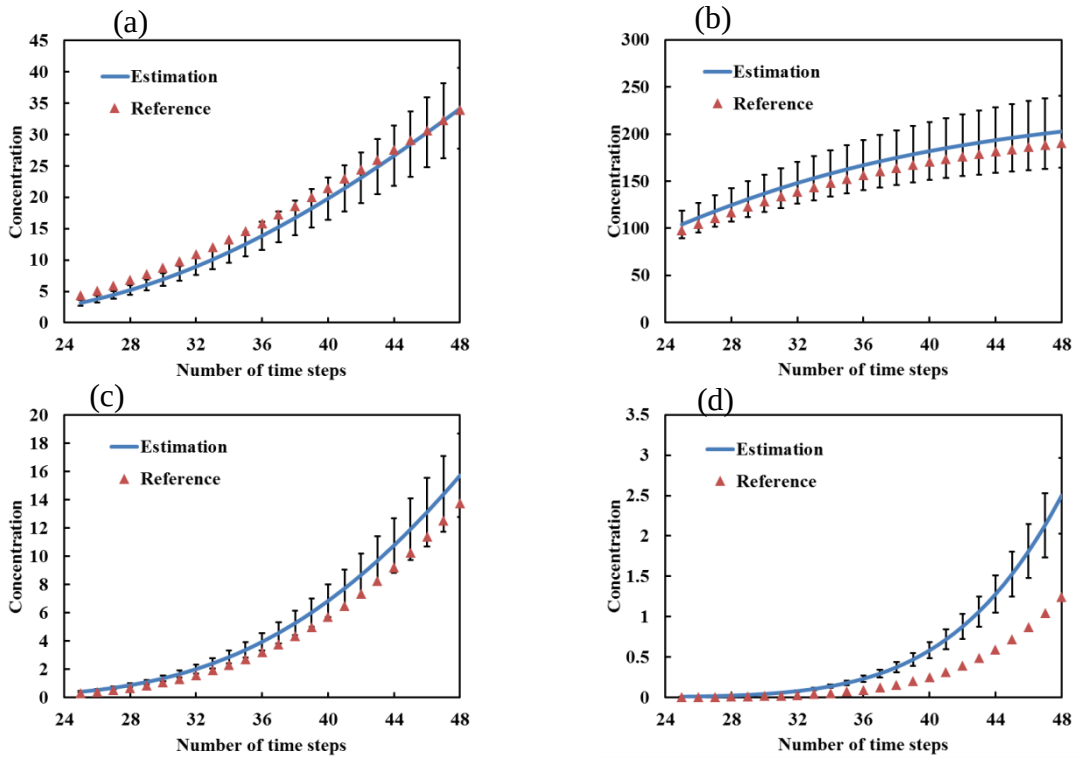
Figure 3. Decay of eigenvalues as function of the number of terms retained in the Karhunen-Loeve (KL) expansion (21).



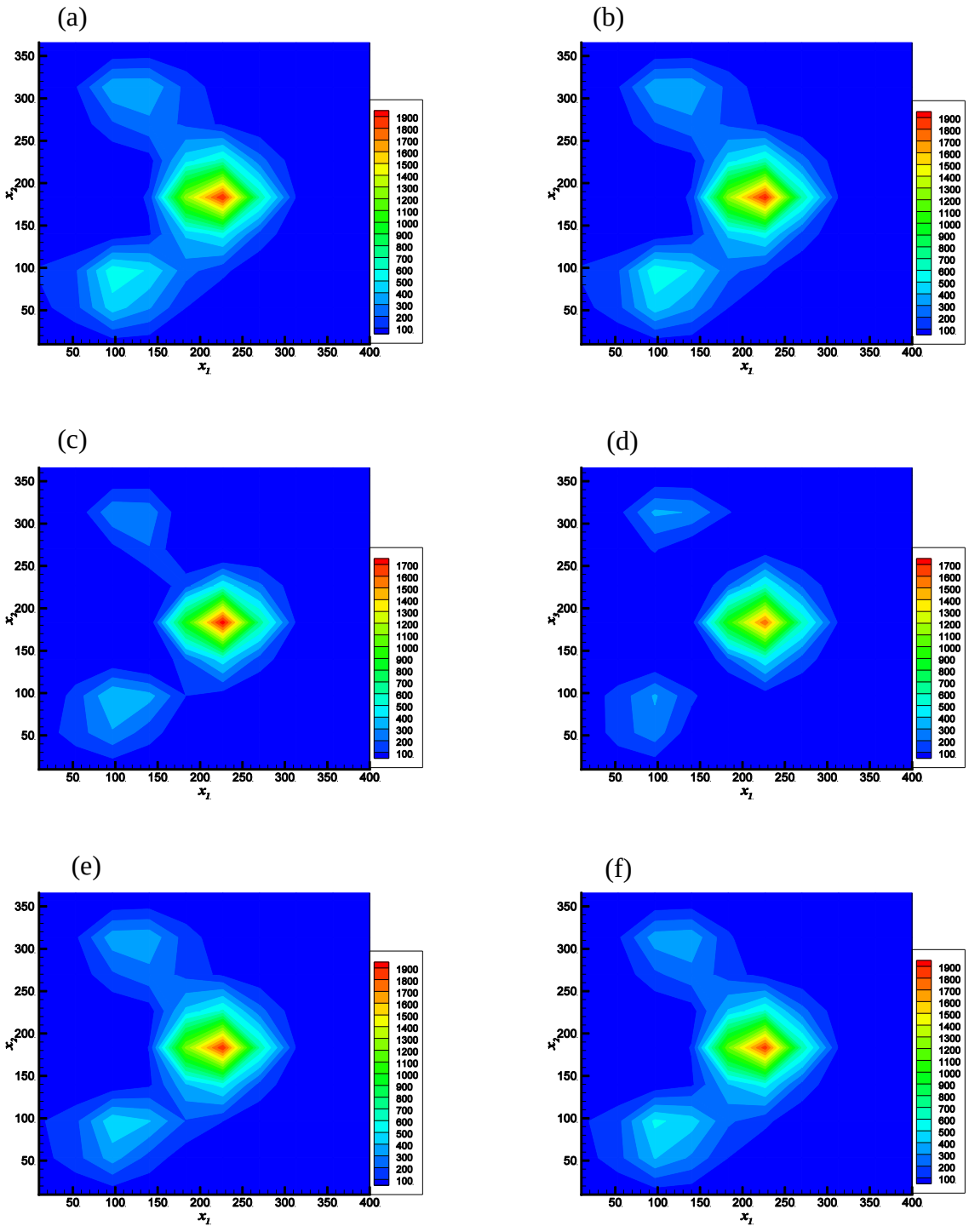
948 Figure 4. (a) Spatial distributions of true (reference) concentration field and (b) mean
949 concentration field obtained via (b) the PCKF, and (c) EnKF at time T_{60}.
950



951 Figure 5. Spatial distribution of the predictive variance of solute concentrations
952 obtained via (a) the PCKF and (b) the EnKF at time T_{60}.
953



954 Figure 6. Solute concentration values observed at observation wells located at (a)（5,
955 5）; (b)（15, 15）; (c)（25, 15）; and (d) （35, 25）at times $T_i = i\,\Delta T$ (i = 24, 25,
956 26,…, 48) with corresponding estimates of mean concentrations and envelopes of
957 width of ±1 standard deviations about the mean obtained based only on prior **D**
958 values.
959

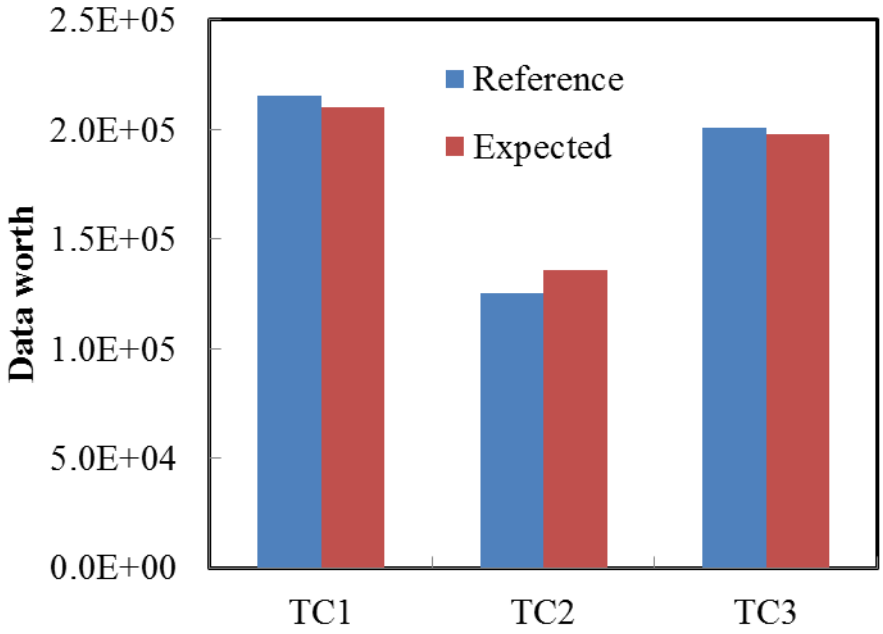


960 Figure 7. Spatial distribution of the expected concentration variance reduction,

961 $Cov_{\mathbf{C}|\mathbf{D}}E\left(\Delta\,|\,\mathbf{D},\mathbf{C}\right)$, for (a) TC1; (c) TC2; and (e) TC3. Spatial distribution of

962 concentration variance reduction, $Var\left(\Delta\,|\,\mathbf{D}\right)-Var\left(\Delta\,|\,\mathbf{D},\mathbf{C}'\right)$, obtained after **C'** has

963 been observed for (b) TC1; (d) TC2; and (f) TC3. Results are depicted for T_{60}.

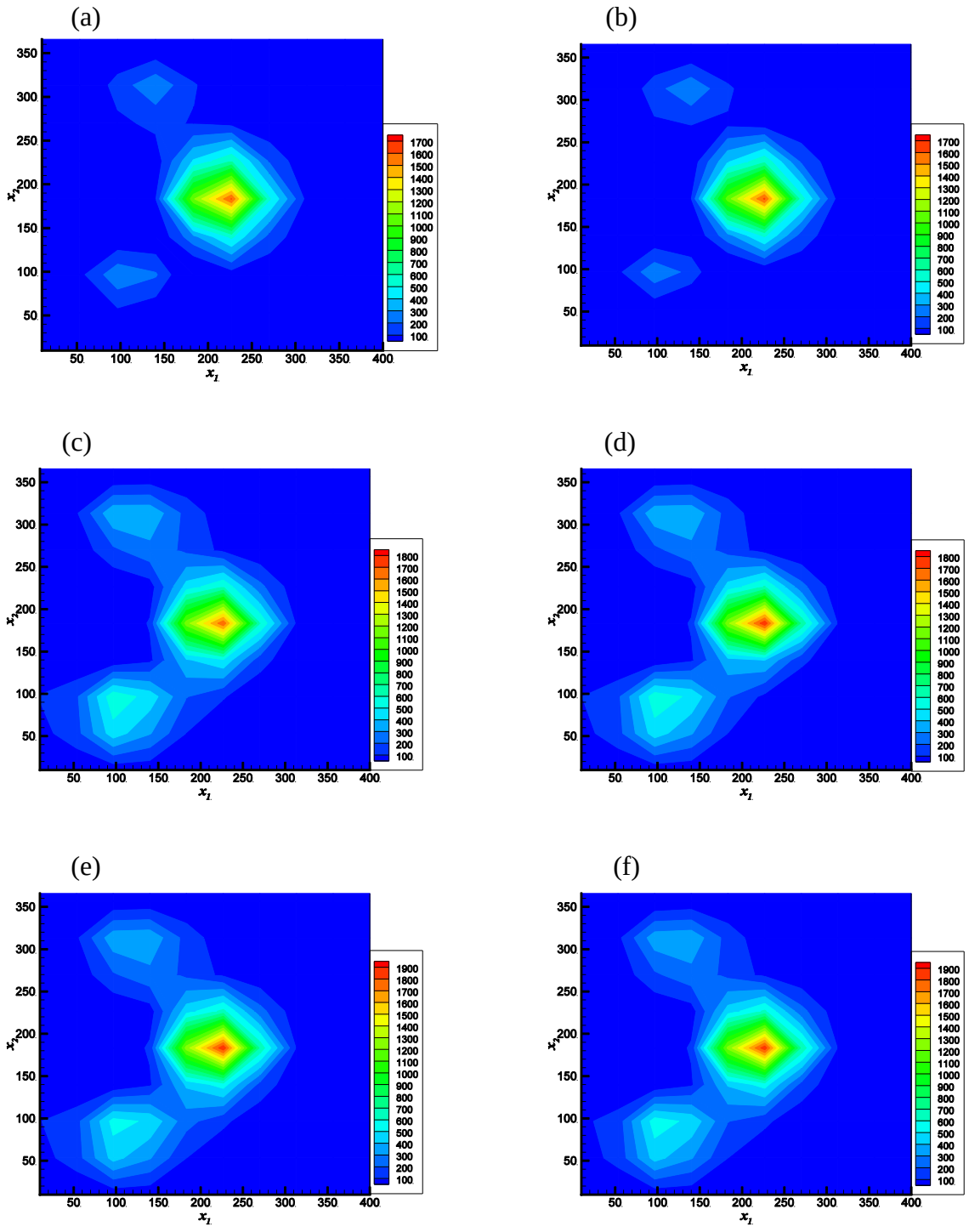


964

965 Figure 8. Scalar measures of the expected and reference data-worth for TC1, TC2, and

966 TC3 at T_{60}.

967



968 Figure 9. Spatial distribution of the expected concentration variance reduction

969 $Cov_{\mathbf{C}|\mathbf{D}}E\left(\Delta\,|\,\mathbf{D},\mathbf{C}\right)$ for (a) TC4; (c) TC5; and (e) TC6. Spatial distribution of

970 concentration variance reduction, $Var\left(\Delta\,|\,\mathbf{D}\right)-Var\left(\Delta\,|\,\mathbf{D},\mathbf{C}^{'}\right)$ obtained after **C'** has

971 been sampled for (b) TC4; (d) TC5; and (f) TC6. Results are depicted for T_{60}.

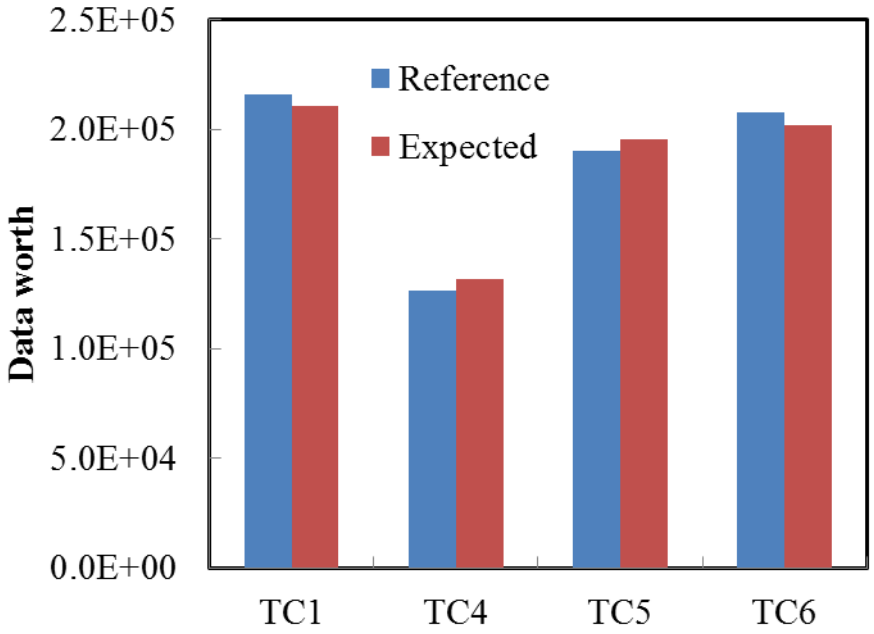


Figure 10. Scalar measures of the expected and reference data-worth for TC1， TC4, TC5, and TC6 at T_{60}.

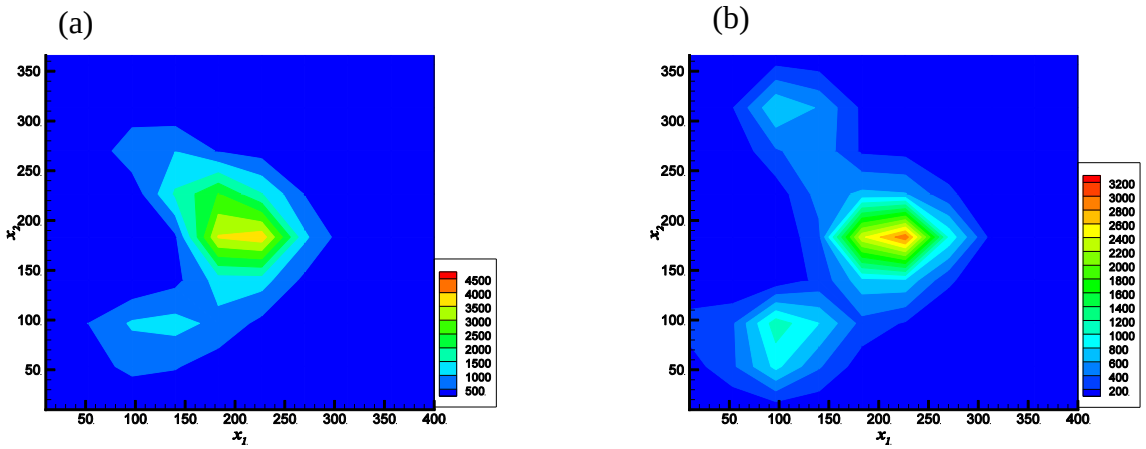
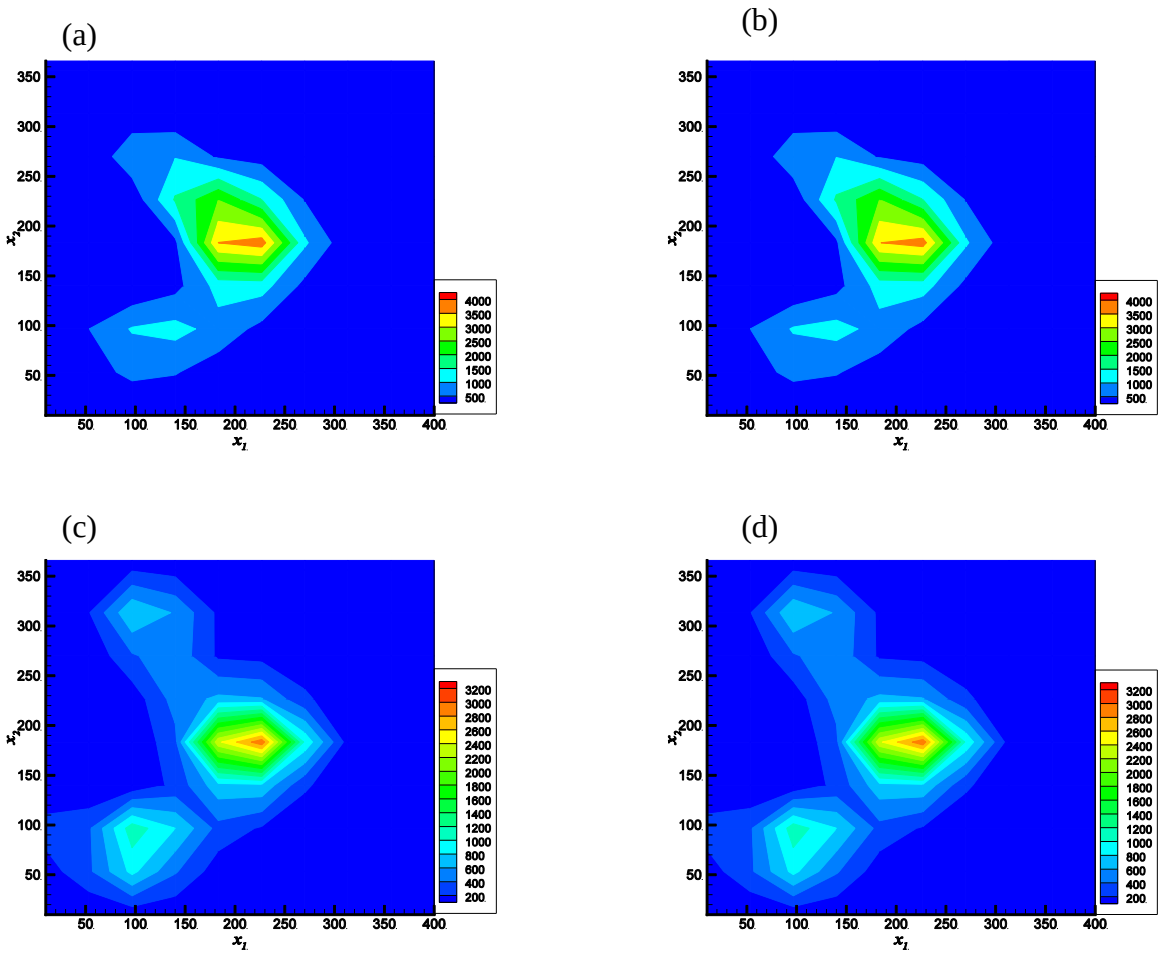


Figure 11. Spatial distribution of the predictive variance of contaminant concentration at time T_{60} conditional on $\mathbf{D}$, $Var\left(\Delta \mid \mathbf{D}\right)$, for (a) TC7; and (b) TC8. Results are depicted for T_{60}.

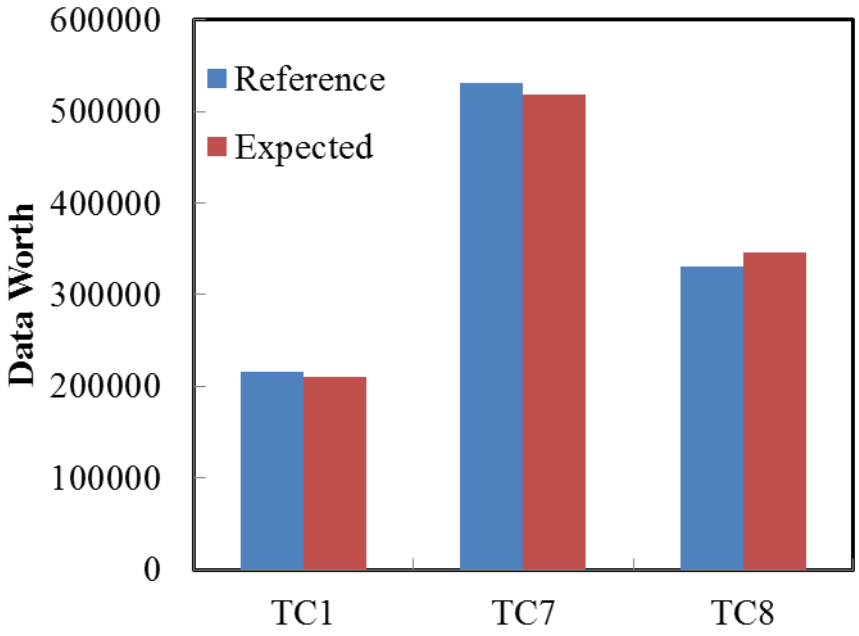


981 Figure 12. Spatial distribution of the expected concentration variance reduction,

982 $Cov_{\mathbf{C}|\mathbf{D}}E\left(\Delta\,|\,\mathbf{D},\mathbf{C}\right)$, for (a) TC7; and (c) TC8. Spatial distribution of concentration

983 variance reduction, $Var\left(\Delta\,|\,\mathbf{D}\right)-Var\left(\Delta\,|\,\mathbf{D},\mathbf{C}'\right)$, obtained after **C'** has been observed

984 for (b) TC7; and (d) TC8. Results are depicted for T_{60}.



985

986 Figure 13. Scalar measures of the expected and reference data-worth for TC1, TC7

987 and TC8 at T_{60}.

988

989 Table 1. Summary of the system setup parameters employed in the two-dimensional
990 synthetic examples.

Parameter	Value (in consistent units)
Discretization and source characteristics	
No. of rows in the domain	40
No. of columns in the domain	40
Domain width	400
Domain length	400
Time step	30
Solute source location $\boldsymbol{x}_0 = (x_{01}, x_{02})$	(20, 200)
Volumetric flux at source (q_s)	100
Concentration at source (C_s)	250
Flow and Transport Parameters	
Head along upstream boundary	12
Head along downstream boundary	7
Diffusion coefficient D_m	0.5
Local-scale dispersivities	(1.5, 0.3)
Porosity	0.15
Log Permeability	
Mean value	5
Variance	1
Correlation Length ($\eta_1 = \eta_2$)	80

991
992

Table 2. Main characteristics of test cases examined.

Test case	No. of time steps in prior data **D**	Monitoring well	Sampling frequency of additional data	Description
TC1	24	COLUMN1-4	ΔT	Base case
TC2	24	COLUMN 1-4	$6\,\Delta T$	Effect of sampling frequency
TC3	24	COLUMN 1-4	$3\,\Delta T$	Effect of sampling frequency
TC4	24	COLUMN 1,2	ΔT	Effect of monitoring network location
TC5	24	COLUMN 3,4	ΔT	Effect of monitoring network location
TC6	24	COLUMN 2,3	ΔT	Effect of monitoring network location
TC7	18	COLUMN 1-4	ΔT	Effect of prior data content
TC8	12	COLUMN 1-4	ΔT	Effect of prior data content

Table 3. Summary of the scalar measures of preposterior and posterior uncertainty reductions for all test cases (results are listed for T_{60}).

Test case	Preposterior predictive uncertainty reduction	Posterior predictive uncertainty reduction
TC1	215,742	210,452
TC2	125,292	135,982
TC3	200,672	197,962
TC4	126,698	131,980
TC5	190,141	195,533
TC6	207,585	201,750
TC7	530,745	518,494
TC8	330,028	345,770

Acknowledgements

This work is partially funded by the National Natural Science Foundation of China (Grant no. 41402199), the Science Foundation of China University of Petroleum, Beijing (Grant no. 2462014YJRC038), the independent research funding of State Key Laboratory of Petroleum Resources and Prospecting (Grant no. PRP/indep-4-1409) and the Platform Construction Project for Researches on the Relationship between Water and Ecology in the Ordos Plateau (Grant no. 201311076). The fourth author acknowledges funding from the European Unions Horizon 2020 Research and Innovation programme (Project ”Furthering the knowledge Base for Reducing the Environmental Footprint of Shale Gas Development” FRACRISK - Grant Agreement No. 640979) and funding from MIUR (Italian ministry of Education, Universities and Research – PRIN2010-11; project: "Innovative methods for water resources under hydro-climatic uncertainty scenarios").