Investigation of Data Assimilation Methods for Soil Parameter Estimation with Different Types of Data

Yuanyuan Zha, Penghui Zhu, Qiuru Zhang, Wei Mao, and Liangsheng Shi*

In the past few decades, different data assimilation methods have been proposed to estimate soil parameters. It is not clear whether a straightforward sampling approach is sufficient or whether a linear filter or even an advanced nonlinear filter is needed to interpret the potential information carried by different data types (e.g., pressure head and water content data from multiple depths, easily available surface soil moisture data, and groundwater level data). In this study, three classical data assimilation methods, i.e., the ensemble Kalman filter (EnKF), the ensemble randomized maximum likelihood filter (EnRML), and the Markov chain Monte Carlo (MCMC), were investigated numerically in terms of the utility to cope with three different types of observations. Results show that, compared with the EnKF approach, EnRML is a superior method to extract the parameter information from observations. The MCMC approach performs well in homogeneous soil but not in heterogeneous soil. Regardless of the data assimilation methods and the soil heterogeneity, point-scale soil water pressure head data are the most valuable in terms of soil parameter estimation, followed by groundwater level data, which require a nonlinear filter to interpret. A smaller observation error for groundwater level data leads to obviously improved parameter estimates by using EnRML with a slight improvement by using EnKF. The stable performance of the EnKF method relies more heavily on a relatively large number of ensembles than does EnRML. In a homogeneous soil column.

Abbreviations: EEF, estimation efficiency; EnKF, ensemble Kalman filter; EnRML, ensemble randomized maximum likelihood filter; MCMC Markov chain Monte Carlo.

Parameter estimation is of great importance in soil hydrology, considering that accurate prediction of soil moisture relies heavily on the quality of model parameters. However, direct parameter measurements are typically difficult and time consuming, especially at large spatial scales. Inverse modeling has been widely used to identify soil hydraulic parameters by minimizing the differences between measured and predicted soil water content or soil water pressure head. During the past decades, many different inverse methods have been developed, including the least-square methods (Bard, 1974), the pilot point method (de Marsily et al., 1984), the maximum likelihood method (Carrera and Neuman, 1986), the ensemble Kalman filter (Evensen, 1994), the self-calibration method (Gómez-Hernández et al., 1997), the Markov chain Monte Carlo (MCMC) method (Oliver et al., 1997), the particle filter (PF), and many other methods. Each of these approaches has its strengths and weaknesses. Reviews of these were presented by Kool et al. (1987), McLaughlin and Townley (1996), Carrera et al. (2005), and Hendricks Franssen et al. (2009). Overall, inverse models have evolved from a single estimate to stochastic Monte Carlo simulation and have become capable of handling a non-multi-Gaussian distribution (Zhou et al., 2014).

In this study, we were particularly interested in the ensemble Kalman filter (EnKF), ensemble randomized maximum likelihood filter (EnRML), and Markov chain Monte Carlo (MCMC). These three methods are all Monte Carlo based and are also typical data assimilation approaches. The EnKF, first introduced by Evensen (1994), has become one
of the most popular methods due to its computational efficiency and simplicity. Developed from the original Kalman filter, EnKF extracts auto-covariances and cross covariances from an ensemble of realizations to avoid large computation (sensitivity analysis during the linearization of the state equation) and storage costs (large parameter auto-covariances). A byproduct of using ensemble approximation in EnKF is that it gains the ability to deal with mild nonlinear problems because the ensemble statistics can be accurately evolved in time with nonlinear models, and higher order terms are retained compared with the original or extended Kalman filter (Nowak, 2009; Burgers et al., 1998). The EnKF is a sequential data assimilation method that can make use of measurements whenever they are available. Due to these merits, EnKF has been successfully applied to hydrological modeling (Clark et al., 2008; Camporese et al., 2009; Xie and Zhang, 2010; Hendricks Franssen et al., 2011), flood forecasting (Neal et al., 2007; Komma et al., 2008), land surface modeling (Crow and Wood, 2003), and crop yield forecasting (de Wit and van Diepen, 2007) among others. Recent applications of EnKF include the work of estimating the hydraulic conductivity of a groundwater aquifer (Chen and Zhang, 2006; Hendricks Franssen and Kinzelbach, 2008; Zhou et al., 2011; Li et al., 2012; Schöninger et al., 2012; Crestani et al., 2013), solute transport parameters (Liu et al., 2008), and soil parameters (Wu and Margulis, 2011; Song et al., 2014; Shi et al., 2015; Erdal et al., 2014, 2015; Bauser et al., 2016). However, EnKF is optimal only when both the model and observation operator are linear, and it is not guaranteed to be the best estimator for nonlinear cases. As revealed by Song et al. (2014), EnKF may collapse in strongly nonlinear soil water cases with a large number of unknown parameters. Recently, Li et al. (2018) showed the significant importance of characterizing the spatial-temporal feature of soil moisture data when using EnKF.

To alleviate the nonlinearity issue, some nonlinear filter approaches have been developed. For example, Zupanski (2005) proposed a maximum likelihood ensemble filter (MLEF), which is a combination of the maximum likelihood and ensemble data assimilation methods. Based on the iterative solution for the minimum of the quadratic cost function by the Newton method, Gu and Oliver (2007) presented the EnRML method. Because the EnRML can handle both nonlinear observations and nonlinear propagation of the ensemble anomalies regarding the ensemble mean, it can be seen as an extension of the MLEF (Sakov et al., 2012). Chen and Oliver (2012, 2013) later introduced an iterative ensemble smoother that assimilates all data simultaneously. Besides the nonlinear filters, some researchers have also proposed improving the capability of the linear filter (i.e., EnKF) by merging additional modifications. For example, Schöninger et al. (2012) proposed using Gaussian anamorphosis to convert a nonlinear dependence of state variables on parameters into a more linear one. Despite the increasing interest in nonlinear data assimilation from the hydrological community, research on parameter estimation by means of nonlinear filters is surprisingly limited and minimal. We emphasize that the nonlinearity in a data assimilation system depends not only on the nonlinearity of the model or observational operators but also on the state of the data assimilation system as an entire system (Verlaan and Heemink, 2001). Depending on the amount and quality of observations, the same nonlinear data assimilation system can be in either a weakly nonlinear or a strongly nonlinear regime (Sakov et al., 2012).

Unlike the optimization methods mentioned above, the basic idea of MCMC is to generate random samples in the manner of a Markov chain (Metropolis et al., 1953; Hastings, 1970). This method allows incorporation of complicated relationships between measurement data and model parameters and does not require the normalizing constant of a probability density function, which is difficult to compute. The MCMC has gained considerable popularity in the hydrological community in recent decades (e.g., Vrugt et al., 2003, 2008; Reis and Stedinger, 2005; Laloy and Vrugt, 2012; Zhang et al., 2015). The key challenge for MCMC methods is the construction of a proposed distribution that provides a close approximation of the target distribution while at the same time being inexpensive to implement. Several strategies have been introduced to alleviate the computational burden of MCMC. One is the adaptive MCMC method, especially the parallel adaptive sampler, which automatically tunes the proposed distribution during the search to the posterior target distribution. The most classical adaptive algorithms include delayed rejection and the adaptive Metropolis algorithm (Haario et al., 2006) and the differential evolution adaptive Metropolis algorithm (Vrugt et al., 2009). Some studies have also sought to accelerate MCMC-based Bayesian inference by incorporating surrogate models (Marzouk et al., 2007; Marzouk and Najm, 2009; Zhang et al., 2013). In the recent work of Zhang et al. (2016) and Man et al. (2017), an MCMC algorithm was combined with adaptively constructed Gaussian process surrogates and an ANOVA-based transformed probabilistic collocation method to avoid expensive simulations of the original model. Despite broad application in multiple disciplines, it is necessary to note that MCMC is a sampling approach rather than an optimization algorithm in the sense that it generates conditional realizations by sampling from the posterior distribution of the parameters samples. Moreover, the tremendous computational burden of MCMC in a complex and nonlinear system is still a significant concern.

One major difficulty for inverse problems is ill-posedness, which is defined as non-uniqueness, nonexistence, and non-steadiness of the solutions (Hadamard, 1902). Most of the inverse problems in hydrology are ill-posed due to the lack of information on system states (e.g., soil water content) and on initial and boundary conditions. Ill-posedness could become very severe when the number of unknown parameters far exceeds that of the observation data. It can also be triggered by solution instability: parameters to be estimated are highly sensitive to changes in the response of the subsurface (Mao et al., 2013). It is clear that the ill-posedness of inverse problems is largely related to the availability of data, and the selection of a data assimilation method becomes critical to successfully identify the global minimum for ill-posed problems.
Most inverse studies on soil parameter estimation have focused on soil water pressure head and soil water content data. In our recent work (Shi et al., 2015; Zhu et al., 2017), we showed that other types of measurements such as surface soil moisture data (from different scales) and groundwater level data also contain valuable information on soil parameters, and these data are easily available in practice. Parameters, as well as system states, are connected to different types of observations in different ways. However, an overly complex relation between the observations and parameters may hinder correct or accurate parameter estimation (Shi et al., 2015).

In this study, three typical data assimilation methods were selected to interpret different types of data. The purpose of this study was not simply comparing the differences of their performance in the circumstance of infusing various data into a soil water system but also contemplating the necessity of introducing advanced nonlinear approach or whether a straightforward sampling approach or linear filter is adequate. As suggested by Zhou et al. (2014), “...the best inverse model should be the one that is stochastic, is capable of dealing with multiple sources of state data governed by a complex state equation, is not limited to multi-Gaussian realizations, and can weight in prior information.” Considering the inherent nonlinearity of soil water flow (Zha et al., 2019), it will be interesting to investigate the capability of typical data assimilation methods in dealing with different data with a probability distribution and representative scale (e.g., pressure head and soil water content, averaged surface soil water content, and groundwater level data). The motivation of this work was thus to investigate the performance of different data assimilation approaches (i.e., EnKF, MCMC, and EnRML) for estimating soil hydraulic parameters in a one-dimensional soil column using three different observations (i.e., surface soil water content, soil water pressure head, and groundwater level level). Because a successful data assimilation depends on several critical issues, including the construction of a state vector, the availability of observation data, and the identifiability of model parameters (Hu et al., 2017), this study will help modelers to revisit the ability of data assimilation methods under a wide range of available data and conditions.

**Methodology**

**One-Dimensional Model of Variably Saturated Flow**

Soil water in this study was simulated using vertically one-dimensional, variably saturated flow, with the governing Richards equation written as

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left( K(h) \left[ 1 + \frac{\partial h}{\partial z} \right] \right)$$

where \( \theta \) is the volumetric soil water content; \( t \) is time \([T]\); \( h \) is the pressure head \([L]\), positive or zero if the soil is saturated and negative if it is unsaturated; \( z \) is the vertical dimension \([L]\); and \( K(h) \) is the unsaturated hydraulic conductivity. A definition of the relationships between soil water content, unsaturated conductivity, and pressure head is required to solve Eq. [1]. The van Genuchten–Mualem model (van Genuchten, 1980; Mualem 1976) is used to describe these constitutive relationships:

$$\theta(h) = \theta_i + (\theta_s - \theta_i) \left(1 + |a h|^n \right)^{-m}$$

$$K(\theta) = K_s \left[ 1 - \left(1 - S_c^{1/m} \right)^m \right]$$

where \( \theta_i \) and \( \theta_s \) are the saturated and the residual soil water contents; \( \alpha \ [L^{-1}] \) and \( n \) are empirical parameters related to the first and second moments of the pore size density function, respectively; \( m \) is related to \( n \) as \( m = 1 - 1/n \); \( K_s \) is the saturated hydraulic conductivity; and \( S_c \) is the effective saturation degree, \( S_c = (\theta - \theta_i)/(\theta_s - \theta_i) \). The widely used HYDRUS-1D model for saturated–unsaturated flow modeling (Simůnek et al., 2005) was used in this study. The soil hydraulic parameters \( \alpha, n, \) and \( K_s \) are treated as unknown parameters to be estimated.

**The Ensemble Kalman Filter**

Being a Monte Carlo variant of the standard Kalman filter, EnKF uses an ensemble of model realizations to approximate the prior covariance of the state vector instead of explicitly computing the covariance matrix. The EnKF updates the unknown parameters and uncertain variables sequentially with time whenever observation data are available. The augmented state vector \( \mathbf{V}_k \) at the \( k \)-th time step can be written as

$$\mathbf{V}_k = \left( \mathbf{m}_k^T, \mathbf{h}_k^T \right)^T$$

where \( \mathbf{m}_k \) is the parameter vector including parameters of all the one-dimensional grid points and \( \mathbf{h}_k \) is the vector of state variables (pressure head in this study). When observation data are acquired, the state vector of every ensemble member will be updated in parallel by

$$\mathbf{V}_{k,i} = \mathbf{V}_{k,i} + \mathbf{K}_k \left( \mathbf{d}_{k,i} - \mathbf{H}_k \mathbf{V}_{k,i} \right)$$

where \( \mathbf{V}_{k,i} \) and \( \mathbf{V}_{k,i} \) represent the predicted and updated state vectors for the \( i \)-th ensemble member at the \( k \)-th time step, \( \mathbf{H}_k \) is the observation operator that maps the model states to the observation space and therefore \( \mathbf{H}_k \mathbf{V}_{k,i} \) is the predicted counterpart of the corresponding actual observation data, and \( \mathbf{d}_{k,i} \) is the \( i \)-th realization of the actual observation data \( \mathbf{d}_k \):

$$\mathbf{d}_{k,i} = \mathbf{d}_k^i + \mathbf{e}_{k,i}$$

where \( \mathbf{e}_{k,i} \) is the independent white noise of the observations, varying among realization members (Burgers et al., 1998). In this study \( \mathbf{e}_{k,i} \) was assumed to be temporally constant, which does not affect our research purpose. The term \( \mathbf{K}_k \) in Eq. [5] is conventionally called the *Kalman gain* and acts as a weighting factor between observations and model predictions:

$$\mathbf{K}_k = \mathbf{P}_k^i \mathbf{H}_k^T \left( \mathbf{H}_k \mathbf{P}_k^i \mathbf{H}_k^T + \mathbf{R}_k \right)^{-1}$$

where \( \mathbf{R}_k \) is the error covariance matrix of the observations at the \( k \)-th step and \( \mathbf{P}_k^i \) is the prior error covariance matrix of the state.
vector, which in EnKF is approximated by the forecast ensemble members:

$$P^{f}_{k} \approx \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \left( \mathbf{V}^{f}_{k,i} - \mu_{k} \right) \left( \mathbf{V}^{f}_{k,i} - \mu_{k} \right)^{T}$$  \[8\]

$$V^{f}_{k} \approx \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} V^{f}_{k,i}$$  \[9\]

where $N_{\text{ens}}$ is the ensemble size, $\mathbf{V}^{f}_{k,i}$ is the ensemble mean of the state vectors of all the ensemble members before assimilation. The updated state vectors are inserted back into the model, based on which all the ensemble members will then run forward in parallel until the next observation time, at which the state vectors will be updated again.

**The Ensemble Randomized Maximum Likelihood Filter**

The EnKF method has achieved great popularity because it is generally effective and easy to use. In the forecast process, the EnKF method can approximately address the nonlinearity of the physical dynamics, but in the analysis step, this method sometimes fails, especially for highly nonlinear problems (Gu and Oliver, 2007). To tackle this issue, Gu and Oliver (2007) proposed an iterative ensemble Kalman filter, namely the EnRML, which is similar to the iterative Kalman filter of Jazwinski (1970) but using the ensemble form instead. In the EnRML method, the iteration steps are added to enforce the nonlinear measurement constraints. A brief introduction of the EnRML method is given here. See Gu and Oliver (2007) for a detailed discussion.

In EnRML, the Gauss–Newton method is applied to minimize the stochastic objective function, and parameters are updated by (the time index $k$ has been omitted for simplicity)

$$m^{l+1} = m^{pr} + (1-\beta_{l}) m^{l} - \beta_{l} C_{M} G_{f}^{T} \left( C_{D} + G_{f} C_{M} G_{f}^{T} \right)^{-1} \times \left( g(m^{l}) - d^{l} - G_{f} \left( m^{l} - m^{pr} \right) \right)$$  \[10\]

where $m^{l+1}$ is the $(l+1)$th iterative estimate of the model parameters for the $i$th ensemble member, $\beta_{l}$ is a coefficient to adjust the step size at each iteration, $m^{pr}$ is the prior parameter value, $C_{M}$ is the prior covariance of parameters estimated from the ensemble (namely $P^{f}_{k}$ in Eq. [8]), which does not change during the Gauss–Newton iteration, $C_{D}$ is the error covariance matrix of the observations, which is expressed as $R$ in Eq. [7], $g(m^{l})$ is the prediction of observations for the $i$th ensemble member, $d^{l}$ is the $i$th realization of the actual observation in Eq. [6], and $G_{f}$ is the ensemble average sensitivity matrix at the $i$th iteration, which reflects changes in the model parameters in response to changes in the computed data:

$$\Delta D^{f} = G_{f} \Delta M^{f}$$  \[11\]

where $\Delta D^{f}$ is an $N_{D} \times N_{\text{ens}}$ matrix representing the deviation of each vector of model parameters from the current mean, and $G_{f}$ is an $N_{D} \times N_{\text{ens}}$ matrix, $N_{\text{ens}}$ is the number of model parameters and $N_{D}$ is the number of data points. The singular value decomposition is used to solve the system because $\Delta M^{f}$ is not invertible generally.

The implementation process of EnRML is very similar to that of the EnKF method, except the iteration steps. Details of the iteration steps can be found in Gu and Oliver (2007) (see also below). It should be mentioned that in this study we used a maximum iteration of 100 to ensure the convergence of the calculation.

**Bayesian Inference and Markov Chain Monte Carlo**

The Bayesian method has been widely used to fit model parameters to data. Bayes’ rule updates the prior distribution when observations become available. We use a nonlinear model $F$ with parameter $m$ to simulate the observed data:

$$d = F(m) + \varepsilon$$  \[12\]

where $F(m)$ is the nonlinear soil water flow model, $m$ is the $N_{m} \times 1$ vector of unknown soil parameters to be estimated from the data, and $\varepsilon$ is the $N_{m} \times 1$ vector of measurement error, which is generally assumed to be a normally distributed random variable with zero mean.

Given the observations $d$, the posterior distribution of model parameters $p(m|d)$ can be described as

$$p(m|d) = \frac{p(m) p(d|m)}{\int p(d|m) p(m) dm}$$  \[13\]

where $p(m)$ is the prior distribution and $p(d|m)$ is the likelihood function, which represents the discrepancy between system states and observations. In this study, the likelihood function was assumed to be Gaussian:

$$p(d|m) \propto \exp \left\{ -\frac{1}{2} \| d - F(m) \|^{T} R^{-1} \| d - F(m) \| \right\}$$  \[14\]

For most nonlinear problems, it is very difficult to calculate the posterior distribution directly. The MCMC method first generates an initial value $m^{0}$, then produces a sequence of dependent parameter values $\{ m^{1}, m^{2}, m^{3}, \ldots \}$ in the manner of the Markov process. The earliest MCMC approach works as follows (Metropolis et al., 1953):

Step 1: A candidate $m^{*}$ is sampled from a symmetric proposal distribution $q(m|\cdot)$.

Step 2: The candidate point is either accepted or rejected using the Metropolis probability ratio:

$$\alpha(m, m^{*}) = \min \left\{ \frac{p(m^{*}|d) q(m|m^{*})}{p(m|d) q(m^{*}|m)}, 1 \right\}$$  \[15\]

Step 3: If the proposal is accepted, the chain moves to $m^{*}$; otherwise, the chain remains at $m$.

Under certain regularity conditions, the generated samples will eventually converge to the stationary, posterior distribution...
of the estimated parameters. The major computational burden of MCMC is in computing the acceptance rate $\alpha(m,m^*)$, which highly depends on the selection of the proposed distribution. A too-wide distribution leads to very few candidate samples accepted, while a too-narrow distribution limits the changing range of the chain and thus requires a large number of iterations to obtain the desired posterior distribution (Laloy et al., 2013). In this study, the differential evolution adaptive Metropolis algorithm developed by Vrugt et al. (2009) was used to generate the conditional samples. Specifically, we set up our experiments following the DREAM(ZS) version of this algorithm by Vrugt (2016). See Vrugt et al. (2009) and Vrugt (2016) for details of the algorithm.

**Numerical Experiments**

Synthetic experiments were conducted to investigate the ability of the EnKF, EnRML, and MCMC methods to assimilate different types of data. A reference modeling was performed first, in which parameters used the reference values. The data-assimilation runs were then implemented during the same time period using the initial soil hydraulic parameters (deviated from the reference parameters). Initial and boundary conditions of the assimilation runs and the reference run were set to be identical because this study focused on uncertainties arising from soil hydraulic parameters. The assimilation runs assimilate observation data drawn from the reference modeling to estimate the unknown parameters. Our analysis included the heterogeneity of soil media, the aforementioned three data assimilation methods, different data types, ensemble size, and different observation errors. Our purpose was to reveal to what extent these factors can influence the estimation of soil hydraulic parameters.

**Flow Domain and Initial and Boundary Conditions**

Two synthetic, one-dimensional soil columns of 250-cm height were constructed, one with homogeneous and the other heterogeneous soil. We used one-dimensional models because the Darcian flux in the vertical direction is much greater than that in the lateral direction without a significant hillslope (Mantoglou, 1992; Chen et al., 1994). The flow domain was discretized into 51 grid points with an equal interval of 5.0 cm. The total simulation time was 92 d. The soil type selected was sandy loam. Detailed parameter settings are discussed below. The top boundary was given with specified rainfall and potential evapotranspiration, as shown in Fig. 1. This boundary condition resulted in infiltration-dominated processes in our study. The bottom boundary was set as a zero-flux boundary. The initial groundwater table was at 218 cm below the soil surface. The initial condition was deterministic, with a total head of −218 cm for all the grid points. Soil water flow was simulated with a time step of 0.1 d, and the observation data were assimilated on a daily basis.

**Observations**

Three types of observation data were considered, namely point-scale soil water pressure head (SWH), surface soil water content (SSWC), and groundwater level (GWL). For the homogeneous soil column, the soil water pressure head was observed at the depth of 125 cm, while for the heterogeneous one, soil water pressure heads were observed at depths of 60 and 125 cm, unless otherwise stated. Surface soil water content is the averaged soil water content at 0 to 22.5 cm, corresponding to the control length of the top five nodes. The groundwater level in the numerical model was determined by interpolating the vertical coordinates of the two adjacent nodes with positive and negative pressure heads.

To consider the influence of measurement precision, we assumed two levels of observation error. For the soil water pressure head data, we considered relative measurement errors of 5 and 20%. For the surface soil water content observation, we set absolute measurement errors of 0.02 and 0.05 cm³ cm⁻³. For the groundwater level data, we used absolute measurement errors of 0.01 and 0.05 m. The observation data were drawn from the reference modeling, so they contain information about the reference parameters. For EnKF and EnRML, data drawn from the reference modeling were perturbed with the above errors. The MCMC method was implemented using reference observations (zero error) without perturbation in this synthetic study.

**Parameter Setting**

Soil parameter statistics for sandy loam from Carsel and Parrish (1988) were used as a basis in this study. Parameters $q_s$ and $q_r$ were taken as known, using the values of 0.41 and 0.065 throughout all study cases. The mean and standard deviation of parameters $\alpha$, $n$, and $K_s$ were first transformed into log-space, and
it was assumed that unknown soil parameters obeyed the lognormal distribution. The original normal-space statistics of Carsel and Parrish (1988) and the logarithmically transformed statistics of parameters $\alpha$, $n$, and $K_s$ are listed in Table 1. For the data assimilation runs, initial realizations of these three parameters were generated using the lognormal distribution with mean $\mu + \sigma$ and standard deviation $\sigma$, where $\mu$ and $\sigma$ are the reference (or true) mean and standard deviation of the logarithmic parameter listed in Table 1. That is, the initial guessed ensemble mean had a shift of one standard deviation from the true distribution. For the homogeneous soil column, only three parameters needed to be initialized. The initial parameter realizations for data assimilation were generated using the MATLAB function of `mvnrnd` with the biased mean and standard deviation $\sigma$ (see Table 1 for values), while the reference mean value $\mu$ was used in the reference modeling. For the heterogeneous soil column, the logarithmic parameters were assumed to be second-order stationary with a covariance function defined by the exponential form:

$$C_y(b) = \sigma_y^2 \exp\left(-\frac{|b|}{\lambda}\right) = \sigma_y^2 \exp\left(-\frac{|z_1 - z_2|}{\lambda}\right)$$  \hspace{1cm} \text{(16)}$$

where $z_1$ and $z_2$ are the one-dimensional coordinates of two grid points, $\sigma_y$ is the variance, and $\lambda$ is the correlation length. We generated the initial realizations of a random field with mean $\mu$ and covariance $C_y(b)$ via the Karhunen–Loeve expansion method (Zha et al., 2018). The prior mean and variance of the logarithmic soil hydraulic field can be found in Table 1. The correlation length of the initial parameter realizations was 120 cm. The reference field was determined by randomly selecting a realization from the initial ensemble realizations. Note that no correlation between different types of parameters was considered in this study because, in practical situations, knowledge about such correlations is difficult to obtain. The model structural errors were ignored because the same model was applied in the reference modeling and the assimilation runs. The data assimilation methods of EnKF and EnRML were implemented with three ensemble sizes (10, 50, and 500) to investigate the influence of the number of ensemble members. As for the MCMC method, we arbitrarily selected chain numbers of $p = 1$ for the homogeneous soil column and $p = 51$ for the heterogeneous column. The decreasing percentage of final RMSE relative to the initial value was calculated for $\alpha$, $n$, and $K_s$ separately to represent the estimation efficiency (EEF, %):

$$\text{EEF} = \frac{\text{RMSE}_{\text{ini}} - \text{RMSE}_{\text{end}}}{\text{RMSE}_{\text{ini}}} \times 100$$  \hspace{1cm} \text{(18)}$$

where RMSE$_{\text{ini}}$ and RMSE$_{\text{end}}$ are the initial and final RMSE values for one particular type of soil parameter. A positive and higher EEF implies improved parameter estimation, while a negative EEF represents worsened estimation after data assimilation.

### Results and Discussion

We analyzed the data assimilation results using different algorithms, data types, estimation accuracies, and ensemble sizes. First, the results using algorithms EnKF, EnRML, and MCMC with data on SWH, SSWC, and GWL are compared. Specifically, we compare the temporal evolution of the estimated parameters using EnKF and EnRML for the homogeneous soil (Fig. 2–3; MCMC performance is given in Table 2), while the final estimated

<table>
<thead>
<tr>
<th>Observation†</th>
<th>RMSE</th>
<th>Estimation efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$n$</td>
</tr>
<tr>
<td>SWH</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>GWL</td>
<td>0.007</td>
<td>0.001</td>
</tr>
<tr>
<td>SSWC</td>
<td>0.007</td>
<td>0.001</td>
</tr>
</tbody>
</table>

† SWH, soil water pressure head; GWL, groundwater level; SSWC, surface soil water content.

### Table 1. Soil parameters used in simulation.

<table>
<thead>
<tr>
<th>Parameter†</th>
<th>Reference mean</th>
<th>SD</th>
<th>In mean ($\mu$)</th>
<th>In SD ($\sigma$)</th>
<th>Initial guess of ln mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_s$</td>
<td>0.41</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$\theta_r$</td>
<td>0.065</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$K_s$, m $d^{-1}$</td>
<td>7.5</td>
<td>3.7</td>
<td>1.906</td>
<td>0.467</td>
<td>2.373</td>
</tr>
<tr>
<td>$n$</td>
<td>1.89</td>
<td>0.17</td>
<td>0.633</td>
<td>0.090</td>
<td>0.722</td>
</tr>
<tr>
<td>$K_s$, m $d^{-1}$</td>
<td>1.061</td>
<td>1.351</td>
<td>–0.423</td>
<td>0.982</td>
<td>0.559</td>
</tr>
</tbody>
</table>

† $\theta_s$ and $\theta_r$, saturated and residual volumetric water content; $\alpha$ and $n$, empirical parameters related to the first and the second moments of the pore size density function, respectively; $K_s$, saturated hydraulic conductivity.

### Performance Assessment

The ensemble mean values of the parameters at the end of the simulation time for EnKF and EnRML, and at the end of the sampling process for the MCMC method, were used to evaluate whether a combination of analysis factors yielded satisfying results. We also present the temporal evolution of the root mean square error (RMSE) with respect to $\alpha$, $n$, and $K_s$ estimation, which is defined as

$$\text{RMSE} = \sqrt{\frac{1}{n_p} \sum_{i=1}^{n_p} (E(x_i) - x_i^{true})^2}$$  \hspace{1cm} \text{(17)}$$

where $n_p$ is the number of unknown parameters, $E(x_i)$ is the ensemble mean value of the $ith$ parameter, and $x_i^{true}$ is the reference value. Those three types of parameters were evaluated separately, and therefore $n_p = 1$ for the homogeneous soil column and $n_p = 51$ for the heterogeneous column. The decreasing percentage of final RMSE relative to the initial value was calculated for $\alpha$, $n$, and $K_s$ separately to represent the estimation efficiency (EEF, %):
parameters using EnKF, EnRML, and MCMC for the heterogeneous soil are shown in Fig. 4 to 6. Next, we focus on the influences of data accuracy (Fig. 7–8) and ensemble size (Fig. 9–11) on parameter estimation using EnKF and EnRML.

Comparison of Different Data Assimilation Methods

The RMSE values of $\alpha$, $n$, and $K_s$ assimilated by using different types of data with the EnRML and EnKF approaches are shown in Fig. 2 (for the homogeneous soil) and Fig. 3 (for the heterogeneous soil). Both results correspond to utilization of an ensemble size of 500 and the smaller measurement errors described above. Each curve in Fig. 2 and 3 represents the temporal evolution of the RMSE for each parameter corresponding to the given combinations of data type and method. For the heterogeneous soil column, we plotted the individual grid-wise soil parameters of the reference values, the initial guess, and the final ensemble mean estimates by the aforementioned three methods in Fig. 4 to 6, using the SWH, GWL, and SSWC data, respectively. The abscissa in Fig. 4 to 6 represents the node order number of the soil column. In each figure, the EEF values are presented for the convenience of comparison. The RMSE and EEF values from the MCMC method are given in Table 2.

It can be seen from Table 2 that the MCMC method accurately yielded the reference values of the three parameters for the homogeneous soil column, with final RMSE values almost equal to zero. For the heterogeneous soil column with a total of 153 parameters to be estimated, the MCMC method is probably not sufficient to give parameter estimation as accurate as those from the EnKF and EnRML methods. According to Vrugt (2016), the MCMC method usually requires millions of samplings for a high-dimensional problem. Although we utilized a total of 50,000 samplings in this study, and visible improvement of the 153 parameter estimates can be achieved (see Fig. 4–6), it is expected that a larger sampling size can lead to a better estimation.

Figures 2, 3, and 6 show that the EnRML and EnKF methods had equal ability to assimilate the SSWC data for both the homogeneous and heterogeneous soil columns. Because the EnRML method requires a much heavier computational burden.

![Fig. 2](image1.png)

Fig. 2. Temporal evolution of RMSE values for empirical parameters (a) $\alpha$ and (b) $n$ and (c) saturated hydraulic conductivity $K_s$ in the homogeneous soil column using soil water pressure head data (SWH), groundwater level data (GWL), or surface soil water content data (SSWC) and either the ensemble randomized maximum likelihood filter (EnRML) or the ensemble Kalman filter (EnKF). Results were obtained with an ensemble size of 500 and the smaller observation errors.

![Fig. 3](image2.png)

Fig. 3. Temporal evolution of RMSE values for empirical parameters (a) $\alpha$ and (b) $n$ and (c) saturated hydraulic conductivity $K_s$ in the heterogeneous soil column using soil water pressure head data (SWH), groundwater level data (GWL), or surface soil water content data (SSWC) and either the ensemble randomized maximum likelihood filter (EnRML) or the ensemble Kalman filter (EnKF). Results were obtained with an ensemble size of 500 and the smaller observation errors.
due to the iterations (in our experiment, an average of nearly 20 iterations at every analysis step was required), the EnKF method is thus recommended when assimilating surface soil moisture data.

If soil water pressure head observations are available, it can be seen in Fig. 2 and 3 that the EnRML method resulted in faster convergence than the EnKF method. However, the final parameter estimates by EnRML and EnKF are close if long-term observations are assimilated (see Fig. 2–3), except for parameter $n$ in Fig. 2b, in which the curve of SWH–EnRML is lower than the curve of SWH–EnKF. For the heterogeneous soil column, the MCMC method also brought reasonably good parameter estimates using SWH data (Fig. 4), but the accuracy of 153 parameter estimates (Fig. 4) is not as good as that from the EnRML and EnKF methods.

The difference between EnRML and EnKF is distinct when the measurement is GWL data, whose relationship with the parameters is complicated and nonlinear. From Fig. 2 and 3, it can be seen that by using the combination of EnRML and GWL data, the estimation of all parameters was greatly improved, even for the heterogeneous soil column. In comparison, EnKF created noticeable value only for parameter $K_s$ estimation but showed limited ability to estimate parameters $\alpha$ and $n$ using GWL data. This conclusion can be reached for both the homogeneous and heterogeneous soil columns. As for the MCMC method, the results are comparable to the EnRML method for parameters $\alpha$ and $K_s$ in the heterogeneous soil column (see Fig. 5).

As revealed by Shi et al. (2015) and Zhang et al. (2018), the change of groundwater level contains useful information on soil properties. Our results demonstrate that the simple EnKF method does not seem competent enough to extract such information (see Fig. 2a, 2b, 3a, and 3b). In contrast, EnRML is a superior method to interpret the GWL data. While previous studies (Shi et al., 2015; Hopmans and Šimůnek, 1999) emphasized that "an increase in the number of estimated parameters entails the need for further measurements of different types," we herein demonstrate the significance of using more advanced methods to extract as much information from the available limited data as possible.
The Effectiveness of Different Types of Measurements on Soil Parameter Estimation

Based on the results presented above, the EnRML method generally led to the best parameter estimation. Therefore, the results based on EnRML were selected to analyze the potential information content of each measurement type on parameter estimation. From Fig. 2 to 6, it can be seen that, generally, all three types of data have more or less value in estimating soil parameters. While surface soil moisture data are useful in simultaneously estimating different types of parameters for the homogeneous soil column (Fig. 2), it seems that soil moisture is less useful in estimating different types of grid-wise parameters of the heterogeneous soil column, especially for parameters $\alpha$ and $n$ (Fig. 3 and 6). This indicates that the downward propagation of surface soil moisture data is difficult. The results are in agreement with former studies such as those of Ines and Mohanty (2008a, 2008b), in which the surface soil moisture alone was found inadequate to identify the parameters of deeper soil layers. Our present study shows that no further information on deep soil parameters can be uncovered from the surface soil moisture data by using the EnRML or MCMC methods.

Figures 2a and 3a show that the SWH and the GWL data are more effective in estimating parameter $\alpha$ for both the homogeneous and heterogeneous soil columns compared with the SSWC data. For the heterogeneous soil column, Fig. 3 reveals that the SWH data at two depths resulted in much better estimates than the SSWC and GWL data, which may result from SWH measurements having been made at two different spatial locations in this study. The results highlight the importance of SWH data in characterizing the soil spatial heterogeneity, although soil water pressure head measurement usually suffers from great uncertainty.

The results also demonstrate that different types of data may possess different information content for different types of parameters. The information content contained intrinsically depends on the explicit vs. implicit and linear vs. nonlinear relationship between the data and soil properties. Besides the governing equation of soil water flow, the relationship is further influenced by the measurement locations. It is reasonable that SSWC observations contain more information about parameters of the shallow soil layers and that point-scale SWH data are more useful for parameter estimation of nearby regions, while GWL data contain considerable information on the overlying soil parameters but less or no information on the underlying soil properties. Different types of measurements may also exhibit synergistic effects if they are assimilated simultaneously. For example, when estimating $K_s$, assimilating all three types of data simultaneously can produce better estimation than those in Fig. 2c and 3c. However, in practical applications, consecutive measurement of SSWC data or GWL data is an easier alternative instead of flux and SWH observation.

The results provide us with some interesting practical implications on soil parameter estimation.

Influence of Observation Error

Theoretically, the magnitude of the observation error affects the relative weights of observations vs. predictions in the analysis step. The effect of observation error was analyzed for the heterogeneous soil column. Figures 7 and 8 present the comparison of RMSE values by assimilating the observation data with different errors utilizing the EnRML and EnKF methods, respectively. For both methods, the change in absolute measurement error from 0.02 to 0.05 cm$^3$ cm$^{-3}$ for SSWC data did not significantly affect the final estimates, except that the convergence speed of $K_s$ was faster with a smaller measurement error. As for the GWL data, obviously improved results were obtained with the smaller observation error using the EnRML method (see Fig. 7), while improvement for the EnKF method is not evident (Fig. 8). Regarding the SWH data, a smaller observation error resulted in a relatively faster drop in the RMSE values of the parameters but we did not find distinct differences between the final results with observation errors of 5 and 20% for the EnRML method (Fig. 7), which is similar to the finding of Shi et al. (2015) in a layered soil column. However, EnKF tends to be easily corrupted
by lower measurement accuracy of GWL data (as shown in Fig. 8c), while the output of the EnRML method depends less on the measurement accuracy (as shown in Fig. 7c). In conclusion, the EnKF and EnRML methods can have similar responses to the observation error of SSWC. The EnKF method receives limited benefit from a smaller observation error of GWL data, while the EnRML method is able to apparently improve the parameter estimates with improved observation accuracy. Degraded SWH data may lead to deteriorated parameter estimation if the EnKF is used, while the performance of the EnRML method is less susceptible to measurement errors for SWH data.

**Influence of Ensemble Size**

The ensemble size may influence the outcome of the ensemble-based data assimilation methods. We tested the impact of ensemble size for both EnRML and EnKF methods on parameter estimation. For the heterogeneous soil column, the temporal evolution of parameter RMSE values based on ensemble sizes of 500 and 50 are shown in Fig. 9 and 10 based on the EnRML and EnKF methods, respectively. Generally, better results were acquired using an ensemble size of 500. For the EnKF method, in the cases of assimilating the GWL and SWH data, the results of using 50 ensemble members seem unstable during the late period (see Fig. 10), while the EnRML method with 50 members produced more stable estimates (see Fig. 9). Because the same initial parameter realizations were used for the EnRML and EnKF methods, the results indicate that random sampling error from limited realization size is more pronounced for the EnKF method, and the performance of the EnKF method depends more on the ensemble size. In contrast, the EnRML method is more robust for resisting the influence of sampling error.

An ensemble size of 10 was further investigated for both methods. For the heterogeneous soil column, the generated RMSE values with 10 ensemble members increased dramatically with time and the simulation was corrupt at the late time (results not presented). For the homogeneous soil column, the changes in parameter RMSE values based on ensemble sizes of 500 and 10 are compared in Fig. 11, using the EnRML method (the figure with regard to the EnKF method is not presented due to the corrupted
It can be seen that an ensemble size as small as 10 produces results comparable to an ensemble size of 500 for the homogeneous soil column. Therefore, it seems that there is no need to use a large ensemble size for cases with a homogeneous medium if the EnRML method is used.

**Conclusions and Future Work**

In this study, we evaluated the capability of three data assimilation methods, i.e., EnKF, EnRML, and MCMC, to extract parameter information from three types of observation data, i.e., SWH, GWL, and SSWC. The EnKF and EnRML belong to the filtering algorithms that assimilate observation data sequentially, and MCMC is a sampling approach that utilizes all observation data simultaneously. While all three methods have been explored in some fields, a comprehensive study on assimilating different types of data is rare for nonlinear soil water problems. A homogeneous and a heterogeneous soil column were designed to investigate the performance of the three methods, with 3 and 153 unknown soil parameters to be estimated, respectively. Through a series of synthetic cases, we evaluated the parameter estimation accuracy with respect to several factors, including data assimilation method, data type, observation error, and ensemble size. Based on the results of the study, several conclusions can be drawn:

1. The EnRML method leads to a better estimation than EnKF due to its iterative nature when the data–model relations are significantly nonlinear (SWH and GWL data vs. unknown parameters $\alpha$, $n$, and $K_s$, especially the former two). In contrast, the EnKF and EnRML methods exhibited almost equal ability to extract information from the SSWC data due to its more linear relation with hydraulic parameters. The MCMC method is superior to the EnKF and EnRML methods when estimating parameters of a homogeneous soil. For the heterogeneous soil with many unknown parameters, the MCMC method does not outperform the other two methods, possibly due to an insufficient number of samplings.

2. All three types of data contain useful information on soil hydraulic parameters. Overall, the point-scale SWH data led to the optimal parameter estimation, followed by SSWC data and...
In our test case, the observation error for SSWC seems not to
were limited to SWH, SSWC, and GWL. In a future study, the
can be extended to a more realistic, three-dimen-
conceptual model, in which the flow was assumed to be one-
saturated hydraulic conductivity $K_s$ in the homogeneous soil column.

GW L data. Interpretation of SWH and GWL data requires a
robust algorithm that can handle nonlinear problems.

3. In our test case, the observation error for SSWC seems not to
be an influential factor affecting the performance of the EnKF
or EnRML approaches. Groundwater level data with high
accuracy had an obviously positive influence on the parameter
estimates from EnRML but little influence on the outcome
of EnKF. Compared with EnRML, the EnKF method tends to
degrade with declining measurement accuracy of SWH data.

4. The performance of the EnKF method might depend more sig-
ificantly on the ensemble size than does the EnRML method.
For the homogeneous soil column with few unknown parameters,
only a small ensemble size (i.e., 10) is needed if the EnRML
method is used.

This study focused on a relatively simplified soil water flow
conceptual model, in which the flow was assumed to be one-
dimensional without root water uptake, and the observations
limited to SWH, SSWC, and GWL. In a future study, the
conceptual model can be extended to a more realistic, three-dimen-
sional model, which involves significant lateral flow and complex
soil heterogeneity. Beyond traditional hydrological observations,
geophysical observations can also be included in the current inves-
tigation framework.

Acknowledgments
This study is supported by the National Natural Science Foundation of China
Grants 51861125202 and 51629901. Y. Zha also acknowledges the grants from the National Natural Science Foundation of China (no. 51609173 and
51779179). The constructive comments and suggestions from two anonymous
reviewers are appreciated.

References
closed-eye period: Towards a consistent aggregation of
Camporese, M., C. Paniconi, M. Putti, and P. Salandin. 2009. Ensemble
Kalman filter data assimilation for a process-based catchment scale
doi:10.1029/2008WR007031
doi:10.1007/s10040-004-0404-7
transient and steady state conditions: 1. Maximum likelihood method
doi:10.1029/WR022i002p00199
24:755–769. doi:10.1029/WR024i005p00755
Chen, Y., and D.S. Oliver. 2012. Ensemble randomized maximum likelihood
doi:10.1007/s11004-011-9376-z
Chen, Y., and D.S. Oliver. 2013. Levenberg–Marquardt forms of the
iterative ensemble smoother for efficient history matching
Chen, Y., and D. Zhang. 2006. Data assimilation for transient flow in
Chen, Z., R.S. Govindaraju, and M.L. Kavvas. 1994. Spatial averaging of
unsaturated flow equations under infiltration conditions over areally
et al. 2008. Hydrological data assimilation with the ensemble
Kalm an filter: Use of streamflow observations to update states in a
doi:10.1016/j.advwatres.2008.06.005
Kalman filter versus ensemble smoother for assessing hydraulic
17:1517–1531. doi:10.5194/hess-17-1517-2013
soil brightness temperature imagery into a land surface model
using ensemble Kalman filtering: A case study based on ESTAR
doi:10.1016/S0309-1708(02)00088-X
Interpretation of interference tests in a well field using
geostatistical techniques to fit the permeability distribution in a
reservoir model. In: G. Verly et al., editors, Geostatistics for natural
doi:10.1007/978-94-009-3701-7_16


