Journal of Hydrology 564 (2018) 76–88

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Sequential data-worth analysis coupled with ensemble Kalman filter for soil water flow: A real-world case study

Yakun Wang[a], Liangsheng Shi[a,*], Yuanyuan Zha[a], Xiaomeng Li[a], Qiuru Zhang[a], Ming Ye[b]

[a] State Key Laboratory of Water Resources and Hydropower Engineering Sciences, Wuhan University, Wuhan, Hubei 430072, China
[b] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL 32306, United States

ARTICLE INFO

This manuscript was handled by C. Corradini, Editor-in-Chief, with the assistance of Zhiming Lu, Associate Editor

Keywords:
Data worth
Soil water flow
Data assimilation
Model structural error

ABSTRACT

Given the high cost of data acquisition in soil water problems, it is becoming increasingly essential to collect the measurements as cost-efficient as possible. By introducing the data-worth analysis framework coupled with Ensemble Kalman Filter (EnKF), this real-world case study attempts to assess the worth of potential soil moisture observations before data collection. A field experiment was implemented to demonstrate the feasibility of quantifying the effect of future data on uncertainty reduction under real circumstances in a sequential way. The data worth of future observations is defined regarding soil hydraulic parameter estimation or soil moisture profile retrieval. Four information measures, including *the trace* (T_r), *Shannon entropy difference* (*SD*), *relative entropy* (*RE*) and *degrees of freedom for signal* (*DFS*), are introduced to quantify the information content. The sequential data worth analysis framework is examined by a number of cases, including under different irrigation intensities, with different prior data (existing observations that have already been collected), and with data of various depths and different measurement errors. We demonstrated the ability, and the challenge as well, of quantifying the data worth sequentially. Our results showed that data worth assessment regarding soil moisture profile retrieval is more difficult than that regarding parameter identification. Variance-type and covariance-type metrics have relatively loose accuracy requirement on potential observations (future possible observations to be collected), while mean-covariance-type metrics require higher accuracy. The vertical covariance of soil moisture is susceptible to the effect of atmospheric boundary condition, which eventually imposes a challenge on the quantification of data worth with covariance involved indices. The match between the expected and reference data worth can be improved by assimilating more prior data. However, more prior data cannot compensate for the damage from possible model structural error due to the changed scenarios between the prior stage and the posterior or preposterior stage. Shallow soil moisture data generally has larger data worth than deep observations in our study, but evaluating data worth with shallow data is subject to considerable uncertainty if covariance-type or mean-covariance-type index is employed. Smaller measurement error does not always lead to improved data worth estimation.

1. Introduction

The unsaturated zone is inextricably involved in many aspects of hydrology: evaporation, groundwater recharge, soil moisture storage, and soil erosion (Renard, 1997; Fetter, 2000; Lettau, 1969; Penman, 1948). Plentiful efforts have been made in the past decades on the development of numerical algorithms for modeling unsaturated flow (Celia et al., 1990; Freeze, 1971; Song et al., 2014; Van Genuchten, 1982). These algorithms usually entail characterizing the spatial distribution of soil hydraulic parameters and factors affecting flow. However, it is difficult to directly measure the soil hydraulic properties due to the limitations of measurement techniques and the spatial heterogeneity of parameters. Inverse modeling approaches and data assimilation methods have thus become popular ways to improve soil moisture estimation and derive soil parameters. Numerous inverse and data assimilation methods have been developed (Man et al., 2016a,b; Alcolea et al., 2006; Reichle et al., 2002; Reichle et al., 2008; Rubin et al., 2010; Zimmerman et al., 1998). Thereinto, sequential data assimilation method has gained a significant amount of traction in various communities because of the capability of sequentially incorporating observations into models (Aanonsen et al., 2009; Chen and Zhang, 2006; Oliver et al., 2008; Shi et al., 2015; Xie and Zhang, 2010; Zhu et al., 2017; Hu et al., 2017). In the recent decade, the interest in the real-time integration of measurements has been further boosted by the

* Corresponding author.
E-mail address: liangshs@whu.edu.cn (L. Shi).

https://doi.org/10.1016/j.jhydrol.2018.06.059
Received 2 April 2018; Received in revised form 6 June 2018; Accepted 21 June 2018
Available online 22 June 2018

increasing ability of online data acquisition and the growing demand for real-time prediction and management.

Observation data from different sources has been merged, including remote sensing data (Shi et al., 2011; Crow and Wood, 2003; Montzka et al., 2011; Moradkhani, 2008; Pipunic et al., 2008) and ground-based measurements (Shuwen et al., 2005; Walker et al., 2001; Li and Ren, 2011; Li et al., 2018). It is no doubt that by incorporating more data into the model, the uncertainty of modeling output can be reduced. However, massive observations from multiple sources would significantly increase the budget of data collection. Moreover, the collection process of soil moisture data is subject to considerable uncertainty. Either for remotely-sensed soil moisture data or data collected by the traditional ground method, the identification of data quality or measurement error is not an easy task (Li et al., 2009; Reichle et al., 2008). This can create an extra uncertainty in data assimilation system. Furthermore, the contribution of each data is not identical when multiple data-sources are assimilated together. Some measurements may make significant contribution to the improvement of simulation accuracy while others may be less useful. To avoid the overloaded monitoring cost caused by redundant measurements, it is essential to develop a framework to assess the value of infused data before data acquisition.

Data worth, sometimes called data utility, data information content or data impact, has been defined in several different ways. It was first defined as the reduction in expected error after sampling (Gates and Kisiel, 1974). In a later study of James and Freeze (1993), data worth was assessed by comparing the cost of data collection against the expected value of the risk reduction. In light of modeling uncertainty, statistical measures including the reduction of parameter or prediction variance, relative entropy, and the possible significance level of hypothesis testing can be used to quantify the data worth (Leube et al., 2012). Data worth and analogous concepts have been extensively studied in optimal monitoring network design, which aims to obtain a maximum gain of information from the optimum number, locations, frequency, and data types of measurements (Wu et al., 2005; Kollat and Reed, 2006; Nowak et al., 2010). A few data worth analysis frameworks have been proposed. Freeze et al. (1992) used the search theory and Bayesian updating to determine if additional data have worth or not, in terms of risk reduction in their study. Vrugt et al. (2002) merged the generalized sensitivity analysis, the Bayesian recursive estimation algorithm, and the metropolis algorithm to identify the most informative measurements for model parameter estimation. Leube et al. (2012) developed a preposterior data impact assessor within the Bayesian framework to assess the expected data worth of proposed sampling design. Recently, Dai et al. (2016) proposed a computationally efficient data worth analysis framework based on Ensemble Kalman filter and probabilistic collocation method. The typical data worth framework consists of prior, posterior, and preposterior stages. The preposterior analysis is most critical, and it evaluates whether a future measurement is effective and how effective before it is taken (Neuman et al., 2012; Freeze et al., 1992; James and Gorelick, 1994).

However, the simultaneous data worth analysis coupled with sequential data assimilation methods (i.e., EnKF) received limited attention to date. Zhang et al. (2005) and Dai et al. (2016) employed the static Kalman filter or EnKF to assess the data worth in a groundwater and solute transport problem. Man et al. (2016a) proposed a sequential ensemble-based optimal design method to improve the performance of parameter estimation for soil water problems. Evaluating the data worth sequentially has three main benefits. Firstly, the parameter and model states are updated in real time, so the data impact to modeling system can be detected instantaneously. Secondly, the sampling strategy can be dynamically adjusted to save the monitoring and analysis cost (Man et al., 2016). Thirdly, data-worth analysis frameworks based on sequential data assimilation methods (i.e., EnKF) are strictly non-intrusive and can be embedded into the forward model in a straightforward way.

To the best of our knowledge, all previous studies are based on synthetic cases, and data-worth analysis in the context of dynamically evolving soil parameters and soil moisture profile has not been studied in a real-world case. As stated by Leube et al. (2012), "*for nonlinear problems, the estimation variance and more sophisticated measures of data utility depend on the actual values of measurements, which are still unknown prior to collection*". It will be more convincing to investigate the data worth under real-world circumstances for nonlinear soil water flow. One primary difficulty for practical unsaturated soil water problems is the identification of multiple parameters in the water retention curve and unsaturated hydraulic conductivity. Another difficulty lies in the definition of the upper boundary condition (such as uncertain and uneven rainfall or irrigation), which has a significant influence on shallow soil moisture content (Cull et al., 1981). Hence, this study focuses on exploring the influence of a variety of uncertain soil parameters on data-worth assessment. Three field plots in a greenhouse were thoroughly investigated. Soil parameters were measured by the double-ring test and lab experiments. A period of 65-day continuous soil moisture data under three irrigation schemes was observed. We were thus able to identify the evolution of data worth under different upper boundary settings. The objective of this study is not to minimize the uncertainty of soil moisture profile estimation by specifying the optimal number and location of measurements to be collected but to explore the applicability of running data-worth analysis in a sequential way. We attempt to answer the following questions with the aid of this field-scale experiment: 1) Given multiple uncertain soil parameters, how does the data worth regarding parameter estimation and soil moisture profile retrieval evolve respectively? 2) How do prior data (existing data that have already been collected), observation depth, and measurement error affect the data-worth? 3) What are the performances of different data worth indices under different irrigation schemes? It is hoped that this study can provide a guidance on the design of prospective monitoring strategy for field-scale unsaturated flow problems.

The data worth framework by Dai et al. (2016) is accommodated to this study. Besides the tracer index (T_r) introduced by Neuman et al. (2012), the *Shannon entropy difference* (*SD*) (Shannon, 1949), *relative entropy* (*RE*) (Kullback, 1959) and *degrees of freedom for signal* (*DFS*) (Fisher, 2003) are also introduced to implement the data worth analysis. It is worthy comparing the performance of different indices.

In the following context, Section 2 presents the principles of EnKF and data-worth analysis frameworks and introduces four information metrics, namely T_r, *SD*, *RE*, and *DFS*. Thereafter, Section 3 presents a set of examples to demonstrate the ability of the data-worth analysis framework to quantify the worth of one given monitoring scheme (in terms of irrigation pattern, spatial location, observation error, and prior data content). Moreover, in this section, the discrepancy of four information metrics is explored. In Section 4, conclusions and discussions are presented.

2. Methodology

2.1. Numerical model of one-dimensional unsaturated flow

In this paper, one-dimensional soil water movement is considered. The flow is described by Richards' equation (Richards, 1931):

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

where $\theta[L^3L^{-3}]$ is the volumetric moisture content; h[L] is the pressure head; t[T] is the time; z[L] is the spatial coordinate, oriented positively downward; $K[LT^{-1}]$ is the unsaturated hydraulic conductivity. In Eq. (1), there are three unknown quantities: θ, h, and K. The constitutive relationship between them can be characterized by the van Genuchten-Mualem model:

$$\theta h = \begin{cases} \theta_r + \frac{\theta_s - \theta_r}{(1+|\alpha h|^n)^m} & h < 0 \\ \theta_s & h \geqslant 0 \end{cases} \tag{2}$$

$$Kh = \begin{cases} K_s S_e^{1/2}\left[1-(1-S_e^{\frac{1}{m}})^m\right]^2 & h < 0 \\ K_s & h \geqslant 0 \end{cases} \tag{3}$$

$$m = 1-\frac{1}{n},\ n > 1 \tag{4}$$

$$S_e = \frac{\theta-\theta_r}{\theta_s-\theta_r} \tag{5}$$

where $\theta_s[L^3L^{-3}]$ and $\theta_r[L^3L^{-3}]$ are the saturated and residual moisture content, respectively; $\alpha[L^{-1}]$ and n[dimensionless] are the shape parameters of the soil water characteristic curve; $K_s[LT^{-1}]$ is the saturated hydraulic conductivity; S_e is the effective saturation. In our paper, we employ the Ross method (Ross, 2006, 2003), which is a noniterative numerical scheme, to obtain a fast solution of one-dimensional Richards' equation. The specific advantages of Ross method can be found in Zha et al. (2013).

2.2. Ensemble Kalman filter (EnKF)

In EnKF, the parameter vector of interest $\boldsymbol{p}$ are augmented with the state variable vector $\boldsymbol{s}$ into a joint state vector $\boldsymbol{y} = [\boldsymbol{p}, \boldsymbol{s}]^T$. A collection of N_1 members of the state vector $\boldsymbol{Y}$ can be written as

$$Y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ ..., \boldsymbol{y}_{N_1}\} \tag{6}$$

The relationship between observations at time step t, $\boldsymbol{d}_t^{obs}$ and their true values $\boldsymbol{y}_t^{true}$ is represented as follows,

$$\boldsymbol{d}_t^{obs} = \boldsymbol{H}\boldsymbol{y}_t^{true} + \varepsilon_t \tag{7}$$

where matrix $\boldsymbol{H}$ is the observation operator which relates the state and observation vectors. $\boldsymbol{\varepsilon}_t$ represents measurement error vector which is assumed to be zero-mean Gaussian with covariance matrix $\boldsymbol{R}_t$.

As a sequential Monte Carlo method, EnKF entails two main steps: forecast step and analysis step. At the forecast step, each state vector in $\boldsymbol{Y}$ is projected from time step $(t-1)$ to time t through the forward numerical/analytical model $\boldsymbol{F}$

$$\boldsymbol{y}_{i,t}^f = \boldsymbol{F}(\boldsymbol{y}_{i,t-1}^a),\ i = 1, 2...N_1 \tag{8}$$

where superscripts f and a refer to forecast and analysis, respectively. Superscripts i is the ensemble member index.

At the analysis step, for any ensemble member i at a given time t, the joint state vector is updated by combining model predictions and observations

$$\boldsymbol{y}_{i,t}^a = \boldsymbol{y}_{i,t}^f + \boldsymbol{K}_t(\boldsymbol{d}_{i,t}^{obs} - \boldsymbol{H}\boldsymbol{y}_{i,t}^f) \tag{9}$$

where the Kalman gain $\boldsymbol{K}_t$ is defined as

$$\boldsymbol{K}_t = \boldsymbol{C}_t^f \boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{C}_t^f\boldsymbol{H}^T + \boldsymbol{R}_t)^{-1} \tag{10}$$

where $\boldsymbol{C}_t^f$ is the covariance matrix of the state vector at time t, which can be updated after data assimilation through

$$\boldsymbol{C}_t^a = (\mathbf{I} - \boldsymbol{K}_t\boldsymbol{H})\boldsymbol{C}_t^f \tag{11}$$

where $\mathbf{I}$ is a unit matrix whose dimension is $N_d \times N_d$; and N_d is the number of available measurements.

2.3. Data-worth analysis framework coupled with EnKF

Following the framework of Neuman et al. (2012) and Dai et al. (2016), data worth analysis of future monitoring scheme is comprised of three stages:

(1) At the prior stage, EnKF with an ensemble size of N_1 is employed to sequentially assimilate the prior data $\boldsymbol{A}$. A set of N_1 hypothetical observations are then generated with $\boldsymbol{B}_i = \boldsymbol{H}\boldsymbol{y}_i^f + \varepsilon_i$, $i = 1, 2, ..., N_1$. It is worth noting that N_1 should be large enough to include as many as possible measurement values and assure the accuracy of estimated mean and covariance.
(2) At the preposterior stage, for each possible data $\boldsymbol{B}_i$, N_2 realizations satisfying Gaussian distribution(the ensemble mean is the value of $\boldsymbol{B}_i$, and the variance is the measurement error variance) are firstly generated. EnKF with an ensemble size of N_2 is again implemented to estimate the resulting mean and other statistics. There are totally $N_1 \times N_2$ realizations being run. Predictive statistics of posterior vector $\boldsymbol{y}_{i,j}$ ($i = 1, 2, ..., N_1$; $j = 1, 2, ..., N_2$), i.e. $E(\boldsymbol{Y}_i|\boldsymbol{A}, \boldsymbol{B}_i)$ and $Cov(\boldsymbol{Y}_i|\boldsymbol{A}, \boldsymbol{B}_i)$, are calculated by jointly conditioning on $\{\boldsymbol{A}, \boldsymbol{B}_i\}$. Then $E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$, $E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$, and $Cov_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$ are yielded through statistics over all $\boldsymbol{B}_i$. Similar procedures repeat until the final time. In the meantime, the expected data worth in the form of quantitative indices is calculated.
(3) At the posterior stage, the actual mean and covariance are obtained by using available real data set $\boldsymbol{B}'$ in a sequential manner. The real (i.e. reference or posterior) data worth and expected (preposterior) data worth is then compared to reveal the effectiveness of this data worth framework. The whole workflow of data-worth analysis framework is depicted in Fig. 1.

The framework described above is essentially a two-layer hierarchical Bayesian model. The first layer considers the possible variation range of potential data (observations to be collected in the future), and the second layer contains the uncertainty from uncertain parameters. By quantifying the contained uncertainty (and sometimes the resulting change of mean behavior for a nonlinear system) in the first layer, the expected data worth for a given observation scheme (i.e., for given number, locations, frequency, and data types of measurements) is then determined. In this study, we introduce four information metrics to quantify the worth of potential observations.

Firstly, the statistics in the prior and preposterior analysis have the following theoretical relations (Neuman et al., 2012):

$$E(\boldsymbol{Y}|\boldsymbol{A}) = E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B}) \tag{12}$$

$$Cov(\boldsymbol{Y}|\boldsymbol{A}) = E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B}) + Cov_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B}) \tag{13}$$

where $E(\boldsymbol{Y}|\boldsymbol{A})$ is the prior mean of state vector $\boldsymbol{Y}$, conditioning on prior

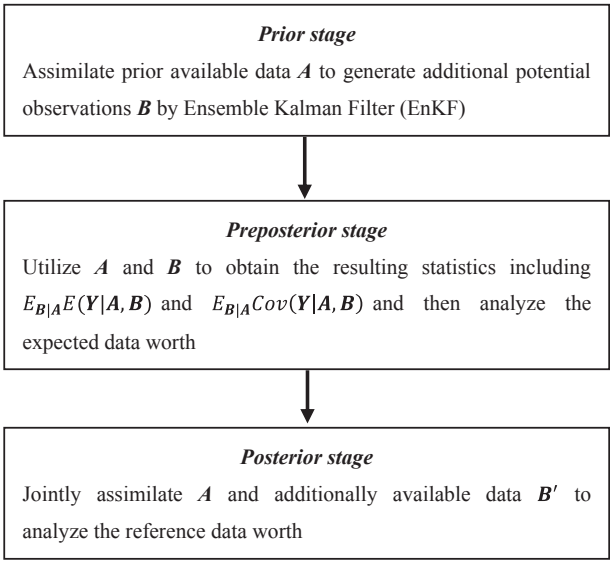

Fig. 1. The workflow of Bayesian data worth analysis framework coupled with Ensemble Kalman Filter (EnKF).

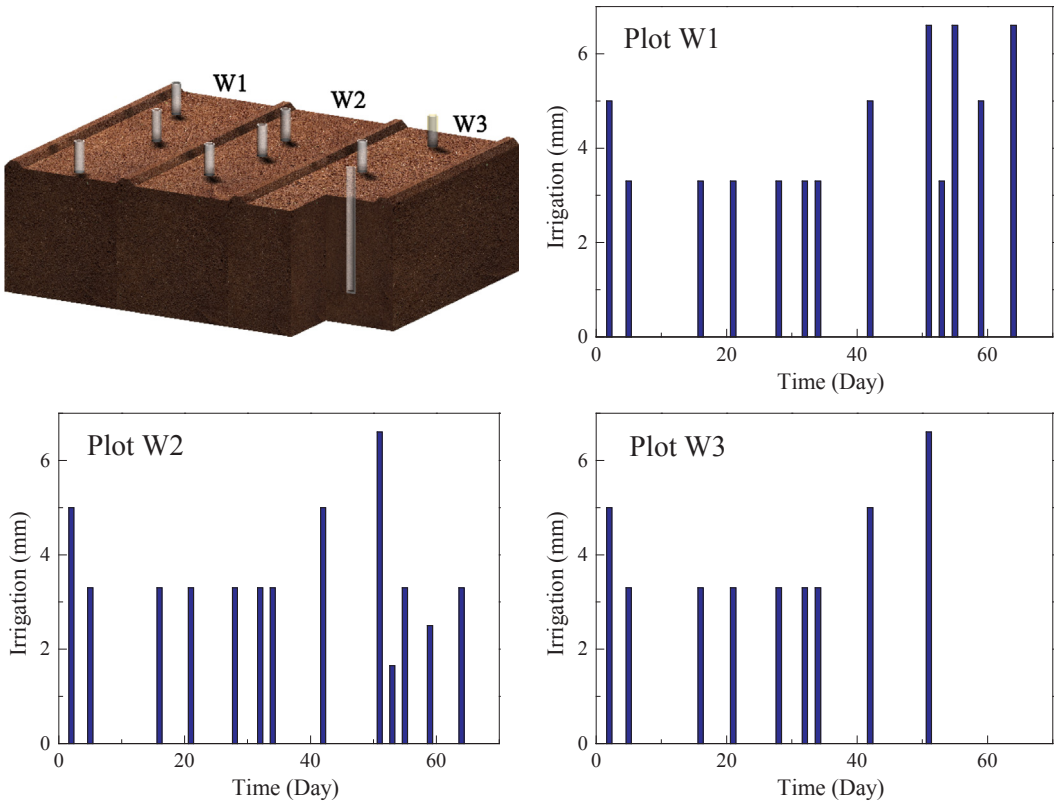

Fig. 2. The layout of experimental plots and irrigation amount versus time in plots W1, W2, and W3.

data $\boldsymbol{A}$. $E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ is the expectation of $E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$. For the sake of simplicity, $E(\boldsymbol{Y}|\boldsymbol{A})$ and $E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ are denoted by $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$, respectively. $Cov(\boldsymbol{Y}|\boldsymbol{A})$, denoted as $\boldsymbol{C}_1$, represents prior predictive uncertainty. Given that potential observation $\boldsymbol{B}$ is generated by conditioning on $\{\boldsymbol{A}\}$, $\boldsymbol{C}_1$ is theoretically equal to $Cov(\boldsymbol{Y}|\boldsymbol{A})$ at the preposterior stage. However, in fact, they may differ from each other due to the sampling error caused by the finite sample size. Moreover, $E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ is the expectation of $Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ conditioning on $\{\boldsymbol{A},\boldsymbol{B}\}$, which represents the predictive uncertainty in preposterior data-worth analysis. Here, it is denoted by $\boldsymbol{C}_2$. At the early time of data worth analysis procedure, the total uncertainty of the system mostly lies in the second layer, i.e. $Cov(\boldsymbol{Y}|\boldsymbol{A}) \approx E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ while $Cov_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}) \approx 0$. With the sequential infusion of potential observations at the following time steps, $Cov_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B})$ starts to increase. In addition, when calculating the reference data worth, $\boldsymbol{C}_2$ refers to $Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')and\boldsymbol{M}_2$ refers to $E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')$.

(1) *Trace*

A scalar indicator *trace* (T_r) is defined by Neuman et al. (2012), as a measure of data worth in terms of uncertainty reduction

$$T_r = T_r(\boldsymbol{C}_1) - T_r(\boldsymbol{C}_2) \tag{14}$$

where T_r indicates the trace (sum of diagonal entries) of a matrix.

(2) ***The Shannon entropy difference*** (*SD*) between the prior and preposterior probability density functions (pdfs) also can be used to quantify the information content extracted from future data. Due to the fact that the prior and preposterior pdfs are both Gaussian in EnKF, *SD* can be expressed as (Shannon, 1949; Xu, 2007),

$$SD = \frac{\ln\det(\boldsymbol{C}_1)}{2} - \frac{\ln\det(\boldsymbol{C}_2)}{2} = \ln\det(\boldsymbol{C}_1\boldsymbol{C}_2^{-1})/2 \tag{15}$$

where det $(*)$ denotes the determinant.

(3) ***The relative entropy*** *(RE)*, known as a signal-dispersion combined index, provides a measure of the information content of an analysis (posterior) pdf with respect to background (prior) pdf (Zhang et al., 2015; Singh et al., 2013). Under the assumption that the background and the analysis pdfs are *n*-dimensional Gaussian, *RE* is defined as (Xu, 2007),

$$RE = J_b + DP \tag{16}$$

$$J_b = (\boldsymbol{M}_2 - \boldsymbol{M}_1)^{\mathrm{T}}\boldsymbol{C}_1^{-1}(\boldsymbol{M}_2 - \boldsymbol{M}_1)/2 \tag{17}$$

$$DP = [\ln\det(\boldsymbol{C}_1\boldsymbol{C}_2^{-1}) + T_r(\boldsymbol{C}_2\boldsymbol{C}_1^{-1}) - n]/2 \tag{18}$$

Unlike *SD* and T_r, the *relative entropy* (*RE*) is not additive (Xu, 2007). The advantage of *RE* index is the ability of measuring both the signal (mean) part (J_b) and dispersion (covariance) part (*DP*), which may be important for nonlinear soil water system. From Eqs (9), (16), (17), and (18), it is emphasized that *RE* not only depends on the observation operator but also depends on observation vector itself. In particular, the signal part J_b mainly depends on the realizations of $(\boldsymbol{d}_{i,t}^{obs} - \boldsymbol{H}_t\boldsymbol{y}_{i,t}^{f})$. For each realization of $\boldsymbol{d}_{i,t}^{obs}$, the signal and dispersion components indicate the improvements made from the mean and covariance, respectively (Xu et al., 2009). The calculation process of J_b implies the influence of actual values of measurements on data worth assessment.

(4) ***The degrees of freedom for signal*** (*DFS*) measures how many degrees of freedom of an observation are related to signal (versus noise) (Xu et al., 2009). It characterizes the total reduction in variance, or the reduction in the number of degrees of freedom of the error resulting from the addition of observations (Singh et al., 2013). *DFS* can be defined as,

$$DFS \equiv \langle 2J_b \rangle \tag{19}$$

where $*$ represents the expectation.

For a linear (or linearized) observation operator, the definition of *DFS* can be transformed into:

$$DFS = n - T_r(\boldsymbol{C}_2\boldsymbol{C}_1^{-1}) \tag{20}$$

In essence, these indices can be divided into three categories: variance-type (T_r), covariance-type (*SD* and *DFS* (Eq. (20)), and mean-

covariance-type (*RE* and *DFS* (Eq. (19)). Among these indices, *RE* (J_b part) and *DFS* (Eq. (19)) consider the effect of actual measurement data values on data worth, while other indices merely include the influence of uncertainty (variance and/or covariance). It is noted that the implementation of above data worth analysis framework is computationally expensive, thus parallel computing is utilized to alleviate this issue. Efficient surrogate system can be further combined to reduce the computational burden (Man et al., 2017; Dai et al., 2016).

3. Experiment and modeling results

3.1. Field experiment

A field experiment was conducted in the Irrigation and Drainage Laboratory of Wuhan University, China. The experimental period was between 29 February and 3 May 2016. There were totally three experimental plots selected in the greenhouse. All plots were bare soil with no vegetation cover. Three plots were named as W1, W2, and W3, respectively. Each plot had a size of 3 m by 1 m, as shown in Fig. 2. Within each plot, three TRIME (TRIME-PTCO-IPH) tubes were installed randomly in space. Average soil water content at multiple depth ranges (0.02–0.18 m, 0.22–0.38 m, 0.42–0.58 m, and 0.62–0.78 m) were measured. The average soil water content of three tubes in each plot were regarded as the soil water content of this plot. The measurement time interval was 3–4 days.

The amount of irrigation water for plots W1, W2, and W3 are given by Fig. 2. During the first 51 days, the irrigation amount of three plots were the same, while during the next 14 days, the irrigation amount at each irrigation time in W2 was half of that in W1 and there was no irrigation in W3. In W1 and W2, the irrigation frequency doubled during the last 14 days. The late stage of W2 was designed to investigate the data worth response to changed irrigation frequency, while the simultaneous effect of changed irrigation frequency and amount was explored in W1, and the data worth response to no irrigation was analyzed in W3.

16 groups of undisturbed soil samples and 3 groups of packed soil samples were collected to measure the soil hydraulic parameters. The saturated moisture content θ_s was determined as 0.44 by using oven drying method. The other three soil parameters (θ_r, α and n) were determined by using the nonlinear *isqcurvefit* function in MATLAB. Their values were equal to 0.02, 1.48(m^{-1}), and 1.27, respectively. In addition, soil saturated hydraulic conductivity K_s was measured by using the double-ring infiltrometer experiment at 10 random sites within the study field, leading to an average value of 0.068 m/d.

3.2. Model setup

The key parameters of the numerical cases are listed in Table 1. Focusing on the vertical soil water flow, each plot is represented by a one-dimensional soil column. The soil column has a height of 0.8 m and is discretized with 80 grids of uniform size. The top and bottom boundaries are set as an atmospheric boundary and constant water content respectively. In addition, K_s, α and n are assumed to be unknown and lognormally distributed during data worth analysis.

Table 1
Summary of key model parameters.

Parameters	Value
Description of soil column	
Soil column height [m]	0.8
No. of nodes	81
Simulation time [d]	65
Number of realizations	
$\boldsymbol{N}_1$	200
$\boldsymbol{N}_2$	200
Distribution of initial soil parameter (state)realizations	
$\ln\alpha[\mathbf{m}^{-1}]$	$N(1.3451, 0.5890^2)$
$\ln\boldsymbol{n}$	$N(0.4779, 0.0625^2)$
$\ln\boldsymbol{K_s}$[m/d]	$N(-0.4194, 0.8294^2)$
Variance of initial soil moisture realizations	0.03^2

Table 2
Summary of designed test cases and main characteristics.

Case	Prior data (d)	Monitoring depth (m)	Observation error	Description
C1	19	0.02–0.18	0.0005	Base case
C2	30	0.02–0.18	0.0005	Prior data content
C3	9	0.02–0.18	0.0005	Prior data content
C4	19	0.22–0.38	0.0005	Monitoring depth
C5	19	0.42–0.58	0.0005	Monitoring depth
C6	19	0.02–0.18	0.0001	Observation error
C7	19	0.02–0.18	0.001	Observation error

A suite of cases (Table 2) is designed to serve our research purposes. Case C1 is set as the base case to verify the feasibility of assessing the data worth within the data-worth analysis framework coupled with EnKF. The prior data entering into case C1 comprises of soil water content measurements taken at diverse depths (i.e. 0.02–0.18 m, 0.22–0.38 m, 0.42–0.58 m, and 0.62–0.78 m) during the first 19 days.

A comparison among cases C1, C2, and C3 is designed to investigate the effect of prior data content on data-worth analysis, which can help us to determine the required prior information content to ensure the accuracy of data-worth assessment. Cases C2 and C3 have 30-day and 9-day prior soil moisture data respectively, differing from 19-day data in case C1.

Many previous studies attempted to explore the potential of using surface soil moisture data, which are most readily available, to retrieve soil moisture profile (Li et al., 2010; Calvet and Noilhan, 2000; Das and Mohanty, 2006) and soil hydraulic properties (Montzka et al., 2011; Shi et al., 2015). However, in the cases where moisture content observations in the topsoil and deep soil are both available without considering technique constraints, is the surface soil moisture more valuable? Therefore, two more test cases (i.e. C4 and C5) are designed to investigate the effect of monitoring depths on data-worth analysis. Test cases C4 and C5 are identical to C1 except that the additional observations are made at the depths of 0.22–0.38 m and 0.42–0.58 m, respectively.

Considering that observed data is always accompanied with error, it is interesting to explore the discrepancy of data-worth under different levels of observation errors. Here, we assume that the measurement errors are Gaussian-distributed and have a constant variance. The measurement error variance of soil moisture for C1 is assumed to be 0.0005, to be compared against the values of 0.0001 and 0.001 in cases C6 and C7, respectively.

3.3. Results and discussions

3.3.1. Validation of data-worth analysis method coupled with EnKF (case C1)

In this case, 19-day prior soil moisture data is provided to generate potential soil moisture at the depth of 0.02–0.18 m. Fig. 3b(1–3) presents a comparison of the real observations (red squares) and the corresponding $N_1 = 200$ potential observation samples (black dots) as well as their ensemble mean (black squares) in all plots. It is seen that the mean values of potential observation realizations show some deviations from the actual measurements with 19-day prior data. With the increase of prior data, the generated potential data are expected to approach to the actual observations, as shown in Fig. 3c (with 30-day prior data). The unsatisfactory performance under inadequate prior data can be attributed to the following reasons: (1) insufficient prior data may result in parameter estimates not close enough to the 'true' values after data assimilation. Fig. 4(a-c) plots the updated mean of $\ln\alpha$, $\ln n$, and

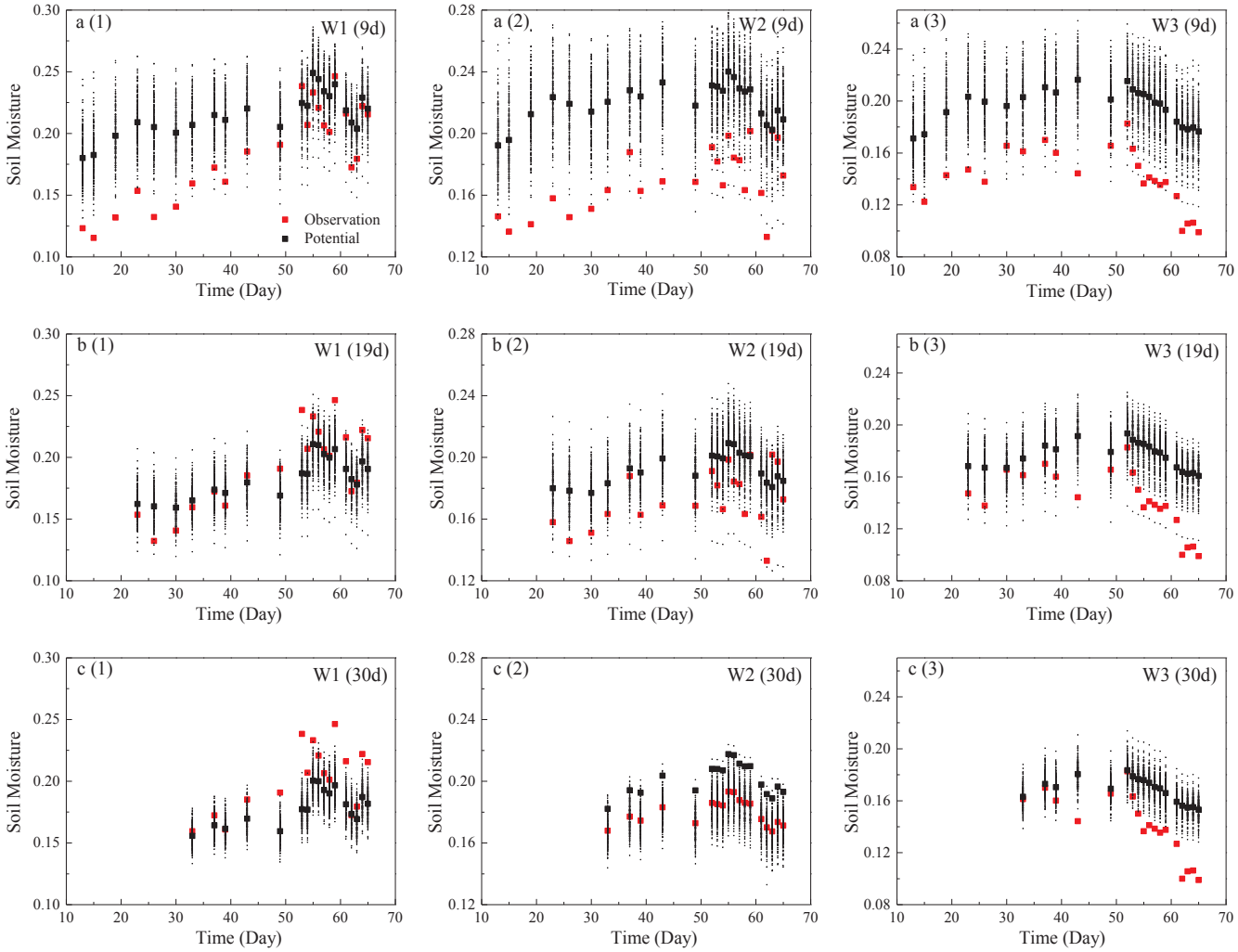

Fig. 3. A comparison of real observations at the depth of 0.02–0.18 m (red squares) and the corresponding potential observation realizations (black dots) as well as the ensemble mean (black squares) in plots W1, W2, and W3, with 9-day (a), 19-day (b),and 30-day (c) prior data respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

lnK_s versus time by assimilating all soil moisture observations sequentially. It is seen that within 19-day prior period, the estimated parameters (especially for ln K_s) have not reached stable values for all plots. (2) The change of future scenario (irrigation amount and frequency) may trigger the potential model structural error (Wagener and Gupta, 2005; Xu et al., 2017). This is particularly obvious in plot W3 with no irrigation at the late time. If one model undergoes different and contrasting scenarios, the model structural error is likely to appear since model parameters updated or calibrated under one scenario have not been examined under another scenario. This can be confirmed by the fluctuated parameter estimates in all plots after 50th day, when the irrigation amount and frequency start to change (Fig. 4). Inaccurate description to forcing term (such as evaporation) or a simplified representation of real soil layer structure (Xu and Valocchi, 2016; Beven, 2005; Crow and Van Loon, 2005; Dee, 2005; Mccuen, 1974) may also introduce model structural error. This potential error eventually increases the difficulty of accurately 'imitating' real observations with potential data.

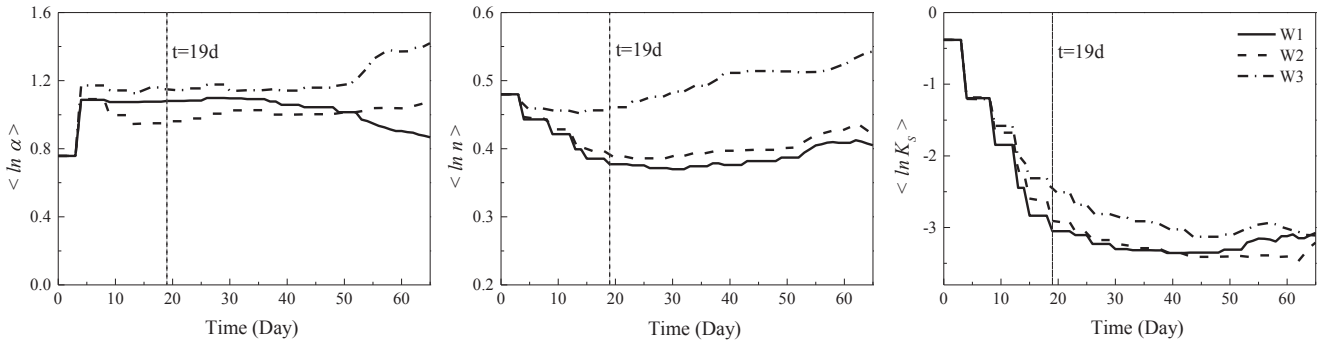

Fig. 4. The temporal change of updated mean ln α, ln n and ln K_s by assimilating all actual observations sequentially in plots W1, W2, and W3.

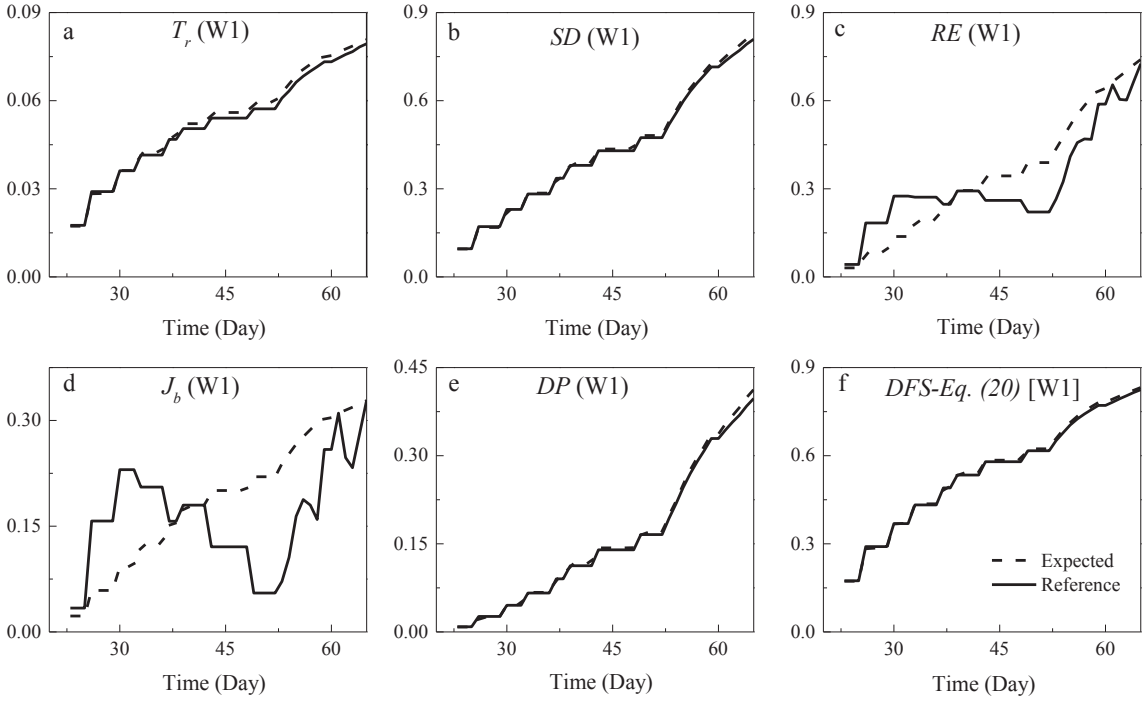

Fig. 5. A comparison of the expected (dash line) and reference (solid line) data worth regarding parameter identification in the form of T_r, *SD*, *RE*, J_b, *DP*, and *DFS* in plot W1.

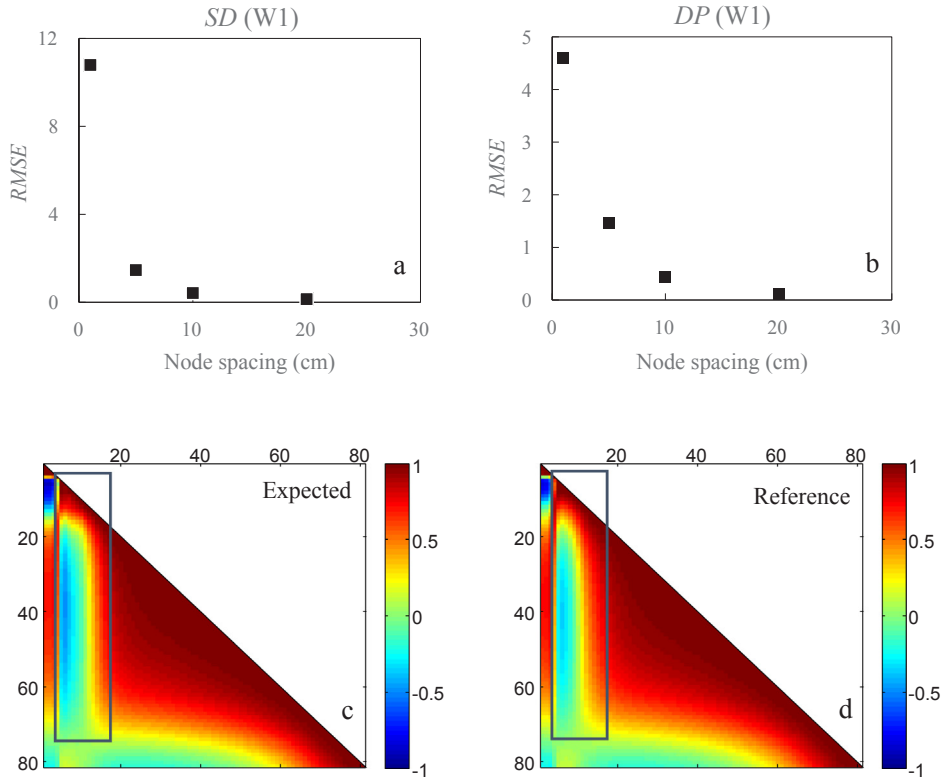

Fig. 6. The root-mean-square error (*RMSE*) between the expected and reference *SD* (a) and *DP* (b) under different node spacings; A comparison of the expected (c) and reference (d) soil moisture covariance matrix at the 65th day in plot W1.

3.3.2. Parameter estimation

Based on the above potential and actual observations, their expected and reference data worth regarding parameter estimation can be quantified in the form of T_r, *SD*, *RE*($= J_b + DP$), and *DFS* (Eq. (20)), as depicted in Fig. 5. Only the results in plot W1 are presented. On the whole, the results lead to two findings: (1) for variance-type (T_r) or covariance-type (*SD*, *DP* and *DFS*(Eq.(20))) indicators, the expected and reference data worth values match very well even though the potential observations and real data do not fit quite well. We recall that such an agreeable match is achieved even if the means of the potential

observations still deviate from the actual observations, as shown in Fig. 3; (2) for *RE* (or J_b), the expected value barely reproduces its reference counterpart. This is because that mean-covariance-type metric further takes into account the effects of mean behavior ($M_2 - M_1$), which is barely reproducible with high accuracy due to the intrinsic nonlinearity of the soil water system (Yeh et al., 1985; Mantoglou and Gelhar, 1987) and the possible model structural error. These results imply that a satisfactory evaluation of data worth with mean-covariance-type metrics places high demands on the potential observations (especially the mean of potential observations) for soil water problems under the real-world circumstance, while the variance-type and covariance-type metrics relatively relax the accuracy requirement to the potential observations but focus on the uncertainty level.

According to Eq. (19) and Eq. (20), *DFS* can be calculated in two ways, i.e. the original definition (similar to J_b, shown in Fig. 5d) and linearized transformed definition (shown in Fig. 5f). Although the linearized form of *DFS* leads to a better match between the expected and reference data worth, it may not be applicable in soil water problems due to the inherent nonlinearity. The results show that such linearization can introduce artificial distortion to data worth analysis, although linearized *DFS* has been popularly used by previous studies (Fisher, 2003; Xu, 2007).

3.3.3. Soil moisture profile retrieval

When we evaluate the data worth of potential measurements on soil moisture profile retrieval, it is requisite to consider at which resolution the soil moisture profile should be reproduced. We run data worth analysis with different numbers of nodes and investigate the influence of meshing size on data worth analysis. Surprisingly, the results show that meshing size has a very large effect on the calculated data worth in terms of *SD* and *DP*. Fig. 6(a) and (b) present the *RMSE* between the expected and reference *SD* and *DP* with respect to different node spacings. It is observed that the *RMSE* for *SD* or *DP* increases drastically when the node spacing is small. Fig. 6(c) and (d) further compare the expected and reference covariance matrixes at the 65th day with a node spacing of 1 cm. The apparent contrast between the expected and reference covariance exists for the cross-covariance between the shallow and deep nodes (highlighted by gray rectangle). This cross-covariance bias is caused by the spurious correlation due to numerical calculation. Further results show that the variance can be estimated relatively easily while a satisfactory cross-covariance estimation is difficult (which can be inferred by comparing Fig. 7a and b). Since covariance-type metrics depend on both the variance (the diagonal elements of the covariance matrix) and the cross-correlation (the non-diagonal elements), the results demonstrate the challenge of data worth analysis for soil moisture prediction if the covariance-type index is employed. There are two solutions for this issue. One is to use a larger node spacing. A larger node spacing seems helpful to alleviate the adverse effect of spurious correlation (see Fig. 7a and b) but at the same time it may lead to false cross-covariance and variance. Another way is to use the localization technique to avoid the spurious cross-covariance.

In our study, only soil moisture at several representative nodes, i.e. at the depths of 10 cm, 30 cm, 50 cm, and 70 cm, are selected to evaluate the data worth. The results of data worth are given in Fig. 7. The expected and reference *SD* matches constantly well. The expected and reference *SD* (and *DFS*) fit well during the early period but start to deteriorate at the late time. Meanwhile, the comparison of Fig. 5 and Fig. 7 reveals an inferior data worth analysis performance for soil moisture retrieval when the covariance-type index is used, compared with data worth analysis for parameter estimation. To understand this phenomenon, the expected and reference posterior covariance matrices at the 30th and 65th day are displayed in Fig. 8. It is found that compared to parameter covariance, the match of expected and reference soil moisture covariance deteriorates obviously at the 65th day. This can explain the deteriorated match of *SD* (or *DFS*) over time for soil moisture retrieval.

One unfavorable aspect regarding data worth analysis for soil moisture retrieval is that the vertical covariance of soil moisture is susceptible to the upper boundary condition. In plot W1, frequent irrigation is applied during the late period, and the intensified drying and wetting processes seem difficult to be described accurately by the updated soil parameters (from the prior period), which consequently brings difficulty of reproducing the real cross-covariance. Thus, the future change of the atmospheric scenario should be paid enough attention, not only due to the possibly introduced model structural error but also due to its vital effect on soil moisture covariance estimation.

3.3.4. Effect of prior data content (C1, C2, and C3)

Fig. 9(a–f) compares the data worth among cases C3, C1, and C2, with 9-day, 19-day and 30-day prior data respectively, with respect to parameter identification. To avoid repetition, only the results of T_r, *SD* and *RE* (representing variance-type, covariance-type, and mean-covariance-type index, respectively) in plots W1 and W2 are shown. Overall, the expected and reference data worth has a better match as more prior data are assimilated, especially for T_r and *SD*. This is in line with the fact that actual observations are more readily 'captured' with potential observation realizations as prior data increases, as depicted in Fig. 3. It is also seen from Fig. 9(a–c) (or Fig. 9(d–f)) that in comparison to mean-covariance-type indicator (*RE*), the fitness between the expected and reference variance-type and covariance-type indicators (T_r and *SD*) are less sensitive to prior data. Different responses of *RE* to prior data are observed in plots W1 and W2. More prior data leads to the improved estimation of *RE* in plot W2, while increasing the prior period from 19 days to 30 days unfortunately worsens the data worth match after the 50th day in plot W1. Considering that W1 was applied with more intensive irrigation after the 50th day, it is found that more prior data does not always lead to better data worth estimation (Fig. 9(c)). In plot W1, the abnormal pattern of *RE* at the late time is possibly caused by more intensified wetting–drying cycle of moisture content than in plot W2, which eventually increases the difficulty of obtaining reasonable parameters to represent the real soil. Thus, on one side, data worth assessment regarding parameter identification can be improved by assimilating more prior data (Xue et al., 2014) if T_r and *SD* are used. On the other side, the deteriorated performance of *RE* in plot W1 reminds us that poor fitting of mean-covariance-type index due to imperfect parameter estimates may hardly be compensated by assimilating more prior data. We note that model structural error and imperfect parameter estimate actually appear simultaneously since parameters are forced to compensate for the bias between observation and simulation if sequential data assimilation is employed.

Similar conclusions can be drawn for soil moisture profile retrieval. Fig. 9(g) presents the *SD* in plot W2. As suggested by comparing Fig. 9(e) and Fig. 9(g), some interesting phenomena can be found: for parameter identification, 19-day prior data is adequate to guarantee the estimation accuracy of *SD*; nevertheless, the accuracy of *SD* regarding soil moisture prediction is satisfactory only when the prior data period reaches 30 days. Such a difference again indicates that more prior information is required to assess data worth for soil moisture profile characterization, compared with soil parameter identification.

3.3.5. Effect of monitoring location (C1, C4, and C5)

Since shallow soil moisture is more sensitive to the atmospheric boundary and shows larger temporal variation than deep soil moisture in our field experiment, it is expected that shallow soil moisture is subject to higher uncertainty and hence observations at shallow depths generate larger data worth (T_r and *SD*). As demonstrated in Fig. 10(a), (b),(d) and (e), assimilating deep soil moisture brings reduced data worth, regardless of parameter identification or soil moisture profile retrieval. This result is consistent with the finding of Dai et al. (2016), which claimed that an effective observation strategy is to acquire new data at a location where the prior predictive uncertainty is large.

However, the response of relative entropy to different depths of soil

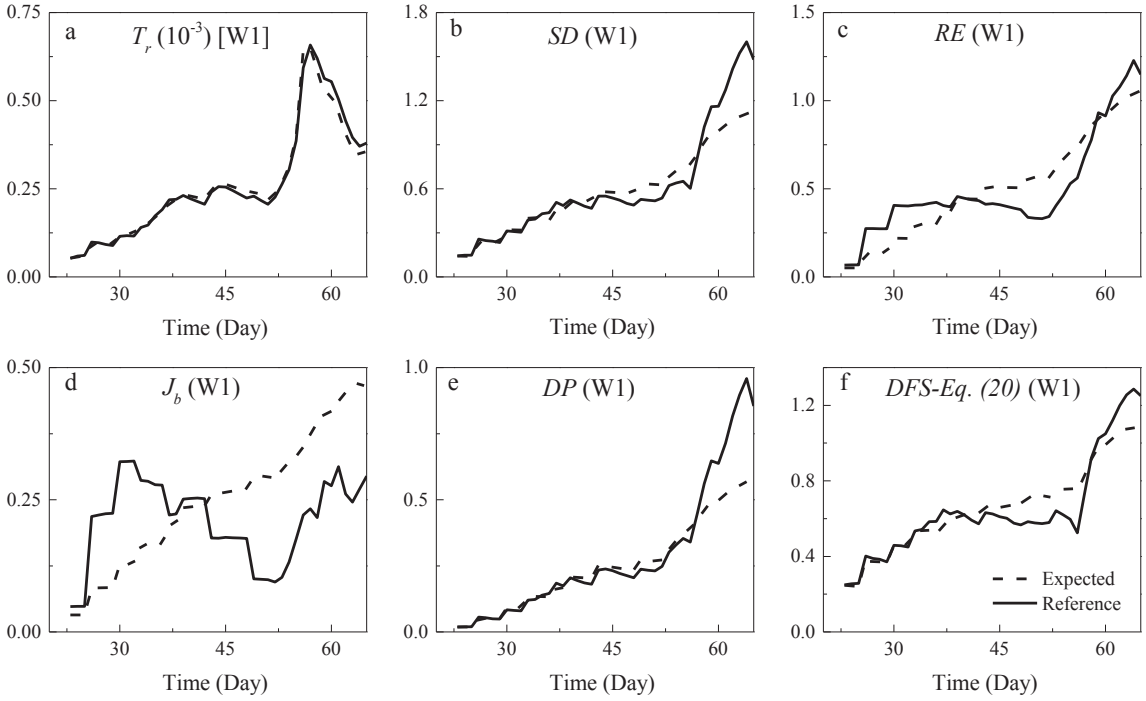

Fig. 7. A comparison of the expected (dash line) and reference (solid line) data worth regarding soil moisture profile retrieval in the form of T_r, *SD*, *RE*, J_b, *DP*, and *DFS* in plot W1.

moisture is rather complicated, especially at the late time when the irrigation frequency was intensified in plot W2. A few phenomena can be observed: (1) Regardless of the observation depth, overall, all expected values of data worth indices deviate from their reference values more and more as time goes on; (2) The moisture data has reduced reference data worth with increased observation depth during the early period (0–50 day), but some unusual patterns can be seen. For example, the reference relative entropy based on the observation at 0.22–0.38 m depth shows abnormal surge during the late period; (3) Despite of the larger data worth from shallow observations, for soil moisture estimation, *SD* and *RE* do not perform well in term of the matching degree of reference and expected data worth if shallow moisture data at the depths of 0.02–0.18 m and 0.22–0.38 m are collected. The reason could be two-sided: the potential observations of shallow depth may suffer more uncertainty sources (for example, spatial or/and temporal variability of surface soil parameters, uncertain atmospheric boundaries, and inaccurate evapotranspiration model); the measurement error of surface moisture data may also receive higher uncertainty. Our results imply that evaluating the data worth of shallow soil moisture observation is subject to considerable challenge if covariance-type or mean-covariance-type index is employed. Similar findings regarding soil moisture data assimilation have been reported by Crow and Van Loon (2005) and Shi et al. (2015); (4) Comparing to T_r and *SD*, the match of reference and expected *RE* is especially not agreeable, which has also been observed in Figs. 5, 7 and 9. Our further analysis showed that the accelerating discrepancy between the reference and expected *RE* is because J_b is the dominant component and there exists a large difference between the reference and expected J_b (figures not presented). These results further emphasize that the influence of atmospheric boundary conditions should be taken into account when determining the future monitoring strategy (location, frequency, and data type) with mean-covariance-type index.

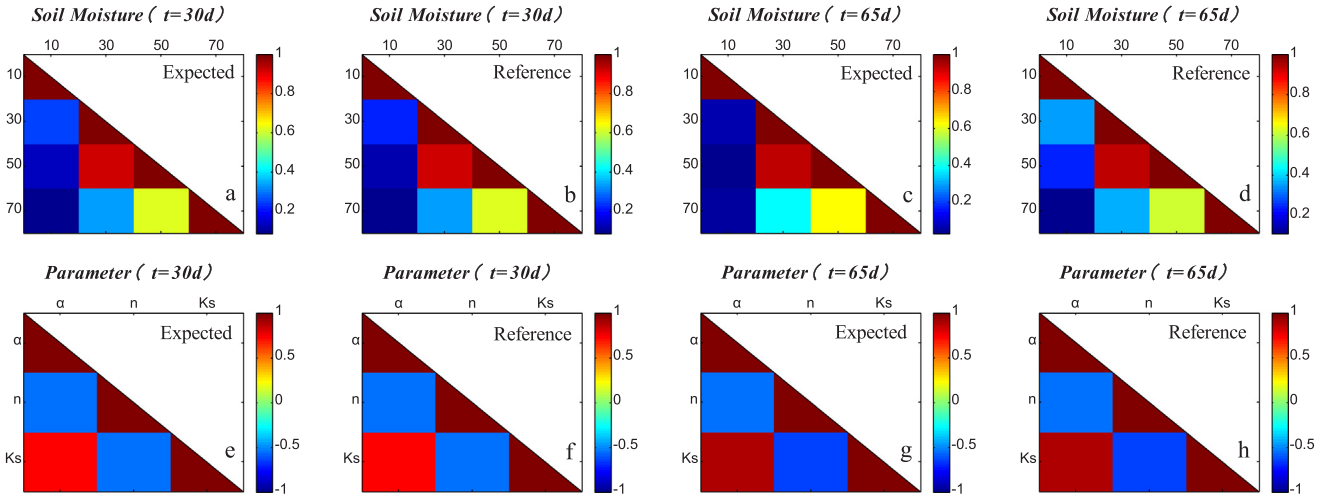

Fig. 8. A comparison of the expected and reference covariance matrices at the 30th and 65th day regarding soil moisture profile retrieval (based on soil moisture at four selected depths, i.e. 10 cm, 30 cm, 50 cm and 70 cm) and parameter identification.

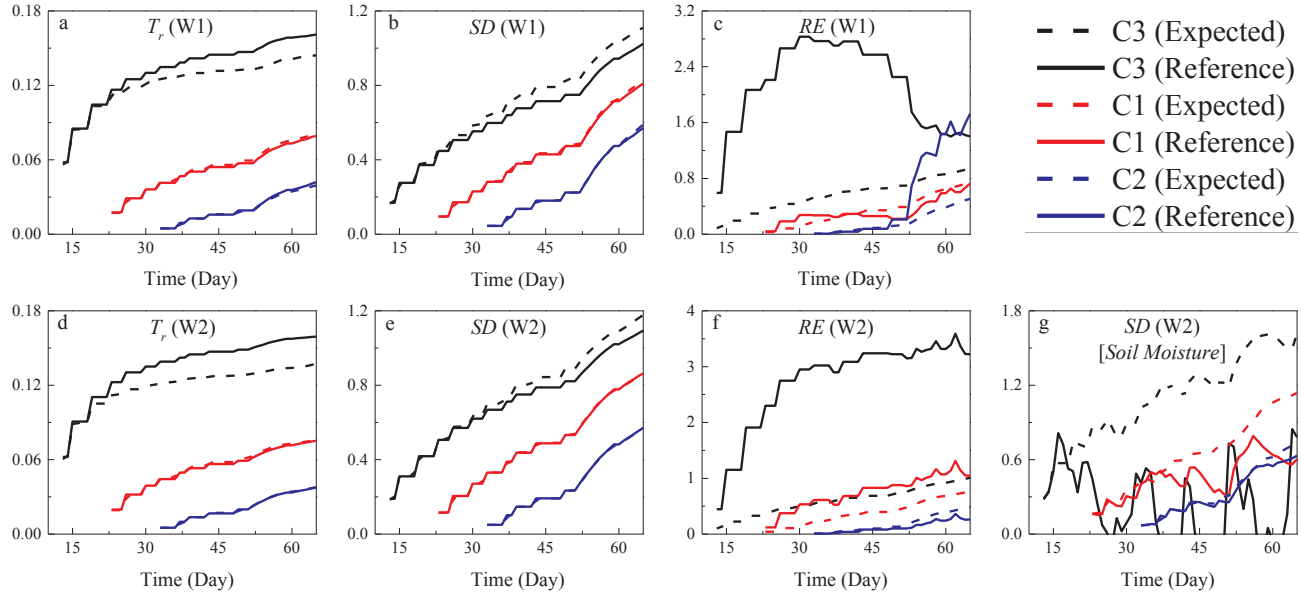

Fig. 9. (a-f) A comparison of the expected and reference data worth regarding parameter estimation for C3, C1, and C2 with 9-day, 19-day, and 30-day prior data respectively in plots W1 and W2; (g) A comparison of the expected and reference *SD* regarding soil moisture profile retrieval for C3, C1, and C2 in plot W2.

3.3.6. Effect of observation error (C1, C6, and C7)

Fig. 11 reveals a comparison of data worth among cases C6, C1, and C7, with observation error variance being equal to 0.0001, 0.0005 and 0.001 respectively. Only the results in plot W1 are shown. Different data worth indices respond very differently to the observation error. T_r increases with increasing error while *SD* shows a opposite trend. Moreover, it is seen that the reference *RE* of case C6 (with the smallest measurement error) shows abrupt rising, and the match between the reference and expected data worth deteriorates significantly after the 50th day, for both parameter estimation and soil moisture profile retrieval. For ease of comparison, the *RMSE* between the expected and reference *RE* for parameter estimation in all three plots are presented in Fig. 12(a). Two findings can be obtained: (1) the smallest observation error leads to the worst data worth estimation in all plots; (2) the data worth accuracy overall degrades when the measurement error increases from 0.0005 to 0.001. The potential observations of cases C6 and C7 are given in Fig. 12 (b) and (c), respectively. It is observed that smaller error indeed generates potential observations closer to real observations before the 50th day, while it also leads to more deviated potential observations at the late time.

Too large observation error certainly will reduce the contained value of measurement data. However, our results demonstrate that too small observation error is also unfavorable in the real-world case. One possible explanation is that relative lager observation error can actually alleviate the adverse effect of model structural error.

4. Discussion and conclusion

In this paper, with the aid of one field experiment, the data worth analysis coupled with EnKF is introduced to sequentially evaluate the

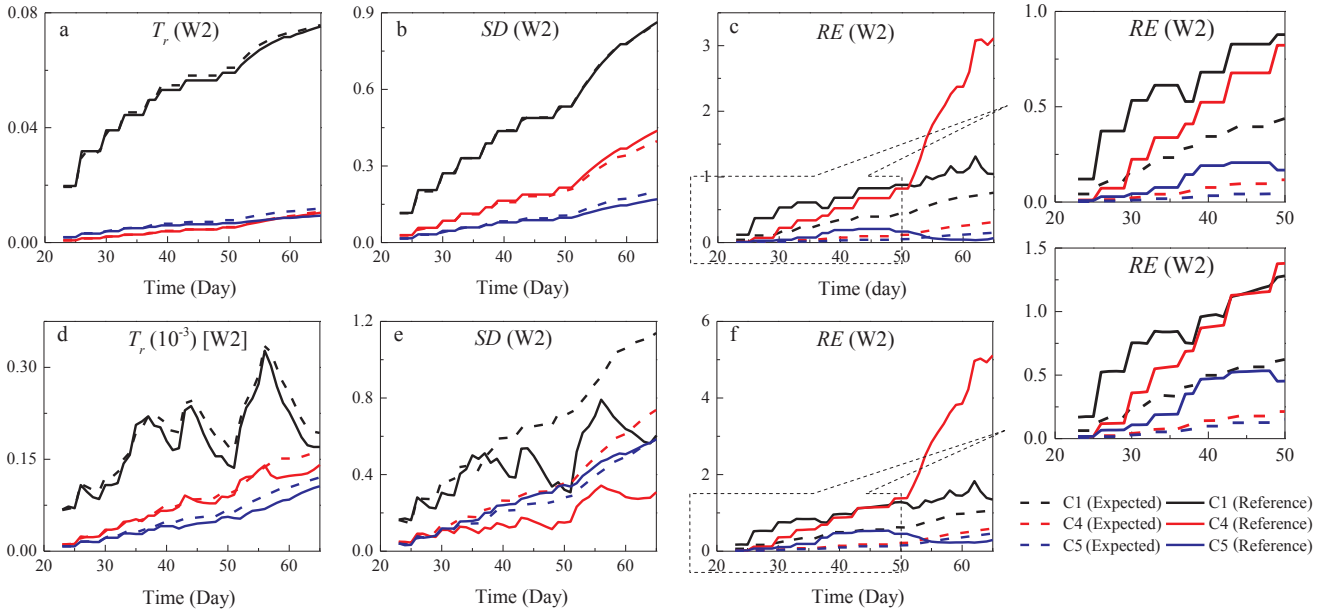

Fig. 10. A comparison of the expected and reference data worth regarding parameter estimation (a、b、c) and soil moisture profile retrieval (d、e、f) for cases C1, C4, and C5 (with potential observations at the depths of 0.02–0.18 m, 0.22–0.38 m and 0.42–0.58 m, respectively).

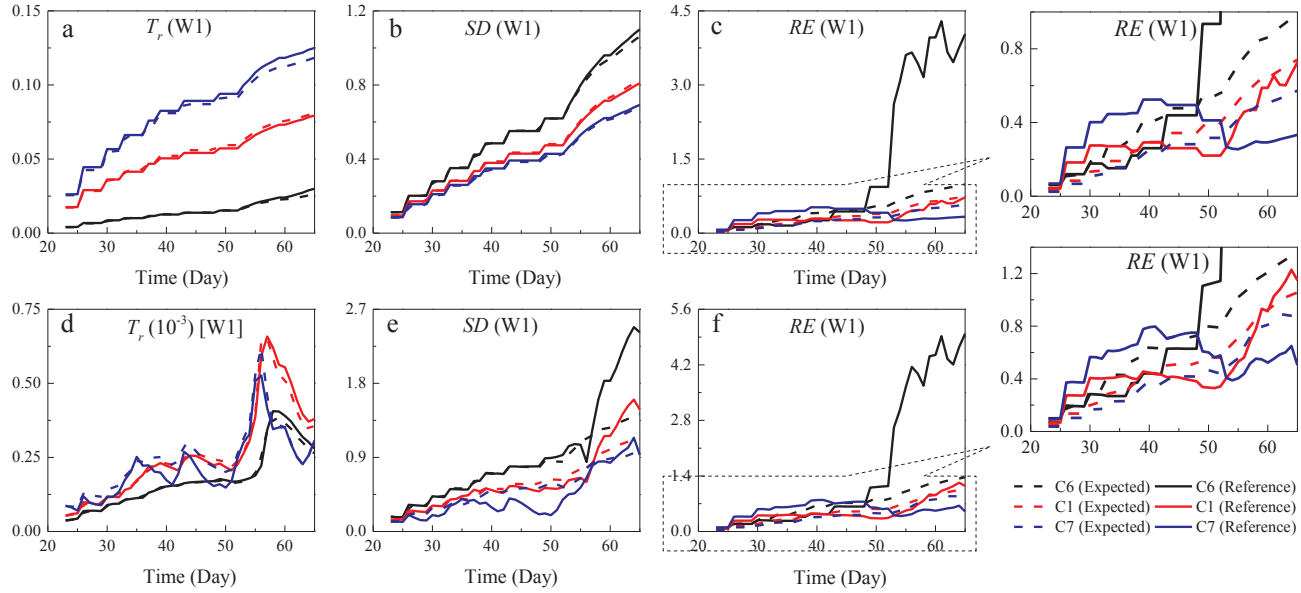

Fig. 11. A comparison of the expected and reference data worth regarding parameter estimation (a,b,c) and soil moisture profile retrieval (d,e,f) for C6, C1, and C7 (with observation error variances of 0.0001, 0.0005 and 0.001, respectively).

potential value of future soil moisture data. Three representative types of indicators, i.e. variance-type (T_r), covariance-type (*SD*, *DP*, and *DFS*) and mean-covariance-type (*RE* and J_b), are employed to quantify the data worth. The variance- and covariance-type metrics focus on the reduction of predictive uncertainty, while the mean-covariance-type index simultaneously considers the time-varying behaviors of mean and covariance. The former two can be estimated with ease while the latter is hardly reproduced due to several reasons, including the high demand on prior information, the inherent non-linearity of unsaturated flow, and the model structural error. Even so, the mean-covariance-type index may be a more objective index since it can characterize the prospective change of future data in a more comprehensive way.

The data worth of given observations to soil parameters (α, n, K_s) estimation and soil moisture profile retrieval are simultaneously assessed in this study. Regarding data worth analysis for soil moisture profile retrieval, we showed that the vertical cross-covariance of soil moisture exerts a significant impact on the estimation accuracy of covariance-type metrics. The manifesting impact is associated with the meshing size and the future scenario (such as atmospheric boundary). Soil moisture profile and associated covariance are susceptible to irrigation and evaporation events. Model structural error is likely to occur under changing scenarios. It is thus challenging to evaluate the worth of future monitoring scheme for soil moisture profile characterization in the real-world circumstance.

A series of illustrative cases were employed to explore the influence of the following various factors on data worth analysis:

(1) ***Scenario condition***. A future scenario obviously different from the historical scenario at the prior stage is likely to bring in model structural error. This raises the difficulty of 'imitating' actual measurements and leads to deteriorated data worth assessment.
(2) ***Prior information content***. Overall, more prior data can improve the match between the reference and expected data worth. The performance of variance- and covariance- type index is less sensitive to prior data volume, while mean-covariance-type indices deteriorate dramatically with insufficient prior data. Despite of the overall benefit by assimilating more prior data during data worth analysis, it cannot compensate the negative damage from unresolved model structural error.
(3) ***Observation location***. The moisture data has reduced reference data worth with increased observation depth for the variance- and covariance-type index. In this study, the surface soil layer is the most valuable location to collect observations in terms of the data worth assessment with variance- and covariance-type index. However, for the mean-covariance-type index, the data worth shows complicated patterns due to the complex features of mean and covariance. Our results demonstrated the challenge of evaluating the data worth of surface soil moisture observations with

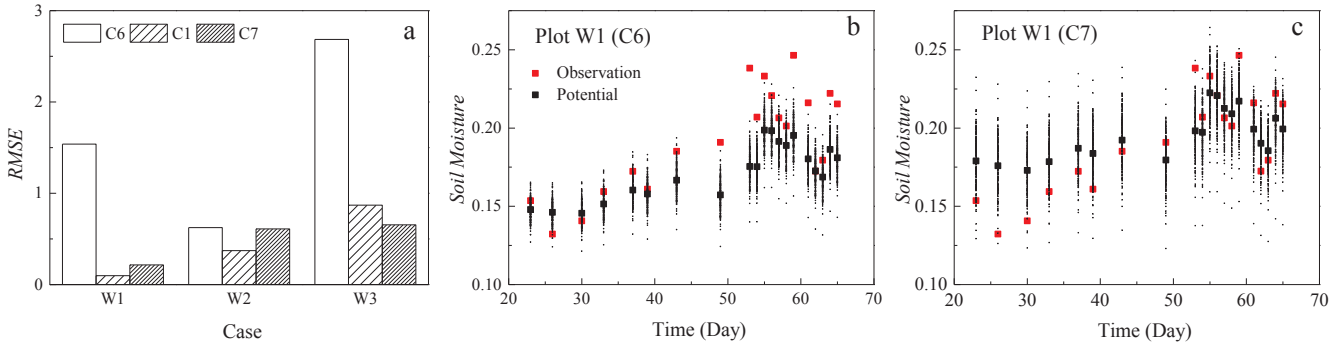

Fig. 12. (a) The *RMSE* between the expected and reference *RE* in plots W1, W2, and W3 for C6, C1, and C7 (with observation error variances of 0.0001, 0.0005 and 0.001, respectively); the actual soil moisture observations at the depth of 0.02–0.18 m and the corresponding potential observation realizations for C6 (b) and C7 (c) in plot W1.

mean-covariance-type index.

(4) ***Observation error***. Small observation error is not always helpful for improving data worth analysis. Relatively high observation error seems beneficial in a real-world case, especially when the future scenario differs with the historical scenario.

In this current study, only soil moisture data are considered. The future study will further evaluate the data worth of other types of data (such as water head, flux, tracer concentration, temperature, remote sensing data, and ground penetrating radar data). It is expected that these data can provide an additional constraint on modeling uncertainty. However, a quantitative assessment to the worth of different data sources remains to be explored. Such analysis can provide modeler with advice on which data to be collected and optimal collection strategy (frequency, location, and accuracy). Besides, how to quantify and alleviate the adverse effect of model structural error and improve the performance of mean-covariance-type index should be investigated for nonlinear soil water problem.

Acknowledgements

This study was supported by the National Natural Science Foundation of China Grant 51479144, 51629901, 51522904, 51609173, and 51779179.

References

Aanonsen, S.I., Nævdal, G., Oliver, D.S., Reynolds, A.C., Vallès, B., 2009. The ensemble Kalman filter in reservoir engineering – a review. Spe J. 14, 393–412.
Alcolea, A., Carrera, J., Medina, A., 2006. Pilot points method incorporating prior information for solving the groundwater flow inverse problem. Adv. Water Resour. 29, 1678–1689.
Beven, K., 2005. On the concept of model structural error: water science & technology a journal of the international association on water. Pollut. Res. 52, 167–175.
Calvet, J., Noilhan, J.L., 2000. From near-surface to root-zone soil moisture using year-round data. J. Hydrometeorol. 1, 393–411.
Celia, M.A., Bouloutas, E.T., Zarba, R.L., 1990. A general mass-conservative numerical solution for the unsaturated flow equation. Water Resour. Res. 26, 1483–1496.
Chen, Y., Zhang, D., 2006. Data assimilation for transient flow in geologic formations via ensemble Kalman filter. Adv. Water Resour. 29, 1107–1122.
Crow, W.T., Wood, E.F., 2003. The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. Adv. Water Resour. 26, 137–149.
Crow, W.T., Van Loon, E., 2005. Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture. J. Hydrometeorol. 7, 421.
Cull, P.O., Hearn, A.B., Smith, R., 1981. Irrigation scheduling of cotton in a climate with uncertain rainfall. Irrigat. Sci. 2, 127–140.
Dai, C., Xue, L., Zhang, D., Guadagnini, A., 2016. Data-worth analysis through probabilistic collocation-based Ensemble Kalman Filter. J. Hydrol. 540, 488–503.
Das, N.N., Mohanty, B.P., 2006. Root zone soil moisture assessment using remote sensing and vadose zone modeling. Vadose Zone J. 5, 296–307.
Dee, D.P., 2005. Bias and data assimilation. Quart. J. R. Meteorol. Soc. 131, 3323–3343.
Fetter, C. W., 2000, Applied hydrogeology: Artificial intelligence a modern approach, pp. 278–289.
Fisher, M., 2003, Estimation of entropy reduction and degrees of freedom for signal for large variational analysis systems, European Centre for Medium-Range Weather Forecasts.
Freeze, R.A., 1971. Three-dimensional, transient, saturated-unsaturated flow in a groundwater basin. Water Resour. Res. 7, 347–366.
Freeze, R.A., James, B., Massmann, J., Sperling, T., Smith, L., 1992. Hydrogeological decision analysis: 4 the concept of data worth and its use in the development of site investigation strategies. Groundwater 30, 574–588.
Gates, J.S., Kisiel, C.C., 1974. Worth of additional data to a digital computer model of a groundwater basin. Water Resour. Res. 10, 1031–1038.
Hu, S., Shi, L., Zha, Y., Williams, M., Lin, L., 2017. Simultaneous state-parameter estimation supports the evaluation of data assimilation performance and measurement design for soil-water-atmosphere-plant system. J. Hydrol.
James, B.R., Freeze, R.A., 1993. The worth of data in predicting aquitard continuity in hydrogeological design. Water Resour. Res. 29, 2049–2065.
James, B.R., Gorelick, S.M., 1994. When enough is enough: the worth of monitoring data in aquifer remediation design. Water Resour. Res. 30, 3499–3513.
Kollat, J.B., Reed, P.M., 2006. Comparing state-of-the-art evolutionary multi-objective algorithms for long-term groundwater monitoring design. Adv. Water Resour. 29, 792–807.
Kullback, S., 1959. Information Theory and Statistics. John Wiley & Sons, Inc. pp. 301.
Lettau, H., 1969. Evapotranspiration climatonomy i. a new approach to numerical prediction of monthly evapotranspiration, runoff, and soil moisture storage. Mon. Wea. Rev. 97, 81–82.
Leube, P.C., Geiges, A., Nowak, W., 2012. Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. Water Resour. Res. 48.
Li, C., Ren, L., 2011. Estimation of unsaturated soil hydraulic parameters using the ensemble Kalman filter. Vadose Z. J. 10, 1205–1227.
Li, F., Crow, W.T., Kustas, W.P., 2010. Towards the estimation root-zone soil moisture via the simultaneous assimilation of thermal and microwave soil moisture retrievals. Adv. Water Resour. 33, 201–214.
Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Quart. J. R. Meteorol. Soc. 135, 523–533.
Li, X., Shi, L., Zha, Y., Wang, Y., Hu, S., 2018. Data assimilation of soil water flow by considering multiple uncertainty sources and spatial–temporal features: a field-scale real case study. Stochast. Environ. Res. Risk Assessm. 1–17.
Man, J., Zhang, J., Li, W., Zeng, L., Wu, L., 2016a. Sequential ensemble-based optimal design for parameter estimation. Water Resour. Res. 52, 7577–7592.
Man, J., Liao, Q., Zeng, L., Wu, L., 2017. ANOVA-based transformed probabilistic collocation method for Bayesian data-worth analysis. Adv. Water Resour. 110, 203–214.
Man, J., Li, W., Zeng, L., Wu, L., 2016b. Data assimilation for unsaturated flow models with restart adaptive probabilistic collocation based Kalman filter. Adv. Water Resour. 92, 258–270.
Mantoglou, A., Gelhar, L.W., 1987. Stochastic modeling of large-scale transient unsaturated flow systems. Water Resour. Res. 23, 37–46.
Mccuen, R.H., 1974. A sensitivity and error analysis cf procedures used for estimating evaporation. JAWRA 10, 486–497.
Montzka, C., Moradkhani, H., Weihermüller, L., Franssen, H.H., Canty, M., Vereecken, H., 2011. Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter. J. Hydrol. 399, 410–421.
Moradkhani, H., 2008. Hydrologic remote sensing and land surface data assimilation. Sensors 8, 2986–3004.
Neuman, S.P., Xue, L., Ye, M., Lu, D., 2012. Bayesian analysis of data-worth considering model and parameter uncertainties. Adv. Water Resour. 36, 75–85.
Nowak, W., Barros, F.P.J.D., Rubin, Y., 2010. Bayesian geostatistical design: task-driven optimal site investigation when the geostatistical model is uncertain. Water Resour. Res. 46, 374–381.
Oliver, D.S., Reynolds, A.C., Liu, N., 2008. Inverse Theory for Petroleum Reservoir Characterization and History Matching. Cambridge University Press.
Penman, H.L., 1948. Natural evaporation from open water, bare soil and grass. Proc. R. Soc. London 193, 120.
Pipunic, R.C., Walker, J.P., Western, A., 2008. Assimilation of remotely sensed data for improved latent and sensible heat flux prediction: a comparative synthetic study. Rem. Sens. Environ. 112, 1295–1305.
Reichle, R.H., McLaughlin, D.B., Entekhabi, D., 2002. Hydrologic data assimilation with the ensemble Kalman filter. Month. Weather Rev. 130, 103–114.
Reichle, R.H., Crow, W.T., Keppenne, C.L., 2008. An adaptive ensemble Kalman filter for soil moisture data assimilation. Water Resour. Res. 44.
Renard, K. G., 1997, Predicting soil erosion by water: A guide to conservation planning with the Revised Universal Soil Loss equation (RUSLE).
Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. Physics 1, 318–333.
Ross, P.J., 2003. Modeling soil water and solute transport—fast, simplified numerical solutions. Agron. J. 95, 1352–1361.
Ross, P. J., 2006, Fast solution of Richards' equation for flexible soil hydraulic property descriptions: Land and Water Technical Report, CSIRO, v. pp. 39.
Rubin, Y., Chen, X., Murakami, H., Hahn, M., 2010. A Bayesian approach for inverse modeling, data assimilation, and conditional simulation of spatial random fields. Water Resour. Res. 46.
Shannon, C.E., 1949. Communication in the presence of noise. Proc. IRE 37, 10–21.
Shi, C., Xie, Z., Qian, H., Liang, M., Yang, X., 2011. China land soil moisture EnKF data assimilation based on satellite remote sensing data. Sci. China Earth Sci. 54, 1430–1440.
Shi, L., Song, X., Tong, J., Zhu, Y., Zhang, Q., 2015. Impacts of different types of measurements on estimating unsaturated flow parameters. J. Hydrol. 524, 549–561.
Shuwen, Z., Haorui, L., Weidong, Z., Chongjian, Q., Xin, L.I., 2005. Estimating the soil moisture profile by assimilating near-surface observations with the ensemble Kaiman filter (EnKF). Adv. Atmosph. Sci. 22, 936–945.
Singh, K., Sandu, A., Jardak, M., Bowman, K.W., Lee, M., 2013. A practical method to estimate information content in the context of 4D-var data assimilation. SIAM/ASA Journal on Uncertainty Quantification 1, 106–138.
Song, X., Shi, L., Ye, M., Yang, J., Navon, I.M., 2014. Numerical comparison of iterative ensemble Kalman filters for unsaturated flow inverse modeling. Vadose Z. J. 13.
Van Genuchten, M.T., 1982. A comparison of numerical solutions of the one-dimensional unsaturated—saturated flow and mass transport equations. Adv. Water Resour. 5, 47–55.
Vrugt, J.A., Bouten, W., Gupta, H.V., Sorooshian, S., 2002. Toward improved identifiability of hydrologic model parameters: the information content of experimental data. Water Resour. Res. 38.
Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stoch. Environ. Res. Risk Assessm. 19, 378–387.
Walker, J.P., Willgoose, G.R., Kalma, J.D., 2001. One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: a simplified soil moisture model and field application. J. Hydrometeorol. 2, 356–373.
Wu, J., Zheng, C., Chien, C.C., 2005. Cost-effective sampling network design for contaminant plume monitoring under general hydrogeological conditions. J. Contamin.

Hydrol. 77, 41–65.
Xie, X., Zhang, D., 2010. Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. Adv. Water Resour. 33, 678–690.
Xu, Q., 2007. Measuring information content from observations for data assimilation: relative entropy versus Shannon entropy difference: Tellus A 59, 198–209.
Xu, Q., Wei, L., Healy, S., 2009. Measuring information content from observations for data assimilations: connection between different measures and application to radar scan design. Tellus A 61, 144–153.
Xu, T., Valocchi, A.J., Ye, M., Liang, F., 2017. Quantifying model structural error: efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. Water Resour. Res. 53.
Xu, T., Valocchi, A.J., 2016. A Bayesian approach to improved calibration and prediction of groundwater models with structural error. Water Resour. Res. 51, 9290–9311.
Xue, L., Zhang, D., Guadagnini, A., Neuman, S.P., 2014. Multimodel Bayesian analysis of groundwater data worth. Water Resour. Res. 50, 8481–8496.
Yeh, T.C.J., Gelhar, L.W., Gutjahr, A.L., 1985. Stochastic analysis of unsaturated flow in heterogeneous soils: 1. Statistically Isotropic Media. Water Resour. Res. 21, 447–456.
Zha, Y., Shi, L., Ye, M., Yang, J., 2013. A generalized ross method for two-and three-dimensional variably saturated flow. Adv. Water Resour. 54, 67–77.
Zhang, J., Zeng, L., Chen, C., Chen, D., Wu, L., 2015. Efficient Bayesian experimental design for contaminant source identification. Water Resour. Res. 51, 576–598.
Zhang, Y., Pinder, G.F., Herrera, G.S., 2005. Least cost design of groundwater quality monitoring networks. Water Resour. Res. 41, 553–559.
Zhu, P., Shi, L., Zhu, Y., Zhang, Q., Huang, K., Williams, M., 2017. Data assimilation of soil water flow via ensemble Kalman filter: Infusing soil moisture data at different scales. J. Hydrol. 555.
Zimmerman, D.A., Marsily, G.D., Gotway, C.A., Marietta, M.G., Axness, C.L., Beauheim, R.L., Bras, R.L., Carrera, J., Dagan, G., Davies, P.B., 1998. A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resour. Res. 34, 1373–1413.