# Article (refereed) - postprint

# Spatiotemporal Sampling Strategy for Characterization of Hydraulic Properties in Heterogeneous Soils

Danyang Yu[1], Yuanyuan Zha [1*], Liangsheng Shi[1], Andrei Bolotov[2], Chak-Hau Michael Tso[3,4]

[1]State Key Laboratory of Water Resources and Hydropower Engineering Sciences, Wuhan University, Wuhan, Hubei 430072, China

[2]Department of Meteorology and Climatology, Russian State Agrarian University - Moscow Timiryazev Agricultural Academy, Moscow, 127550, Russia.

[3] UK Centre for Ecology and Hydrology, Lancaster, UK

[4] Lancaster Environment Centre, Lancaster University, Lancaster, UK

*Corresponding author, E-mail: zhayuan87@whu.edu.cn

**Highlights:**

- Cross-correlation analysis is performed for optimizing sampling strategy.

- Spatial/temporal sampling interval should be less than the correlation length.

- More observations should be sampled in wet period rather than dry period.

**Abstract:**

Accurate characterization and prediction of soil moisture distribution and solute transport in vadose zone require detailed knowledge of the spatial distribution of soil hydraulic properties. Since the direct measurements of these unknown properties are challenging, many studies invert the soil hydraulic parameters by incorporating observation data (e.g., soil moisture and pressure head) at selected point sampling locations into soil moisture flow models. However, a cost-effective sampling strategy for where and when to collect the data, which is vital for saving the costs for monitoring and data interpretation, is relatively rare compared to the direct parameter retrieving efforts. Here, an optimal spatial–temporal sampling strategy was proposed based on cross-correlation analysis between observed state variables and soil hydraulic parameters. Besides, the effects of meteorological condition, observation type, bottom boundary condition, and correlation scale of soil hydraulic parameters are also demonstrated. The proposed sampling strategy was assessed by both synthetic numerical experiments and a real-world case study. Results suggest the retrieval accuracy of heterogeneous soil is acceptable if the spatial/temporal sampling interval is set to be one spatial/temporal correlation length of soil moisture. Besides, surface observation contains the most plentiful information which could be used to derive root-zone soil moisture/parameters, but this ability depends on the correlation scale of soil hydraulic parameters. Besides, the temporal value of soil moisture depends on

meteorological condition. It is not necessary to sample repeatedly during dry periods, but more attention should be paid to the observations after rainfall events.

**Key words:** Variably saturated flow; Soil heterogeneity; Cross-correlation analysis; Sampling strategy; Data assimilation

# 1. Introduction

Soil hydraulic properties is essential for accurate predictions of water movement and solute distributions in the vadose zone (Vereecken et al. 2016). However, obtaining the detailed knowledge of soil properties is challenging, since the direct measurement of hydraulic and pneumatic flow parameters of small-scale samples at a large number of locations is time-consuming, costly, and impractical (Huang et al., 2009; Illman and Tartakovsky, 2005; Yeh and Liu, 2000). Nowadays, datastream of the soil hydraulic states (e.g., soil moisture and pressure head) can be collected conveniently via in-situ sensors. Incorporating these observations into stochastic vadoze zone flow models helps to reduce the uncertainty of the parameters and prediction, usually known as inverse modeling or model calibration (Evensen 2009; ELSheikh et al. 2013).

Observation sampling strategy (e.g., observational type, time, and location, and combination of multiple observations of different attributes) is important in inverse modeling, since it determines the costs for data collection and interpretation, as well as the accuracy of parameter and state estimation (Leube et al. 2012; Li et al. 2018). Different types of observations have been adopted for soil profile retrieval, including soil moisture (Zijlstra and Dane 1996), pressure head (Kool and Parker 1988), tracer test data (Tong et al. 2010), cumulative infiltration data (Šimůnek and van Genuchten 1996), evapotranspiration rate (Jhorar et al. 2002) and the outflow flux (van Dam et al. 1992). Nevertheless, these studies used homogeneous or prescribed multi-layer conceptual model as a surrogate for the heterogeneous soil, which may lead to

suboptimal vadoze zone flow modeling (Erdal et al. 2014). The real-world soil hydraulic properties could be better represented via geostatistical approach (Webster and Oliver 1990; Christakos 2017). Some studies treated the hydrological parameters of vadose zone as spatially correlated, statistically independent, stochastic processes for representing heterogeneity of porous media, and have successfully characterized spatial variation of soil property by incorporating soil moisture/pressure head observations into the model (Hughson and Yeh 2000; Erdal et al. 2012; Man et al. 2020). However, to the best of our knowledge, there is no study yet comprehensively exploring the sampling strategy for geostatistical-based inversion/data assimilation in soil water flow.

Previous research has shown that the inverse modeling improves with higher observation frequency (Chen et al., 2015; Li and Ren, 2011) while this effect becomes insignificant when the frequency reached a certain threshold (Dai et al., 2016). This threshold is often defined as the optimal measurement frequency (or equivalently the temporal observation interval) that minimizes the sampling cost while retains the inverse accuracy. The optimal temporal sampling interval has been intensively studied, but its value is problem-specific, ranging from one day (Hoeben, 2000; Li & Islam, 1999), two to three days (Walker et al. 2002; Pauwels et al. 2007), to up to one to two weeks (Calvet et al. 1998; De Lannoy et al. 2007). The question that then naturally arises is whether we could calculate the optimal sampling frequency in a unified way for various scenarios. One approach to solve this problem involves the use of temporal characteristics of soil moisture, such as temporal autocorrelation analysis (De Lannoy

et al. 2007). However, a quantitative method to determine the optimal observation frequency is still lacking. Moreover, precipitation and evapotranspiration conditions may significantly change the value of soil moisture which should be investigated further.

Another critical issue in observation design is the selection of observational locations. Due to the accessibility of surface soil moisture, previous studies have reported the success of estimating soil hydraulic parameters by assimilating the surface soil moisture (Montzka et al. 2011; Bandara et al. 2013). However, some studies argued that surface soil moisture might be limited to extrapolate the subsurface soil moisture state and estimate hydraulic parameters owing to the physical decoupling of surface and underlying deep soil layers (Vereecken et al. 2010). The essence of this debate is whether the inverse modeling fully considers the soil heterogeneity, which will be a focus in this study. Man et al. (2016) conducted a data worth analysis in a one-dimensional layered soil column and found that the observational locations should be placed near the interfaces of two adjacent soil layers. Moreover, Chaudhuri et al. (2018) stated that the choice of spatial sampling interval for soil water flow modeling strongly depends on the correlation length of soil hydraulic property. However, a quantitative spatial sampling criterion has not been proposed yet.

Sensitivity analysis is useful to develop optimal strategies (Šimůnek and van Genuchten, 1996). For a single-parameter inversion problem, several studies suggest that the observations should be placed at the points so that the sensitivity of the observation to the unknown parameter is the highest (Kool and Parker, 1988; Rocha et

al., 2006). However, this strategy is not suitable for a multiple-parameter inverse problem, since observations from repeated sampling at highest-sensitivity points are highly correlated (or redundant) and lead to poor estimation. As demonstrated by Yeh and Liu (2000), the key to the success of the geostatistical inverse modeling (which has many unknown parameters) is to find a combination of observations that are less correlated and provide maximum information content about the unknowns. Compared with sensitivity analysis concerning the local change of parameters, the cross-correlation analysis implicitly considers the geostatistical feature of the heterogeneous soil (Li and Yeh 1998; Sun et al. 2013), and are widely used for the optimal observation design in groundwater inverse modeling (Mao et al., 2013; Sun et al., 2013; Wu et al., 2005). Nevertheless, a systematic analysis for the spatial and temporal features of cross-correlation function between observation and the heterogeneous hydraulic parameters for vadose zone flow modeling is still lacking.

The main thrust of this paper is to: a) analyze the spatiotemporal features of the cross-correlation between hydraulic parameters and various types of state variables in strongly heterogeneous soil; b) explore the effects of meteorological condition, bottom boundary condition and correlation scale on the characterization of the soil heterogeneity; and c) propose a quantitative method to determine the optimal spatiotemporal sampling strategy for vadose zone flow modeling.

## 2. Method

### 2.1 Data assimilation framework

*2.1.1 One-dimensional model of variably saturated flow*

One-dimensional (1-D) vertical flow plays the dominant role for soil water movement in the unsaturated zone with insignificant hillslope (Chen et al., 1994) and has been widely investigated in the last decade (Walker et al. 2001; Das and Mohanty 2006; Younes et al. 2017). Considering the consistency with previous relevant work and computational costs, here we focus on 1-D soil water flow, which can be described by 1-D Richards' equation:

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

(1)

where $\theta$ is the volumetric soil moisture; $h$ is the soil water pressure head; $K$ is the unsaturated hydraulic conductivity which depends on $h$ for unsaturated soils; $t$ represents the time; $z$ denotes the coordinate in the vertical direction, assumed positive upward. To solve Eq. 1, the constitutive relationship between function $K$ and $h$ as well as $h$ and $\theta$ must be given *a prior*. van Genuchten-Mualem model (van Genuchten, 1980; Mualem, 1976) is often used to describe the relationship as follows:

$$\theta(h) = \begin{cases} \theta_r + \dfrac{\theta_s - \theta_r}{[1+|\alpha h|^n]^m} & h < 0 \\ \theta_s & h \geq 0 \end{cases}$$

(2)

$$K(h) = \begin{cases} K_s S_e^{1/2}\left[1-\left(1-S_e^{1/m}\right)^m\right]^2 & h < 0 \\ K_s & h \geq 0 \end{cases}$$

(3)

$$m = 1-1/n \quad n > 1$$

(4)

$$S_e = \frac{\theta - \theta_r}{\theta_s - \theta_r} \qquad (5)$$

where $\theta_s$ and $\theta_r$ indicate the saturated and residual soil moistures respectively; $\alpha$ and $n$ are the shape parameters of the soil moisture characteristic curve; $K_s$ is the saturated hydraulic conductivity; $S_e$ represents the effective saturation degree. It should be noted that the effects of source/sink term and hysteresis are not taken into the consideration in this work.

Initial and boundary conditions are necessary for the simulation of soil moisture movement. The initial condition is the state of the soil moisture,

$$\theta(z,t)\big|_{t=0} = \theta_0(z) \qquad (6)$$

where $\theta_0(z)$ is the initial soil moisture profile. The top boundary is the state-dependent, atmospheric boundary condition (Neuman et al. 1974), which could be described as,

$$|q| = \left| -K \frac{\partial h}{\partial z} - K \right| \le |E_p - P_p| \qquad (7)$$

$$h_m > h > h_c \qquad (8)$$

where $q$ is the Darcian flux at the soil surface. $E_p$ denotes the potential evaporation, which is calculated using the Food and Agriculture Organization (FAO) Penman-Monteith equation (Allen et al. 1998). $P_p$ represents the precipitation intensity, and $h_m$ and $h_c$ are the maximum and minimum pressure heads allowed at the soil surface, respectively. The value of $h_m$ is set to 0, whereas $h_c$ is determined as $-100$ m.

The bottom boundary condition could be the free drainage boundary (Eq. 9) or zero-flux boundary (Eq. 10):

$$\frac{\partial h}{\partial z}\bigg|_{z=z_N} = 0 \qquad (9)$$

$$q\big|_{z=z_N} = 0 \qquad (10)$$

where $z_N$ is the depth of bottom boundary.

*2.1.2 Data assimilation algorithm (IES)*

Traditional inverse methodology in vadose zone, such as the ensemble Kalman filter (EnKF) (Evensen 2009), are limited to mildly nonlinear systems and inapplicable to identify small-scale heterogeneity. Recently, an iterative data assimilation method (i.e., iterative ensemble smoother (IES)) has been developed by Chen and Oliver (2013) and shown robust performances in strongly nonlinear problems in hydrology (Crestani et al., 2013; Emerick and Reynolds, 2013), which is suitable for our study.

In this section, the details of the IES algorithms are presented. The observation vector for each ensemble member is,

$$\mathbf{d}_j^{obs} = \mathbf{d}^{obs} + \boldsymbol{\varepsilon}_j \tag{11}$$

Where $j$=1, 2,…, $N_e$, with $N_e$ the number of ensemble size taken as 300 in this study (Ju et al. 2018); $\mathbf{d}^{obs}$ is a vector of the true observation; the dimension of $\mathbf{d}^{obs}$ is $N_d$, denoting the total number of all available observations; $\boldsymbol{\varepsilon}_j$ are independent white noises for ensemble index $j$; $\mathbf{d}_j^{obs}$ is the "perturbed" observation (i.e., the observation with measurement error). The model parameter vector at iteration $r$, $\mathbf{m}^r$, can be updated by combining the observations and predictions,

$$\mathbf{m}_j^{r+1} = \mathbf{m}_j^r + \mathbf{K}^r (\mathbf{d}_j^{obs} - \mathbf{d}_j^{f,r}) \tag{12}$$

where $\mathbf{m}_j^r$ is the initially guessed or estimated parameters for realization $j$ at iteration $r$; $\mathbf{m}_j^{r+1}$ is the updated estimates at iteration $r$+1 (i.e., the prior parameters at iteration $r$+1) for realization $j$; $\mathbf{d}_j^{f,r}$ is a vector of predicted data at iteration $r$. $\mathbf{K}$ is Kalman gain, which is given by,

$$\mathbf{K}^r = \mathbf{C}_{md}^r (\mathbf{C}_{dd}^r + \mathbf{C}_d + \lambda diag(\mathbf{C}_{dd}^r))^{-1} \qquad (13)$$

where $\mathbf{C}_{md}^r$ indicates the cross-covariance matrix between the parameter vector $\mathbf{m}_j^r$

and the predicted data vector $\mathbf{d}_j^r$; $\mathbf{C}_{dd}^r$ denotes the auto-covariance matrix of

predicted data vector $\mathbf{d}_j^r$ with the dimension of $N_d$ x $N_d$. $\mathbf{C}_d$ represents the error

covariance matrix of measurements. $\lambda$ is a dynamic stability multiplier (the prior value

is 10) and $diag(\mathbf{C}_{dd}^r)$ is a diagonal matrix with the same diagonal elements as $\mathbf{C}_{dd}^{f,r}$.

Mathematically, the dynamic stabilizer is determined by Levenberg-Marquardt

approach (Pujol 2007) which facilitates the solution switching between the Gauss-

Newton solution and the steepest-descent method.

## 2.2 Statistical analysis method

### 2.2.1 Cross-correlation analysis

We conduct Monte-Carlo simulations to analyze the cross-correlation between soil

hydraulic parameters (e.g., saturated hydraulic conductivity) and observations of state

variables (e.g., soil moisture content). We first generate random fields based on the

geostatistical parameters of the soil and then run the soil water movement model

individually to obtain different sets of state variables. The cross-correlation $\rho$ can be

described by the statistics of soil hydraulic parameters and model state variables as

$$\rho(z_p, z_d, t_d) = \frac{C_{md}(z_p, z_d, t_d)}{\sigma(\mathbf{m}_{z_p})\sigma(\mathbf{d}_{z_d}^{t_d})} \qquad (14)$$

$$C_{md}(z_p, z_d, t_d) = \frac{1}{N_e - 1} \sum_{j=1}^{Ne} (m_{z_p,j} - \overline{m}_{z_p})(d_{z_d,j}^{t_d} - \overline{d}_{z_d}^{t_d}) \qquad (15)$$

where $m_{z_p,j}$ and $d_{z_d,j}^{t_d}$ indicate the parameter at a depth of $z_p$ and the state variable at

depth the of $z_d$ at time $t_d$ for ensemble index $j$ during the simulation, while $\overline{m}_{z_p}$ and

$\overline{d}_{z_d}$ denote the average values of $m_{z_p,j}$ and $d_{z_d,j}^{t_d}$, respectively; $C_{md}(z_p, z_d, t_d)$ is the

covariance between ensemble $\mathbf{m}_{z_p}$ and $\mathbf{d}_{z_d}^{t_d}$, and $\sigma(.)$ stands for the standard deviation

of the ensemble.

*2.2.2 Spatial and temporal correlation ratio*

To investigate the impacts of the spatial/temporal density of the measurement

points and the correlation length of soil moisture on the resolution of the estimation, we

adopt correlation length ratio $Ra$ to indicate the spatial/temporal sampling index, which

is defined as:

$$Ra = \frac{\lambda}{d} \tag{16}$$

where $d$ is the spatial/temporal interval between measurement points (i.e., the distance

between two adjacent observation locations or the observation frequency); $\lambda$ is the

spatial or temporal correlation length of soil moisture. Among them, the spatial

correlation length is given as the scale of constructed parameter field which is equal to

0.25 m, while the temporal correlation length is determined through the calculation of

the soil moisture autocorrelation (De Lannoy et al., 2006).

*2.2.3 Quantitative indexes for evaluating inversion results*

To evaluate the accuracy of the estimated parameters, the coefficient of

determination ($R^2$), the mean square error ($RMSE$) and the relative error ($RE$) are used

as quantitative indexes, which can be calculated as follows:

$$R^2 = \left[ \frac{\frac{1}{n} \sum_{i=1}^{n} (m_i - \overline{m_i})(m_i^t - \overline{m_i^t})}{\sigma(m)\sigma(m^t)} \right]^2 \tag{17}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (m_i - m_i^t)^2} \tag{18}$$

where $n$ is the total number of the estimated values; $m_i^t$ and $m_i$ represent the true and final estimated parameter or state variable values predicted using the estimated parameter (for the case where the true parameters are unknown), while $\overline{m_i^t}$ and $\overline{m_i}$ indicate the average values.

## 3. Synthetic numerical experiments

A number of synthetic numerical simulations are conducted in this study. In section 3.1, we present a general description of the model settings in the variably saturated flow models. In order to explore the relationship between soil hydraulic parameters ($K_s$, $\alpha$ and $n$) and various types of observations, the cross-correlation analysis is performed based on the Monte-Carlo method outlined in section 3.2. In section 3.3, various series of observations are assimilated into the model to validate the conclusions of the cross-correlation analysis and to determine the optimal spatial and temporal sampling intervals.

### 3.1 General description of model settings

In this part, only one parameter (e.g., $K_s$, $\alpha$ or $n$) is assumed to be uncertain at a time in the Monte-Carlo simulations or parameter inversions, and the other parameters are set to be known. With a given correlation scale (i.e., 0.25 m), 300 random fields for

each soil hydraulic parameter ($K_s$, $\alpha$ or $n$) are generated following logarithmic normal distributions. The soil is assumed to be the sandy, with mean values of 1.3 m d$^{-1}$, 5.2 m$^{-1}$ and 2 according to Carsel and Parrish (1988), respectively and corresponding (log-transformed) variances of 0.5, 0.3 and 0.01.

To investigate the effects of boundary conditions on cross-correlation analysis, three typical meteorological conditions and two common bottom boundary conditions are considered. The precipitation and reference evaporation data are chosen to be representative of three different regions (arid, semi-arid and humid regions) in China, as shown in Fig. 1. The investigated bottom boundary conditions include free drainage and zero-flux boundary.

Unless otherwise specified, the initial condition is determined by spinning up the model with one-year meteorology data (Yu et al. 2019). The default upper and bottom boundary conditions are semi-arid climate and free drainage boundary condition, respectively. Besides, the soil profile is set to be 2 m thick, and the flow domain is discretized into 40 grids with a grid size of 5 cm. The spatial discretization error has been assessed by sensitivity test and it shows that the error does not affect the general conclusions in this study. Each grid is assigned with a parameter value using the aforementioned random fields. The simulation time is 128 days in the synthetic experiments. Other details and assumptions for our model simulation runs are given in Table 1.

**3.2 Cross-correlation analysis**

In order to investigate the spatiotemporal features of cross-correlation with various model settings, Case 1, with nine scenarios, has been performed in this section, as presented in Table 2. Scenarios A1 to A3 in Case 1 explore the cross-correlation between the hydraulic parameters ($K_s$, $\alpha$ and $n$) and surface soil moisture. The relationship between parameters ($K_s$ and $\alpha$) and soil moisture at the depth of 100 cm is considered in Scenarios A4 and A5. Scenarios A6 and A7 are used to investigate the effects of observational types in the cross-correlation analysis, while A8 and A9 examine the impact of bottom boundary condition and correlation length on the cross-correlation map.

Fig. 2 presents the temporal change of cross-correlation between observations and hydraulic parameters under a variety of scenarios in Case 1. In general, the observation carries the most considerable information about parameters around the observation point, while it has no significant correlation with the parameters further away (Figs. 2a to 2g). The results agree with the conclusions made by Mao et al. (2013) and are often explained by physical decoupling of adjacent soil layers (Capehart and Carlson, 1997), the degree of which is determined by soil hydraulic properties. More generally, the degree of decoupling is affected by the correlation scale of soil property. As shown in Fig. 2i, the zone length with a high correlation value (between $\ln K_s$ and surface soil moisture) increases when the auto-correlation length of $\ln K_s$ grows from 25 cm to 50 cm, indicating that the decoupling degree of cross-correlation would decrease with the

increase of correlation scale of hydraulic parameter.

During a dry period ($1^{st}$ to $15^{th}$ day), the surface soil moisture (at a depth of 10 cm) is *negatively* correlated with $\ln K_s$ above the observed point (Fig. 2a), whereas *positively* correlated with $\ln K_s$ below (10-50 cm). The results are reasonable, since the change of surface soil moisture is mainly caused by evaporation due to an upward vertical soil moisture gradient (Fig. 3a). A larger hydraulic conductivity above the observation (0-10 cm) would promote the evaporation loss at the observation while a larger $K_s$ below can supply sufficient water from underlying layer and prevent the reduction of observed soil moisture. In addition, although significantly affected by precipitation and evaporation, the cross-correlation pattern is not solely determined by the upper boundary condition. For instance, the correlation between surface soil moisture and $\ln K_s$ at around 0 to 50 cm is positive from days 16 to 20 but negative from days 51 to 53, while both periods are under precipitation (see Fig. 1(b)). This shift is attributed to the difference in the preceding soil moisture profile (see Fig. 3a). From days 16 to 20, an increase of $K_s$ facilitates the increase of the observed soil moisture, since the preceding soil moisture is small and the observed location is gaining infiltrated water. In contrast, the soil moisture profile at day 51 has excessive water and the observation location is draining water even under precipitation from days 52 to 54, resulting a negative $\rho$ value.

Parameters $\alpha$ and $n$ are associated with water-holding capabilities of the soil, and a larger value of $\ln \alpha$ or $\ln n$ means less water retention ability in the soil, leading to a

decrease in surface soil moisture. Therefore, surface soil moisture shows a strong negative correlation with parameter $\ln\alpha$ or $\ln n$ over the top 10 cm and the cross-correlation for $\ln\alpha$ or $\ln n$ has relatively small variation over time (Figs. 2b and 2c), meaning that combining soil moisture observations at different time may be redundant when inverting $\alpha$ or $n$. Nevertheless, the significant changes of cross-correlation values during infiltration periods in Figs. 2b and 2c indicate that the combination of observations at this period is critical to invert $\alpha$ or $n$.

Unlike surface observations, the cross-correlation with deep observations (at the depth of 100 cm) is temporally more stable (Figs. 2d, 2e, 2g and 2h). The result from 0 to 95th day for $\ln K_s$ (Fig. 2d) is similar to the change of cross-correlation during the dry period in Fig. 2a, because the soil moisture at $z_d = 100$ cm is not affected by rainfall in this period and is changed slowly by long-term soil evaporation (Fig. 3a). Similarly, for $\ln\alpha$ (Fig. 2e) and $\ln n$ (result not shown), the deep soil moisture has the strongest correlation with the parameters near the observation location. From Figs. 2d, 2e, and 2g, we find that deep soil moisture observation after a heavy rainfall (e.g., day 96) may provide useful hydraulic property information at deep layer, since a significant change of deep soil moisture is observed at $t = 98\sim100$ days when the wetting front moves to $z_d = 100$ cm (Fig. 3a).

Figs. 2f and 2g explore the effects of observation type on cross-correlation analysis. Different from soil moisture (Figs. 2b and 2e), the observation of pressure head ($h$) is positively related with the parameter $\ln\alpha$, as self-explained by Eq. 2. The cross-

correlation results between pressure head and $\ln n$ are similar to that for $\ln\alpha$. While for $\ln K_s$, the observation type shows little effects on cross-correlation pattern (not shown), because the soil water retention equation (Eq. 2) does not involve parameter $K_s$. In addition, Fig. 2h presents $\rho$ ($z_p$, 100, $t_d$) between $\ln\alpha$ and deep pressure head under a zero-flux bottom boundary condition. Compared with the result under free drainage boundary condition (Fig. 2g), the pressure head (or soil moisture) of deep soil under zero-flux boundary condition gradually approaches to zero (i.e., soil is saturated) with the accumulation of infiltration (Fig. 3b). Therefore, the deep pressure head shows no significant correlation with the deep soil parameters, implying that the information of observation may not be enough to estimate the parameters of deep soil accurately under the zero-flux boundary condition.

**3.3 Parameter identification**

The aforementioned cross-correlation analysis leads to the following four questions about parameter estimation in a heterogeneous soil column: (1) Are the results of noise-free soil moisture and head observations the same to estimate parameter $K_s$, but different for the estimation of parameter $\alpha$ and $n$ if they are sampled at the identical locations and times? (2) Would it be more challenging to estimate the hydraulic parameters at the deep soil with a zero-flux boundary than those with a free drainage boundary? (3) How many spatial/temporal observation points should be used to obtain good estimates of parameter field while minimizing computational effort for the heterogeneous field retrievals? (4) Is it necessary to sample the observations repeatedly during a long-term

dry period (since they contain a lot of redundant information)?

To answer these questions, several cases are designed for parameter inversions which incorporate the information of observation from various perspectives. In response to the first and second question, different types of observations are assimilated into the model with various model settings and we then compare the results of parameter estimation in section 3.3.1. For the third and fourth questions, we explore the relationship between inversion results and the spatial correlation scales in Section 3.3.2. Besides, the least observation days and temporal sampling strategies are proposed in Section 3.3.3.

One realization of parameter random field ensemble (i.e., prior parameters $K_s$, $\alpha$ or $n$ defined in the Section 3.2) is selected as the true value of unknown parameters, while the true values of known parameters are set as those of loam $(\theta_s = 0.43 \text{ cm}^3/\text{cm}^3, \theta_r = 0.078 \text{ cm}^3/\text{cm}^3)$ according to Carsel and Parrish (1988). The synthetic observations used for heterogeneous field retrievals are generated by running the model with the true parameters and disturbed by the observation errors (a standard deviation of 0.01). Unless otherwise specified, the spatial observation interval is approximately 50 cm (i.e., at the depths of 5 cm, 50 cm, 100 cm, 150 cm, and 200 cm) and the observed soil moisture is assimilated every 16 days, starting from day 16.

*3.3.1 Effects of observation types and errors with various model settings*

As shown in Table 3, Case 2 (with six scenarios) is performed in this part. Scenarios B1 to B3 explore the impacts of soil moisture on the estimation of $\ln K_s$, $\ln n$, and $\ln \alpha$

respectively, while Scenario B4 presents the inversion results of ln$\alpha$ by assimilating the observation of pressure head. To find out the role of bottom boundary conditions in parameter inverse problems, B5 is conducted to estimate ln$\alpha$ with a zero-flux boundary. Besides, B6 investigates the influence of observation error on the parameter inversion results.

Fig. 4 presents the *RMSE* result of parameter estimations for each parameter under a variety of scenarios in Case 2. As shown in Figs. 4a to 4c, the parameters of soil profile can be well estimated by assimilating soil moisture without observation error. Regarding the observation of pressure head, the estimation of ln$K_s$ and ln$n$ are also satisfactory (results not shown), but the *RMSE* of ln$\alpha$ increases from 0.0039 to 0.091, showing a significant discrepancy between estimated parameters and true parameters (Fig. 4d). The result might be explained by the cross-correlation maps in Figs. 2e and 2g. For one thing, pressure head shows a lower correlation to ln$\alpha$ compared with soil moisture. For another thing, pressure head changes continuously over the profile which might be limited to provide the information for a dramatic change of spatial parameter. Fig. 4e presents the results of ln$\alpha$ estimation under zero-flux boundary condition. The results of parameter estimation show a significant deterioration in the deep soil profile, implying the difficulties in characterizing the capillary fringe (Fig. 2h). Besides, different types of observations show slight effects on the estimation of ln$\alpha$ under zero-flux boundary, since the main bias is induced by the boundary condition. Yet, utilizing the pressure head to estimate parameters is more conducive to numerical stability when

under the zero-flux boundary condition, since pressure head observation changes continuously (negative in unsaturated soil and positive in saturated soil), while soil moisture cannot provide the information of saturated zone. Fig. 4f plots the results of $\ln K_s$ estimation when assimilating soil moisture observations with an observation error. Compared with Fig. 4a, there is remarkable increase of *RMSE* from 0.0016 to around 0.28, indicating that the observation error can deteriorate the parameter estimation results distinctly.

*3.3.2 Effects of spatial sampling strategies*

To quantify the relationship between optimal spatial sampling interval and correlation scale. Six correlation length ratios (i.e., the *Ra* of 1/8, 1/4, 1/2, 1, 2.5, and 5) are adopted in this part and corresponds to the spatial interval *d* of 2 m, 1 m, 0.5 m, 0.25 m, 0.1 m, and 0.05 m respectively (Eq. 16). The estimation results of individual parameters with different correlation length ratios are presented in Fig. 5. We repeated the numerical experiments three times for each scenario by varying different observational locations but with the same measurement interval. As shown in the figure, *RMSE* value between estimated parameters and true parameters decay exponentially with increasing correlation length ratio. As the value of *Ra* increases from 1/8 to 1, the mean and standard deviation of *RMSE* decreases significantly, indicating a considerable effect of spatial measurement interval on the accuracy of parameter estimation. Nevertheless, when *Ra* grows up from 1 to 5, the decreasing rate of *RMSE* drops, implying that the improvement of spatial density of the measurement points become

less critical for parameter estimation at this stage.

A similar result could also be found in Fig. 6 which normalize the accuracy of parameter estimation for $\ln K_s$, $\ln\alpha$ and $\ln n$ based on determination coefficient $R^2$ index. When $Ra \geq 1$, the mean values of $R^2$ is large (more than 0.8), and the improvement for the accuracy of parameter estimation is marginal. Therefore, we recommend that the optimal sampling interval should be equal to the correlation scale of the dominant heterogeneity considering the expensive cost of sampling and minimal improvement of parameter estimation.

To explore the impacts of observation locations on estimated parameter field, the retrieved $\ln K_s$ field with various $Ra$ values (1/8, 1 and 5) is plotted in Fig. 7. When there is only one observation point at the surface (in Fig. 7a), the heterogeneity around this point (5 to 50 cm) can be characterized well, but anywhere else shows no improvement from the prior values. According to Fig. 2(a), the observations at the surface can only provide the information of soil parameters within 0.5 m depth. Therefore, surface observation is not sufficient to describe the heterogeneity of the soil profile, making it necessary to incorporate the observations at various soil depths. By fusing the information from various depths, the characterization of heterogeneity of the soil profile can be improved significantly (Fig. 7b). Despite the *RMSE* decreases with the increase of the observation locations, the improvement is limited when $Ra$ increases from 1 to 5 (Fig. 7c). The results imply that thesoil moisture at the top 5 cm obtained by remote sensing might be limited in utility to derive the information of root-zone soil moisture

in heterogeneous soils, and we have to collect the measurements at the various depths.

*3.3.3 Effects of temporal sampling strategies*

In order to determine the characteristic time scale of soil moisture and explore the effects of precipitation and evaporation, the autocorrelation for increasing time lags of the daily averaged time series of soil moisture is calculated under different meteorological conditions (Fig. 1), as presented in Fig. 8. Overall, the autocorrelation of soil moisture decreases rapidly at surface, while the value declines more slowly at deeper layers. For example, in humid climate (Fig. 8c), the autocorrelation shows stationarity for time lags of approximately 10 and 25 days for soil moisture at 5 and 50 cm depth respectively. In contrast, a slight increase of the characteristic time scale is observed at deeper layers, with around 35 days for soil moisture at 100, 150 and 200 cm. The growth in the characteristic time scale is reasonable, since the surface soil moisture is strongly affected by precipitation and evaporation, while the soil moisture changes more slowly at deep layers, as discussed in Section 3.2. In addition, the meteorology condition plays an important part in autocorrelation of soil moisture. With the reduction of precipitation, the characteristic time scale at depth of 5 cm increases from around 10 days (Fig. 8c) to 40 days (Fig. 8a). At the same time, the autocorrelation at deeper layers (i.e., 100, 150 and 200 cm depth) keeps an almost constant value of 1 during the whole period (Fig. 8a), indicating that the characteristic time scale of soil moisture at deeper layers exceeds 50 days under the arid climate.

The temporal correlation length is determined as the time lag for which the

autocorrelation becomes $1/e$ according to De Lannoy et al. (2006). To maximize information for all observation depths, the minimum correlation length (i.e., 18, 12 and 2 days in arid, semi-arid and humid climate respectively in Fig. 8) is used as $\lambda$ to calculate the value of correlation ratio $Ra$ (Eq. 16). Similar with spatial sampling strategies, five different temporal sampling strategies (with $Ra$ value of 1/8, 1/4, 1/2, 1, and 2 separately) and an intensive sampling scenario (i.e., with the temporal interval of 1 day) are adopted in this part. Besides, three repeated simulations are conducted for each scenario by changing the observation dates but with the same sampling interval.

$RMSE$ results of $\ln K_s$ estimation with various temporal sampling intervals and meteorological conditions are presented in Fig. 9. Overall, the $RMSE$ value declines with the increase of observation days and approach to the reference value of intensive sampling strategy gradually. The performance deteriorates when $Ra$ is lower than 1/2 while the rate of decreasing stabilizes at around 1 when observation frequency is equal to one minimum temporal correlation length. In addition, various meteorological conditions lead to somewhat different results. The $RMSE$ value as well as its standard deviation under humid climate is much lower than that under arid and semi-arid climate. Moreover, with the growth of observation days, the accuracies of estimated parameters under semi-arid and humid climate are less different ($RMSE = 0.187$ and 0.195 respectively for the intensive sampling scenario), which outperform the conditions under arid climate ($RMSE = 0.267$). The results imply that the occurrence of rainfall event facilitates the retrieval of the heterogeneous parameter field, due to the change of

observation information during the processes of accumulation and infiltration as discussed in Section 3.2.

In practice, the temporal sampling interval is often irregular. Also, the value of temporal observation point may vary over time owing to the evaporation and infiltration processes according to the cross-correlation analysis. To investigate the effects of the different temporal sampling points on the estimation, we conduct a total of 50 inversion cases for estimation of $\ln K_s$, each of which has 16 temporal observation points randomly selected from the simulation period. Therefore, 50 $RMSE$ values can be obtained for the 50 inversions. As presented in Fig. 10, given in ascending order, these $RMSE$ values are divided into three categories: optimal solutions (first 20%: $1^{st}$ to $10^{th}$), sub-optimal solutions (40% to 60%: $21^{th}$ to $30^{th}$) and the worst solutions (80% to 100%: $41^{th}$ to $50^{th}$) of which the mean value is 0.220, 0.229, and 0.248 respectively, showing an obvious disparity of the accuracy of the parameter estimation.

To figure out the differences between the optimal solutions and the worst solutions, the recurrence number of observation day versus precipitation with three parameter estimation accuracy levels are shown in Fig. 11. It should be noted that the low-frequent observation days (no more than two times) are not shown here for the sake of brevity. For the optimal solutions (red circle), the highly-frequent observation days rarely appear in the dry period (0 to 40 days), with one exception at $t = 30^{th}$ day. They most likely emerge during the precipitation period (40 to 128 days). Moreover, these temporal points do not appear after the rain immediately but occur 4 to 8 days after the

precipitation, which is approximately the time for the infiltrated water move to the deeper soil column. That is because the soil moisture observation provides the information of hydraulic properties at the lower soil profile, which would be discussed further below.

Regarding the simulations leading to the worst solutions (blue squares in Fig. 11), there are five highly-frequent observation days appearing during the dry period. The result confirms that multiple observations at different times during the dry period contain redundant information and leads to a lower resolution of parameter inversions. This is attributed to the less change in the covariance of soil moisture and parameters in the dry period (i.e., there is a strong "soil moisture memory" as demonstrated by Koster and Suarez (2001)). In addition, the highly-frequent observation days in the rainy days often appear on 1 to 2 days before or after the rainfall. At this time, the observations can only reflect the soil hydraulic properties near the soil surface, while the information about hydraulic parameters at deeper soil is lacking.

For the sub-optimal inversion results (green triangle in Fig. 11), the sampling dates in the whole period are scattered with no apparent features. In summary, it is not suggested to sample frequently in the dry period because of redundant information, while the observations after a period of the rainfall are more informative. The length of the period between the observed day and rainy day may be related to the permeability of the soil and the amount of rainfall.

To further investigate the reason leading to the difference of the optimal solutions

and the worst solutions, the prior probability density distribution of soil moisture ensemble at various spatial locations ($z_d$ = 25 cm and 150 cm) and temporal sampling points is presented in Fig. 12. It is shown that, with the optimal solutions (Figs. 12a and 12b), the prior probability density distributions of soil moisture ensemble are apparently different at various highly-frequent observation days (4 times or more) whether for the locations at the surface ($z_d$ = 25 cm) or deep layer ($z_d$ = 150 cm). The observation time of the 109th day appears for five times for the optimal solutions in Fig. 11, showing the importance of this observational date. As presented in Fig. 12b, the observations at this day provide the new information of the parameters at the deep layer and thus are important for the parameter estimation, which further confirms the strong effects of precipitation and drainage process on parameter field inversions. On the contrary, with the worst solutions, the prior probability density distributions of soil moisture ensemble are visually similar (Figs. 12c and 12d). The observation of deep soil moisture at $z_d$ = 150 cm (Fig. 12d) bring little new information, which could partly explain the inferior results of the worst solutions.

## 4. A real-world case study

### 4.1 Data and site description

To validate our results, a real-world case study has been conducted. The soil moisture measurements are obtained from the Lincoln_11_SW station (40.6954N, 96.8541W, located in Gage County, Nebraska, USA) of USCRN network (U.S. Climate Reference Network) and can be downloaded from the *International Soil Moisture*

*Network (ISMN)* which establishes and maintains a global in-situ soil moisture database (https://ismn.geo.tuwien.ac.at/en/). The upper boundary condition is atmospheric boundary, and except for the precipitation data provided by the *ISMN*, other meteorological data (e.g., radiation, air temperature and wind speed) are downloaded from the NASA Prediction Of Worldwide Energy Resources (https://power.larc.nasa.gov/). Furthermore, the reference evapotranspiration can be calculated using the Penman-Monteith equation (Allen et al. 1998) recommended by Food and Agriculture Organization (FAO).

The meteorology and soil moisture data from $30^{th}$ March 2018 to $25^{th}$ September 2018 (a total of 180 days) are used in this study, as shown in Fig. 12. The first 120 days are calibration period (with a dry period from $1^{th}$ day to $60^{th}$ and a wet period from $61^{th}$ day to $120^{th}$ day), and the last 60 days are used for prediction. Besides, the volumetric soil moisture was measured at five depths (5, 10, 20, 50 and 100 cm) with the temporal interval of 1 day and the measurement error is assumed to be 0.01 $cm^3/cm^3$.

**4.2 Model setup**

To avoid the effects of bottom boundary, the 1-D numerical domain is set as 120 cm and discretized in 60 grids with a size of 2 cm. The upper boundary condition is set as an atmospheric boundary using the data in Fig. 13a, and the bottom boundary is set to be free drainage. According to the soil property information provided by the *ISMN*, the soil column is assumed to be heterogeneous with sandy loam material. The saturated soil moisture $\theta_s$ and residual soil moisture $\theta_r$ are given as 0.52 and 0.08 according to

the maximum and minimum values of soil moisture observations, while the other hydraulic parameters (i.e., $\ln K_s$, $\ln\alpha$ and $\ln n$) are assumed to be uncertain random fields following the settings of synthetic numerical experiments in Section 3.1 (since the information about variance and correlation length about the hydraulic parameter is not known). In this real-world case, we estimate all uncertain parameters simultaneously.

**4.3 Data assimilation results**

Three different temporal sampling strategies (i.e., ISS, RSS and IRSS) are used to verify the worth of observations and superiority of our optimal scheme. As summarized in Table 4, ISS is an intensive sampling strategy with the temporal interval of 1 day, which means the soil moisture data in a total of 120 days are all used to calibrate the model. In contrast, RSS and IRSS strategies are determined according to the temporal correlation length. First, the autocorrelation of soil moisture data in Lincoln_11_SW station is analyzed and the minimum temporal correlation length is elected as 6 days (see Fig. S1). Then, the optimal number of sampling days could be calculated as the ratio between total observation days (i.e., 120 days) and the minimum temporal correlation length (i.e., 6 days), which is equal to the 20. Regarding RSS strategy, the temporal interval is fixed with the value of 6 days during the whole calibration period. Whereas for IRSS strategy, we sample the less temporal points in dry period with the interval of 12 days, and more temporal points in wet period with the interval of 4 days.

For the real-world case, the performance of three sampling strategies is assessed by comparing the predicted (using the estimated parameters) and observed soil moisture

during the prediction period using the estimated parameters. The temporal changes of soil moisture predictions and observations as well as the *RMSE* results of each depth are presented in Fig. 14. Generally speaking, the accuracy of soil moisture predictions is similar for all three strategies, which agrees well with our results in Section 3. As supposed, the RSS strategy yields a slightly larger *RMSE* value for soil moisture predictions than the ISS strategy, but still guarantees robust estimates of soil moisture at each depth. At the same time, the number of observations used in RSS is only the 1/6 of those for ISS, which saves a lot of computational costs. Surprisingly, IRSS has the smallest *RMSE* among the three strategies at all depths, and the results of IRSS are even better than those of ISS. The simulation of real-world problems may have uncertainties that are not considered in data assimilation. For instance, the collection process of soil moisture data involves considerable uncertainties. Incorporating all data into the model may create an extra uncertainty in data assimilation system. In contrast, IRSS fully absorbed non-redundant information of soil moisture and avoids additional observation errors, which makes a better estimation of hydraulic parameter field.

## 5. Discussion

Our results showed that the key of improving the retrieval of heterogeneous field is to add the observations which contain the different information into the model, as presented in Fig. 12. Specifically, when the spatial/temporal sampling interval is set to be one spatial/temporal correlation length, the accuracy of parameter estimation is satisfactory with the affordable observation and computational costs (Figs. 6 and 9).

When comparing our results to those of previous studies (De Lannoy et al. 2007), it must be pointed out that our conclusions provide a quantitative method for the calculation of optimal spatial/temporal interval, which could be used in many different dimensions.

**5.1 The potential application of spatial sampling strategy**

The information of spatial correlation scale aids to determine the required spatial density of observation points: if the spatial correlation is large, a less dense network of soil moisture monitoring sites is sufficient. A similar conclusion was reached by De Lannoy et al. (2006) and Chaudhuri et al. (2018). However, in this study, the spatial correlation scale is given as a prior, which is always unknown in practice. As well as direct parameter measurements (Sudicky et al. 2010), the spatial structure information of soil property could also be determined by many non-invasive geostatistical methods, such as electrical resistivity tomography (ERT) method (Hübner et al. 2015) and apparent electrical conductivity (ECa) mapping method (Mertens et al. 2008). Based on this geological information and our results, a simple soil moisture monitoring network could be designed before installing the sensors.

In remote sensing, the utilization of surface soil moisture to retrieve the soil moisture/parameters in the root zone has been widely investigated, but it remains a challenge. Capehart and Carlson (1997) concluded that the information of remote sensing of soil moisture is limited and may not be useful to derive the soil state information in the deep layer. On the contrary, Ragab (1995) analyzed the soil moisture

data for two grass sites from 1992 and 1993 and found that the surface soil moisture has a great correlation with that in the root-zone. In this study, we conclude that surface observations do carry the considerable information about the parameters due to the alternative stimuli from precipitation and evaporation (Fig. 2), while the ability of deriving root-zone soil moisture/parameters from surface data depends on the correlation scale of soil hydraulic parameters (Fig. 7).

**5.2 The potential application of temporal sampling strategy**

With the development of sensor technology, soil moisture monitoring network could provide observations with unprecedented temporal resolution. For instance, the soil moisture data could be collected every 10 minutes by EnviroSCAN, which means more than fifty thousand of data in one year just at one depth (De Lannoy et al., 2006). However, assimilating all data into the model to retrieve the heterogenous field of vadose zone is impractical, owing to the large computational cost as well as the uncertainty of the data itself (Wang et al. 2018). Furthermore, soil moisture at adjacent time points has considerable redundancy of information (Fig. 8), making it unnecessary to incorporate the soil moisture data at very high frequency. Thus, our temporal strategy could provide the guidance in the selection of observation for big data analysis. And the results show that although the amount of observed data is reduced to 1/6 of the original data, the accuracy of soil moisture predictions is slightly affected (Fig. 14).

Moreover, sampling after a rainfall event improves the accuracy of the parameter estimation. With the same sampling times, the *RMSE* value can be reduced significantly

when the observations are collected after different rainfall events (Fig. 11). Therefore, we suggest sampling less in a dry period, and more in a wet period. This suggestion is reasonable owing to the non-stationarity of temporal correlation length of soil moisture. Clearly, the drainage process promotes the redistribution of soil moisture so that the new information on parameters for the deep soil can be provided (Fig. 12). In contrast, the autocorrelation of soil moisture is extremely large in dry periods, making the observations superfluous. Furthermore, this suggestion is also validated in a real-world case by using IRSS strategy (Fig. 14).

**5.3 Uncertainty and limitation of current work**

It should be noted that there are some uncertainties and limitations of the current study. Spatial correlation scale is assumed to be known and stable over the whole random field, which means the uncertainty of spatial correlation scale has not been considered in this study. Despite Yeh and Liu (2000) demonstrated that the uncertainty in correlation scale has little effects on estimates, the effects of horizontal correlation scale should be further investigated in two-dimensional soil water simulation. Moreover, the spatial sampling strategy has not been validated in reality owing to the constraint of our current experiment and data. In our future work, we will investigate the spatial structure information of vadose zone based on non-invasive geological inversion methods (e.g., ERT) and examine the applicability of the spatial strategies in practice.

The frequency and amount of precipitation and evaporation is irregular in a relatively short time series (128 days or 120 day), making it difficult to obtian a

representative temporal correlation length of soil moisture. A long, typical meteorology dataset may favor the use of temporal sampling strategy. Besides, soil property and boundary condition play a significant role in the autocorrelation of soil moisture which nevertheless has not been studied systematically. For example, the effects of hysteresis are ignored in this study, which may have imparcts on the cross-correlation analysis. Although we expect our major conclusions to hold, some differences could emerge if different model inputs (e.g., calibration period, soil property and bottom boundary condition) are used. Furthermore, the sampling frequency is set to be the same at all depths according to the minimum temporal correlation length. However, the stimuli origated from precipitation and evaporation would decline with the increase of soil depth, lead to a longer soil moisture memory at deeper layer (Fig. 8). Thus, sampling according to the minimum temporal correlation length may lead to the redundant information of soil moisture data at deep layers. A sampling scheme for dynamic changes in observation frequency with depth or finding a more representative temporal correlation length over the whole soil profile deserves further inverstigation.

Compared with pressure head, soil moisture is cheaper and easier to collect, and may be more sensitive to heterogeneity (Fig. 4). However, we only assess the worth of these observations independently, and the case of utilizing different observations simultaneously has not been investigated. Moreover, geological data are useful for subsurface water flow analysis at highly heterogeneous sites (Zhao and Illman 2018). It would be an interesting problem to utilize lithology information (e.g., ERT data) to

further constrain the inverse problem.

## 6. Conclusions

This paper proposes an optimal spatiotemporal sampling strategy for characterizing hydraulic properties of heterogeneous soils based on cross-correlation analysis. The effects of various model inputs such as observation types, meteorological conditions, bottom boundary conditions and correlation scale of soil are discussed. The results of the cross-correlation analysis and sampling strategies are validated by both synthetic and field numerical experiments for parameter field retrieval problems. Our work leads to the following major conclusions:

1. The value of observations varies in different meteorological conditions, bottom boundary conditions and observed depths. For the bare soil case, the surface observation under the free drainage boundary condition during the precipitation period can provide the most useful information, while the observation at the deep soil with a zero-flux boundary during the dry period contains highly redundant information.

2. The selection of observational types depends on water flow status over the soil profile, which is further affected by the bottom boundary conditions. Soil moisture provides a better retrieval result of soil heterogeneity than pressure head in unsaturated flow conditions (i.e., free drainage boundary), while in saturated flow areas, pressure head is preferred for the numerical stability of the simulations (i.e., zero-flux boundary).

3. Observations at the surface soils could be utilized to derive the root zone soil moisture when the vertical correlation scale of soil hydraulic parameter is large enough.

While in strongly heterogeneous soil, it is necessary to collect observations at various depths. The optimal number of spatial monitoring locations is equal to the ratio between the soil column length and the vertical correlation scale.

4. The following strategy is recommended for temporal sampling: (1) analyze the autocorrelation of soil moisture data and get the minimum temporal correlation length, (2) calculate the optimal number of temporal sampling points which is equal to the ratio between total number of the observation days and minimum temporal correlation length, (3) sample less during dry period but more in wet period according to the number of optimal temporal sampling points.

Further research may examine the performance of these sampling strategies in two- or three-dimensional variably saturated flow conditions. An experiment would be conducted to explore the spatial structure information of heterogeneous soil based on geological inversion methods, such as ERT method.

## References:

Allen RG, Pereira LS, Raes D, Smith M (1998) Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome 300:D05109

Bandara R, Walker JP, Rüdiger C (2013) Towards soil property retrieval from space: A one-dimensional twin-experiment. J Hydrol 497:198–207. https://doi.org/10.1016/j.jhydrol.2013.06.004

Cai J-S, Yan E-C, Yeh T-CJ, Zha Y (2017) Sampling schemes for hillslope
    hydrologic processes and stability analysis based on cross-correlation analysis.
    Hydrol Process 31:1301–1313. https://doi.org/10.1002/hyp.11101

Calvet J-C, Noilhan J, Bessemoulin P (1998) Retrieving the Root-Zone Soil Moisture
    from Surface Soil Moisture or Temperature Estimates: A Feasibility Study Based
    on Field Measurements. J Appl Meteorol 37:371–386.
    https://doi.org/10.1175/1520-0450(1998)037<0371:RTRZSM>2.0.CO;2

Carsel RF, Parrish RS (1988) Developing joint probability distributions of soil water
    retention characteristics. Water Resour Res 24:755–769.
    https://doi.org/10.1029/WR024i005p00755

Chao L, Ren L (2011) Estimation of Unsaturated Soil Hydraulic Parameters Using the
    Ensemble Kalman Filter. Vadose Zo J 10:1205.
    https://doi.org/10.2136/vzj2010.0159

Chaudhuri A, Franssen HJH, Sekhar M (2018) Iterative filter based estimation of fully
    3D heterogeneous fields of permeability and Mualem-van Genuchten
    parameters. Adv Water Resour 122:340–354.
    https://doi.org/10.1016/j.advwatres.2018.10.023

Chen W, Huang C, Shen H, Li X (2015) Comparison of ensemble-based state and
    parameter estimation methods for soil moisture data assimilation. Adv Water
    Resour 86:425–438. https://doi.org/10.1016/j.advwatres.2015.08.003

Chen Y, Oliver DS (2013) Levenberg-Marquardt forms of the iterative ensemble
    smoother for efficient history matching and uncertainty quantification. Comput
    Geosci 17:689–703. https://doi.org/10.1007/s10596-013-9351-5

Chen Z, Govindaraju RS, Kavvas ML (1994) Spatial averaging of unsaturated flow
    equations under infiltration conditions over areally heterogeneous fields 2.
    Numerical simulations. Water Resour Res 30:535–548

Christakos G (2017) Spatiotemporal random fields: theory and applications. Elsevier

Crestani E, Camporese M, Baú D, Salandin P (2013) Ensemble Kalman filter versus

ensemble smoother for assessing hydraulic conductivity via tracer test data assimilation. Hydrol Earth Syst Sci 17:1517–1531. https://doi.org/10.5194/hess-17-1517-2013

Dai C, Xue L, Zhang D, Guadagnini A (2016) Data-worth analysis through probabilistic collocation-based Ensemble Kalman Filter. J Hydrol 540:488–503. https://doi.org/10.1016/j.jhydrol.2016.06.037

Das NN, Mohanty BP (2006) Root Zone Soil Moisture Assessment Using Remote Sensing and Vadose Zone Modeling. Vadose Zo J 5:296. https://doi.org/10.2136/vzj2005.0033

De Lannoy GJM, Houser PR, Pauwels VRN, Verhoest NEC (2007) State and bias estimation for soil moisture profiles by an ensemble Kalman filter: Effect of assimilation depth and frequency. Water Resour Res 43:1–15. https://doi.org/10.1029/2006WR005100

De Lannoy GJM, Verhoest NEC, Houser PR, et al (2006) Spatial and temporal characteristics of soil moisture in an intensively monitored agricultural field (OPE3). J Hydrol 331:719–730. https://doi.org/10.1016/j.jhydrol.2006.06.016

ELSheikh AH, Pain CC, Fang F, et al (2013) Parameter estimation of subsurface flow models using iterative regularized ensemble Kalman filter. Stoch Environ Res Risk Assess 27:877–897. https://doi.org/10.1007/s00477-012-0613-x

Emerick AA, Reynolds AC (2013) Ensemble smoother with multiple data assimilation. Comput Geosci 55:3–15. https://doi.org/10.1016/j.cageo.2012.03.011

Erdal D, Neuweiler I, Huisman JA (2012) Estimating effective model parameters for heterogeneous unsaturated flow using error models for bias correction. Water Resour Res 48:1–19. https://doi.org/10.1029/2011WR011062

Erdal D, Neuweiler I, Wollschläger U (2014) Using a bias aware EnKF to account for unresolved structure in an unsaturated zone model. Water Resour Res 50:132–147. https://doi.org/10.1002/2012WR013443

Evensen G (2009) The ensemble Kalman filter for combined state and parameter estimation. IEEE Control Syst Mag 29:83–104. https://doi.org/10.1109/MCS.2009.932223

Hoeben R (2000) Assimilation of Active Microwave Measurements for Soil Moisture Profile Retrieval Under Laboratory Conditions. Ieee 36:1271–1273

Huang C, Hu BX, Li X, Ye M (2009) Using data assimilation method to calibrate a heterogeneous conductivity field and improve solute transport prediction with an unknown contamination source. Stoch Environ Res Risk Assess 23:1155–1167. https://doi.org/10.1007/s00477-008-0289-4

Hübner R, Heller K, Günther T, Kleber A (2015) Monitoring hillslope moisture dynamics with surface ERT for enhancing spatial significance of hydrometric point measurements. Hydrol Earth Syst Sci 19:225–240. https://doi.org/10.5194/hess-19-225-2015

Hughson DL, Yeh T-CJ (2000) An inverse model for three-dimensional flow in variably saturated porous media. Water Resour Res 36:829–839. https://doi.org/10.1007/978-3-030-18383-7_6

Illman WA, Tartakovsky DM (2005) Asymptotic analysis of cross-hole pneumatic injection tests in unsaturated fractured tuff. Adv Water Resour 28:1217–1229. https://doi.org/10.1016/j.advwatres.2005.03.011

Jhorar RK, Bastiaanssen WGM, Feddes RA, Van Dam JC (2002) Inversely estimating soil hydraulic functions using evapotranspiration fluxes. J Hydrol 258:198–213. https://doi.org/10.1016/S0022-1694(01)00564-9

Ju L, Zhang J, Meng L, et al (2018) An adaptive Gaussian process-based iterative ensemble smoother for data assimilation. Adv Water Resour 115:125–135. https://doi.org/10.1016/j.advwatres.2018.03.010

Kool JB, Parker JC (1988) Analysis of the inverse problem for transient unsaturated flow. Water Resour Res 24:817–830. https://doi.org/10.1029/WR024i006p00817

Koster RD, Suarez MJ (2001) Soil Moisture Memory in Climate Models. J

Hydrometeorol 2:558–570. https://doi.org/10.1175/1525-7541(2001)002<0558:SMMICM>2.0.CO;2

Leube PC, Geiges A, Nowak W (2012) Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. Water Resour Res 48:1–16. https://doi.org/10.1029/2010WR010137

Li B, Yeh TCJ (1998) Sensitivity and moment analyses of head in variably saturated regimes. Adv Water Resour 21:477–485. https://doi.org/10.1016/S0309-1708(97)00011-0

Li J, Islam S (1999) On the estimation of soil moisture profile and surface fluxes partitioning from sequential assimilation of surface layer soil moisture. J Hydrol 220:86–103. https://doi.org/10.1016/S0022-1694(99)00066-9

Li X, Shi L, Zha Y, et al (2018) Data assimilation of soil water flow by considering multiple uncertainty sources and spatial–temporal features: a field-scale real case study. Stoch Environ Res Risk Assess 32:2477–2493. https://doi.org/10.1007/s00477-018-1541-1

Man J, Zhang J, Li W, et al (2016) Sequential ensemble-based optimal design for parameter estimation. Water Resour Res 52:7577–7592. https://doi.org/10.1002/2016WR019285.Received

Man J, Zheng Q, Wu L, Zeng L (2020) Adaptive multi-fidelity probabilistic collocation-based Kalman filter for subsurface flow data assimilation: numerical modeling and real-world experiment. Stoch Environ Res Risk Assess 34:1135–1146. https://doi.org/10.1007/s00477-020-01815-y

Mantoglou A (1992) A theoretical approach for modeling unsaturated flow in spatially variable soils: Effective flow models in finite domains and nonstationarity. Water Resour Res 28:251–267. https://doi.org/10.1029/91WR02232

Mao D, Yeh TCJ, Wan L, et al (2013) Cross-correlation analysis and information content of observed heads during pumping in unconfined aquifers. Water Resour

Res 49:713–731. https://doi.org/10.1002/wrcr.20066

Mertens FM, Pätzold S, Welp G (2008) Spatial heterogeneity of soil properties and its
mapping with apparent electrical conductivity. J Plant Nutr Soil Sci 171:146–
154. https://doi.org/10.1002/jpln.200625130

Mertens J, Stenger R, Barkle GF (2006) Multiobjective Inverse Modeling for Soil
Parameter Estimation and Model Verification. Vadose Zo J 5:917–933.
https://doi.org/10.2136/vzj2005.0117

Montzka C, Moradkhani H, Weihermüller L, et al (2011) Hydraulic parameter
estimation by remotely-sensed top soil moisture observations with the particle
filter. J Hydrol 399:410–421. https://doi.org/10.1016/j.jhydrol.2011.01.020

Mualem Y (1976) Wetting front pressure head in the infiltration model of Green and
Ampt. Water Resour Res 12:564–566.
https://doi.org/10.1029/WR012i003p00564

Neuman SP, Feddes RA, Bresler E, (1974) Finite element simulation of flow in
saturated - unsaturated soils considering water uptake by plants, Third Annual
Report, Project No. A10-SWC-77, Hydraulic Engineering Lab., Technion, Haifa,
Israel

Paltineanu IC, Starr JL (1997) Real-time Soil Water Dynamics Using Multisensor
Capacitance Probes: Laboratory Calibration. Soil Sci Soc Am J 61:1576.
https://doi.org/10.2136/sssaj1997.03615995006100060006x

Pan L, Wu L (1998) A hybrid global optimization method for inverse estimation of
hydraulic parameters: Annealing-simplex method. Water Resour Res 34:2261–
2269

Pauwels VRN, Verhoest NEC, De Lannoy GJM, et al (2007) Optimization of a
coupled hydrology-crop growth model through the assimilation of observed soil
moisture and leaf area index values using an ensemble Kalman filter. Water
Resour Res 43:1–17. https://doi.org/10.1029/2006WR004942

Pujol J (2007) The solution of nonlinear inverse problems and the Levenberg-

Marquardt method. Geophysics 72:W1–W16

Ragab R (1995) Towards a continuous operational system to estimate the root-zone soil moisture from intermittent remotely sensed surface moisture. J Hydrol 173:1–25

Rocha D, Abbasi F, Feyen J (2006) Sensitivity analysis of soil hydraulic properties on subsurface water flow in furrows. J Irrig Drain Eng 132:418–424. https://doi.org/Doi 10.1061/(Asce)0733-9437(2006)132:4(418)

Šimůnek J, van Genuchten MT (1996) Estimating Unsaturated Soil Hydraulic Properties from Tension Disc Infiltrometer Data by Numerical Inversion. Water Resour Res 32:2683–2696. https://doi.org/10.1029/96WR01525

Sudicky EA, Illman WA, Goltz IK, et al (2010) Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: From measurements to a practical application of stochastic flow and transport theory. Water Resour Res 46:1–16. https://doi.org/10.1029/2008wr007558

Sun R, Yeh TCJ, Mao D, et al (2013) A temporal sampling strategy for hydraulic tomography analysis. Water Resour Res 49:3881–3896. https://doi.org/10.1002/wrcr.20337

Tong J, Hu BX, Yang J (2010) Using data assimilation method to calibrate a heterogeneous conductivity field conditioning on transient flow test data. Stoch Environ Res Risk Assess 24:1211–1223. https://doi.org/10.1007/s00477-010-0392-1

van Dam JC, Stricker JNM, Droogers P (1992) Inverse Method for Determining Soil Hydraulic Functions from One-Step Outflow Experiments. Soil Sci Soc Am J 56:1042. https://doi.org/10.2136/sssaj1994.03615995005800030002x

van Genuchten MT (1980) A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils1. Soil Sci Soc Am J 44:892. https://doi.org/10.2136/sssaj1980.03615995004400050002x
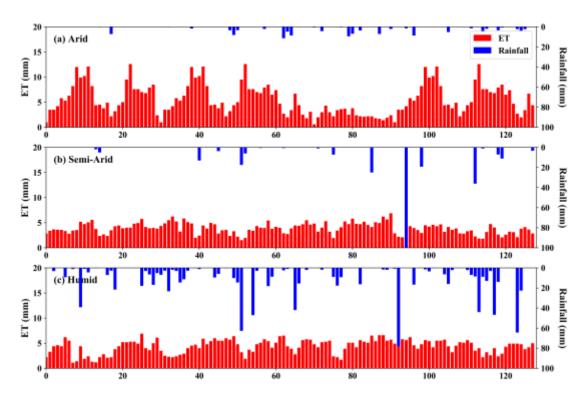
Vereecken H, Huisman JA, Bogena H, et al (2010) On the value of soil moisture

measurements in vadose zone hydrology: A review. Water Resour Res 46:1–21. https://doi.org/10.1029/2008WR006829

Vereecken H, Schnepf A, Hopmans JW, et al (2016) Modeling Soil Processes: Review, Key Challenges, and New Perspectives. Vadose Zo J 15:. https://doi.org/10.2136/vzj2015.09.0131

Walker JP, Willgoose GR, Kalma JD (2002) Three-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: Simplified Kalman filter covariance forecasting and field application. Water Resour Res 38:37–1. https://doi.org/10.1029/2002WR001545

Walker JP, Willgoose GR, Kalma JD (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. https://doi.org/10.1175/1525-7541(2001)002<0356:ODSMPR>2.0.CO;2

Wang Y, Shi L, Zha Y, et al (2018) Sequential data-worth analysis coupled with ensemble Kalman filter for soil water flow: A real-world case study. J Hydrol 564:76–88. https://doi.org/10.1016/j.jhydrol.2018.06.059

Webster R, Oliver MA (1990) Statistical methods in soil and land resource survey. Oxford University Press (OUP)

William J. Capehart and Toby N. Carlson (1997) Decoupling of surface and near-surface soil water content: A remote sensing perspective. Water Resour Res 33:1383–1395

Wu CM, Yeh TCJ, Zhu J, et al (2005) Traditional analysis of aquifer tests: Comparing apples to oranges? Water Resour Res 41:1–12. https://doi.org/10.1029/2004WR003717

Yeh T-CJ, Liu S (2000a) Hydraulic tomography: Development of a new aquifer test method. Water Resour Res 36:2095. https://doi.org/10.1029/2000WR900114

Yeh TCJ, Liu S (2000b) Hydraulic tomography: Development of a new aquifer test method. Water Resour Res 36:2095–2105.

https://doi.org/10.1029/2000WR900114

Younes A, Mara T, Fahs M, et al (2017) Hydraulic and transport parameter assessment using column infiltration experiments. Hydrol Earth Syst Sci 21:2263–2275. https://doi.org/10.5194/hess-21-2263-2017

Yu D, Yang J, Shi L, et al (2019) On the uncertainty of initial condition and initialization approaches in variably saturated flow modeling. Hydrol Earth Syst Sci 23:2897–2914

Zhang Y, Schaap MG, Guadagnini A, Neuman SP (2016) Inverse modeling of unsaturated flow using clusters of soil texture and pedotransfer functions. Water Resour Res 52:7631–7644. https://doi.org/10.1002/2016WR019016

Zhao Z, Illman WA (2018) Three-dimensional imaging of aquifer and aquitard heterogeneity via transient hydraulic tomography at a highly heterogeneous field site. J Hydrol 559:392–410. https://doi.org/10.1016/j.jhydrol.2018.02.024

Zijlstra J, Dane JH (1996) Identification of hydraulic parameters in layered soils based on a quasi-Newton method. J Hydrol 181:233–250. https://doi.org/10.1016/0022-1694(95)02909-5

**Fig. 1.** Synthetic rainfall (blue bars) and reference evapotranspiration (red bars) of three typical climates including (a) arid climate, (b) semi-arid climate, and (c) humid climate.
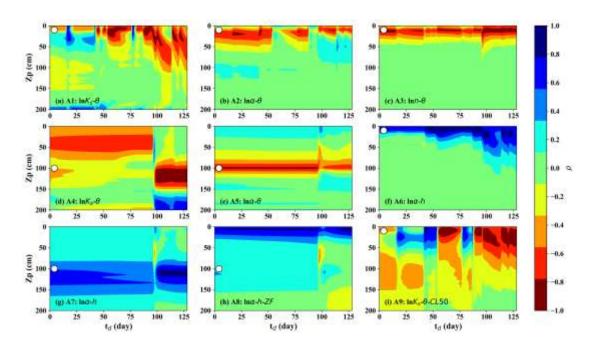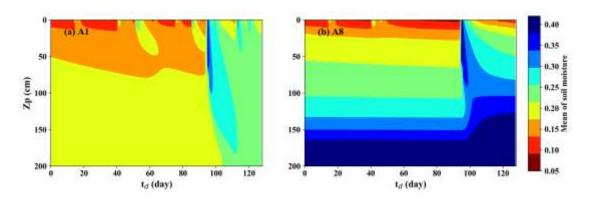
**Fig. 2.** Cross-correlation map $\rho$ ($z_p$, 10, $t_d$) or $\rho$ ($z_p$, 100, $t_d$) with different bottom boundary conditions and correlation lengths (Case 1). $z_p$ is the depth of hydraulic parameter which is from 0 to 200 cm, while 10 and 100 (cm) indicate the depth of observation (white circle in the figure). $t_d$ is the simulation time which is from 0 to 128 days.

**Fig. 3.** Temporal change of the soil moisture profile with (a) free drainage boundary for A1, (b) zero-flux boundary for A8.

**Fig. 4.** The results of the individual parameter estimation for Case 2.

**Fig. 5.** *RMSE* versus correlation length ratios (of 1/8, 1/4, 1/2, 1, 2.5, and 5 respectively) for (a) ln*Ks*, (b) ln*α* and (c) ln*n*. *RMSE* is calculated as the differences between estimated and reference parameters with results obtained from 3 runs (see text for details); red square indicates the mean *RMSE* value and the solid lines represent standard deviations of *RMSE*.
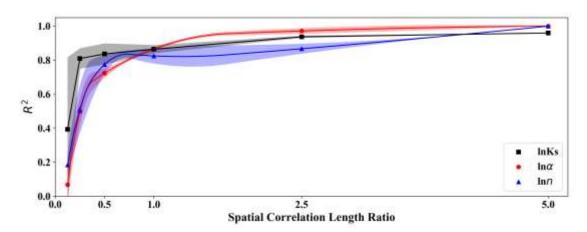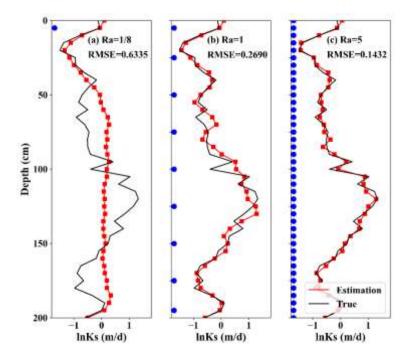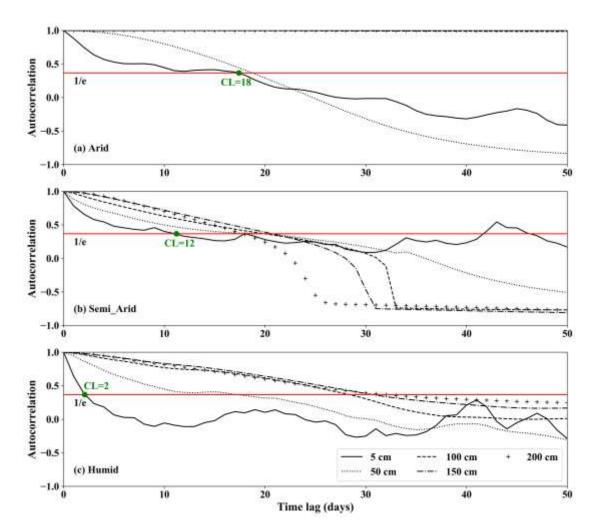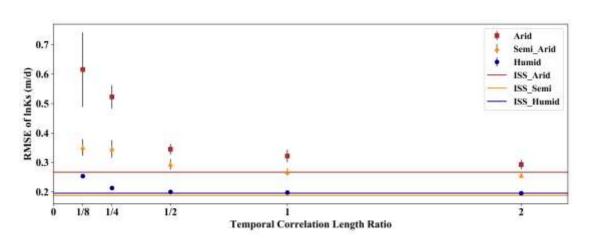
**Fig. 6.** Mean values of determination coefficient $R^2$ between the estimated and reference parameters for ln$K_s$ (black square), ln$\alpha$ (red circle) and ln$n$ (blue triangle) versus correlation length ratios. The results are from 3 runs. The shadowed area below or above the point indicates one standard deviation of *RMSE*.
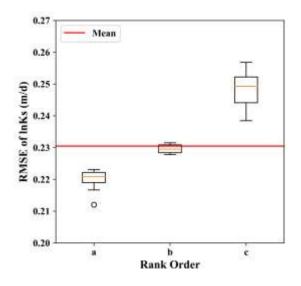
**Fig. 7.** Results of estimated ln$K_s$ random field with various *Ra* values ((a) 1/8, (b) 1 and (c) 5).

It should be noted that the blue circles indicate the locations of observations.
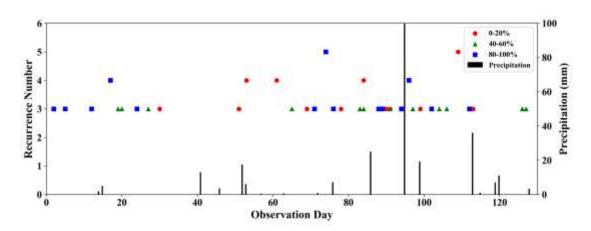
**Fig. 8.** Temporal autocorrelation and corresponding correlation length (CL) of soil moisture under different meteorological conditions. Remark that the autocorrelation function at 100, 150 and 180 cm depth are overlapping for arid condition.

**Fig. 9.** *RMSE* results of ln$K_s$ estimation with different temporal strategies (values on x axis) and meteorological conditions (i.e., arid, semi-arid, and humid). The results are statistical values from three simulations (see text for details). The colored point indicates the mean *RMSE* value and the solid black line represents one standard deviation of *RMSE*. The horizontal line represents the estimated *RMSE* value from the intensive sampling strategy (ISS) with the temporal interval of 1 day. Remark that the solid black line in humid climate is overlapping.

**Fig. 10.** (a) optimal solutions (first 20%: $1^{st}$ to $10^{th}$), (b) sub-optimal solutions (40% to 60%: $21^{th}$ to $30^{th}$), and (c) the worst solutions (80% to 100%: $41^{th}$ to $50^{th}$) in 50 random field inversions results. The classification is determined by ranking the *RMSE* values in ascending order. The red line indicates the mean *RMSE* value of all ensemble.

**Fig. 11.** The relationship between precipitation and recurrence number of observation day with three various parameter estimation accuracy levels, including optimal solutions (red circles), sub-optimal solutions (green triangles) and the worst solutions (blue squares). The left vertical axis indicates the recurrence number of observation days, while the right axis denotes the amount of precipitation. Low-frequency observation days (less than three times) are not shown here for the sake of brevity.
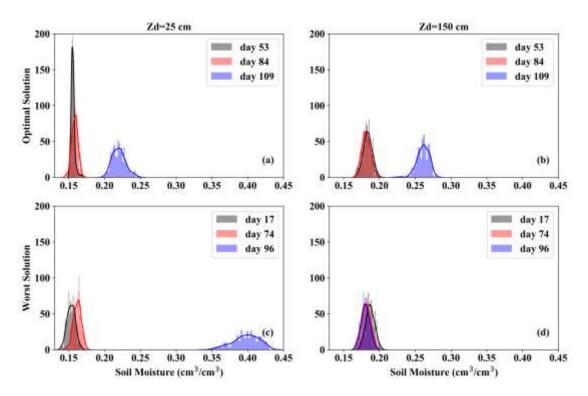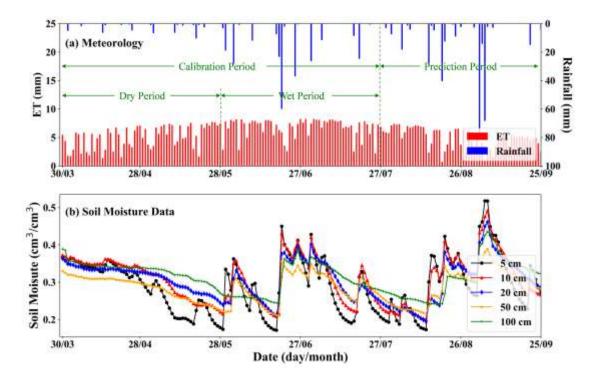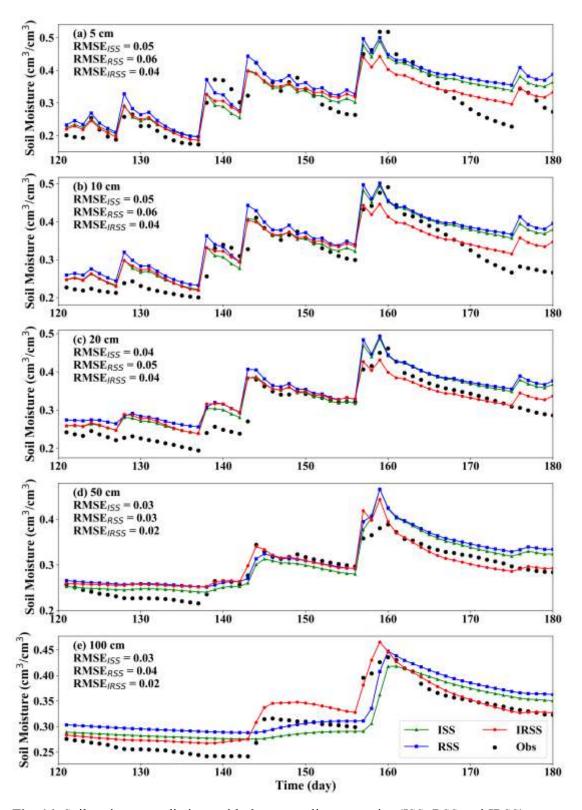
**Fig. 12.** The prior probability density distribution of soil moisture ensemble at various spatial locations (graphs at columns: $z_d$ = 25 cm and 150 cm respectively) and temporal sampling time (recurrence number greater than three) of (a-b) the optimal solutions (53rd, 84th and 109th days) and (c-d) the worst solutions (17th, 74th and 96th days).

**Fig. 13.** The meteorological information and observed soil moisture in the real-world experiment. The first 120 days are calibration period (with a dry period from 1th day to 60th and a wet period from 61th day to 120th day), and the last 60 days are used for prediction.

**Fig. 14.** Soil moisture predictions with three sampling strategies (ISS, RSS and IRSS) versus observations at five measurement depths in a real-world case study.

**Table 1.** The default settings used in variably saturated flow models.

| Model inputs | Settings |
| --- | --- |
| Initial condition | Spinning up the model with one-year meteorology data |
| Upper boundary | Semi-arid climate |
| Bottom boundary | Free drainage |
| Correlation length | 25 cm |
| Number of soil layers | 40 |
| Thickness of soil zone | 2 m |
| Number of grids | 40 (with a size of 5 cm) |
| Simulation time | 128 days |

Table 2. Case summary for cross-correlation analysis (Case1).

| Scenario | Parameter | Observation type | Observation depth (cm) | Bottom boundary | Correlation length (cm) |
|----------|-----------|------------------|------------------------|-----------------|-------------------------|
| A1 | $\ln K_s$ | $\theta$ | 10 | - | - |
| A2 | $\ln \alpha$ | $\theta$ | 10 | - | - |
| A3 | $\ln n$ | $\theta$ | 10 | - | - |
| A4 | $\ln K_s$ | $\theta$ | 100 | - | - |
| A5 | $\ln \alpha$ | $\theta$ | 100 | - | - |
| A6 | $\ln \alpha$ | $h$ | 10 | - | - |
| A7 | $\ln \alpha$ | $h$ | 100 | - | - |
| A8 | $\ln \alpha$ | $h$ | 100 | Zero-flux (ZF) | - |
| A9 | $\ln K_s$ | $\theta$ | 10 | - | 50 |

Note: Ungiven parameters use the default value listed in Table 1.

**Table 3.** Case summary for parameter estimation with various observation types, errors, and bottom boundary conditions (Case 2).

| Scenario | Parameter | Observation type | Observation interval (m) | Observation error | Bottom boundary |
|----------|-----------|------------------|--------------------------|-------------------|-----------------|
| B1 | $\ln K_s$ | $\theta$ | 0.5 | 0 | - |
| B2 | $\ln n$ | $\theta$ | 0.5 | 0 | - |
| B3 | $\ln\alpha$ | $\theta$ | 0.5 | 0 | - |
| B4 | $\ln\alpha$ | $h$ | 0.5 | 0 | - |
| B5 | $\ln\alpha$ | $h$ | 0.5 | 0 | Zero-flux |
| B6 | $\ln K_s$ | $\theta$ | 0.5 | 0.01 | - |

Note: - represents the default values listed in Table 1.

**Table 4.** Case summary for soil moisture predictions with various temporal sampling

strategies.

| Scenario | Number of sampling day | | Description |
|---|---|---|---|
| | Dry period | Wet period | |
| ISS | 60 | 60 | Intensive sampling strategy with the temporal interval of 1 day. |
| RSS | 10 | 10 | Regular sampling strategy with the temporal interval equal to the minimum temporal correlation length (i.e., $Ra$ =1). |
| IRSS | 5 | 15 | Irregular sampling strategy with less sampling points in dry period. The number of sampling dates is equal to the RSS strategy. |