


ORIGINAL RESEARCH ARTICLE

A robust data-worth analysis framework for soil moisture flow by hybridizing sequential data assimilation and machine learning

Yakun Wang¹ | Liangsheng Shi¹  | Lin Lin¹ | Mauro Holzman² | Facundo Carmona² | Qiuru Zhang¹

¹State Key Lab. of Water Resources and Hydropower Engineering Sciences, Wuhan Univ., Wuhan, Hubei, 430072, China

²Instituto de Hidrología de Llanuras “Dr. Eduardo J. Usunoff,” CONICET, UNCPBA-IHLLA, Azul-Tandil, Argentina

Correspondence

Liangsheng Shi, State Key Lab. of Water Resources and Hydropower Engineering Sciences, Wuhan Univ., Wuhan, Hubei 430072, China.

Email: liangshs@whu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 51779180, 51861125202

Abstract

As the collection of soil moisture data is often costly, it is essential to implement data-worth analysis in advance to obtain a cost-effective data collection scheme. In previous data-worth analysis, the model structural error is often neglected. In this paper, we propose a robust data-worth analysis framework based on a hybrid data assimilation method. By constructing Gaussian process (GP) error model, this study attempts to alleviate biased data-worth assessments caused by unknown model structural errors, and to excavate complementary values of multisource data without resorting to multiple governing equations. The results demonstrated that this proposed framework effectively identified and compensated for complex model structural errors. By training prior data, more accurate potential observations were obtained and data-worth estimation accuracy was improved. The scenario diversity played a crucial role in establishing an effective GP training system. The integration of soil temperature into GP training unraveled new information and improved the data-worth estimation. Instead of traditional evapotranspiration calculations, the direct inclusion of easy-to-obtain meteorological data into GP training yielded better data-worth assessment.

1 | INTRODUCTION

During the past few decades, soil moisture measurement technology has experienced a surge in development. Direct observations with ground instruments (Li, Shi, Zha, Wang, & Hu, 2018; Walker, Willgoose, & Kalma, 2001) and indirect observations with remote sensing techniques (Crow & Wood, 2003; Wagner, Lemoine, & Rott, 1999) offer unique opportunities to advance soil hydrology. Massive data of different types, scales, frequencies, and accuracies are accumulating at an unprecedented rate, which hinders the efficient use

of such data. Clearly, we need to implement data-worth analysis to selectively collect the most informative data at the lowest possible cost. As defined by Neuman, Xue, Ye, and Lu (2012) and Geiges, Rubin, and Nowak (2015), data-worth analysis quantifies the amount of relevant information that a proposed data collection program is expected to provide. In the past few decades, data worth has been widely investigated in the field of hydrology (Ben-Zvi, Berkowitz, & Kesler, 1988; Davis, Kisiel, & Duckstein, 1972; Man, Zhang, Li, Zeng, & Wu, 2016). The basic principle of data-worth analysis is Bayesian analysis that considers the impact of data on model predictive uncertainty (Neuman et al., 2012). Various information indicators, including the trace, Shannon entropy difference, relative entropy, and degrees of freedom for a signal, have been used to quantify the data worth (Man et al.,

Abbreviations: EnKF, ensemble Kalman filter; GP, Gaussian process; ISMN, International Soil Moisture Network; MOL-RAO, Meteorological Observatory Lindenberg-Richard Aßmann Observatory.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Vadose Zone Journal* published by Wiley Periodicals, Inc. on behalf of Soil Science Society of America

2016; Wang et al., 2018). A typical data-worth framework consists of prior, posterior, and preposterior stages (Dai, Xue, Zhang, & Guadagnini, 2016). The preposterior analysis evaluates the anticipated worth of future observations, for which possible distributions are predicted in advance by conditioning on prior data.

As recognized in the literature (Beven, 2005; Oreskes, Shrader-Frechette, & Belitz, 1994; Shi, Song, Tong, Zhu, & Zhang, 2015; Xu, Valocchi, Choi, & Amir, 2014; Zha, Zhu, Zhang, Mao, & Shi, 2019), model predictions are inherently uncertain due to the uncertainties from input data, model parameters, model structures, missing physics, and numerical implementation (numerical algorithms, spatial and temporal discretization; Beven, 2005). Once data are used to calibrate a soil moisture model, it is commonly assumed that the model has a fixed form of presentation. In our previous study (Wang et al., 2018), we assessed the data worth of observations in a sequential way by allowing soil hydraulic parameters to be updated. In this sequential data-worth analysis framework, only the most informative observation is selected, which can effectively avoid data redundancy and reduce the cost of data collection (Dai et al. 2016; Man et al., 2016). Data-worth analysis can also prevent creating an extra uncertainty from excessive observations in data assimilation system (Wang et al., 2018). We found that potential observations that will be collected in the future and generated based on prior available data may obviously deviate from actual measurements due to the existence of model structural errors. The accuracy of data-worth evaluation degrades significantly when facing complex real-world conditions, especially for the mean-covariance-type index (e.g., relative entropy). It is necessary to remove potential adverse effects from inaccurate representation or omission of physical processes at the preposterior stage.

A few approaches have been proposed to deal with model structural errors within the framework of data assimilation, such as inflation of the background covariance (Anderson & Anderson, 1999; Hamill & Whitaker, 2005) and bias correction methods (De Lannoy, Houser, Pauwels, & Verhoest, 2007; Drécourt, Madsen, & Rosbjerg, 2006). Nevertheless, in both “covariance inflation” and “bias estimation” approaches, model structural errors are inclined to interact with other sources of model error due to the introduction of total or lumped model error (Pathiraja, Moradkhani, Marshall, Sharma, & Geenens, 2018; Zupanski & Zupanski, 2006). Furthermore, “covariance inflation” approaches require ad hoc tuning factors, which are difficult to determine. Almost all “bias estimation” methods rely on the assumption that the model error covariance is proportional to the state error covariance (Pauwels & De Lannoy, 2015).

To circumvent these drawbacks, a family of statistical learning techniques was proposed to build an error model from an inductive, data-driven perspective. Artificial neural network models were trained to forecast the residual of

Core Ideas

- A new data-worth analysis framework was proposed.
- The new hybrid approach can alleviate biased data-worth assessment caused by model structural error.
- The hybrid method offers an effective approach to excavate complementary value of multisource data.

conceptual rainfall-runoff models (Abebe & Price, 2003). Demissie, Valocchi, Minsker, and Bailey (2009) developed a framework to handle systematic error in a physical-based groundwater model, in which several error-correcting, data-driven models were considered. Xu and Valocchi (2016) adopted a fully Bayesian approach that integrates a Gaussian process (GP) error model with uncertainty analysis of groundwater flow. In Zhang et al. (2019), we proposed a sequential data assimilation scheme by hybridizing an iterative ensemble Kalman filter and GP (EnKF-GP). The main advantage of these statistical learning approaches lies in not requiring explicit representation of the model residual distribution. Instead, they learn complex relationships between the dependent variable (i.e., model structural error) and select predictors from historical data. Therefore, they are good candidates to statistically characterize the model structural error.

In this study, we further integrate the hybrid data assimilation approach (Zhang et al., 2019) into our previous sequential data-worth analysis framework (Wang et al., 2018). Once the potential predictions are obtained, data-worth analysis is implemented to quantify the worth of alternative monitoring strategy (in terms of observation location, frequency, data type, etc.). Consequently, the most informative observations can be collected to reduce monitoring costs. The objective of using this hybrid approach is to alleviate the possible damage of model structural error on data-worth assessment, so that a more robust data-worth analysis can be made in complex environments.

In our previous paper (Zhang et al., 2019), only information directly related to soil water movement was used to train the GP system. However, with the innovation of measurement technology, multisource data can now be collected simultaneously. Some variables (e.g., soil temperature) can be easily observed, and these indirect observations may contain valuable information on soil moisture dynamics. This study further quantifies the worth of several indirect data by not involving them in the equation. We hope to relax the requirement of physical model equations by showing that direct (soil moisture) and indirect data (soil temperature and

meteorological data) can both make important contributions to data-worth estimation under unresolved model inadequacy.

In this context, Section 2 presents the principles of the modified restart EnKF, GP, and hybrid data-worth analysis framework. The experimental data from Falkenberg Station of the Meteorological Observatory Lindenberg-Richard Aßmann Observatory (MOL-RAO) network (from the International Soil Moisture Network [ISMN]) and model setup are described in Section 2.5. Section 3 presents a set of examples to demonstrate the ability of the proposed framework to improve data-worth estimation accuracy (covering aspects of scenario diversity, prior data content, training input augmentation, and replacement of evapotranspiration calculation). Finally, conclusions are drawn in Section 4.

2 | MATERIALS AND METHODS

In Zhang et al. (2019), two variants of a hybrid data assimilation method (EnKF-GP) were proposed, including no feedback of error correction for model reinitialization (named EnKF-GP1) and feeding back error-corrected state variables for model reinitialization (named EnKF-GP2). This paper uses the EnKF-GP2 approach during data-worth analysis because of its simpler implementation.

2.1 | Governing equation of one-dimensional soil moisture flow

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

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

where θ [$L^3 L^{-3}$] 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; and K [$L T^{-1}$] is the unsaturated hydraulic conductivity.

The solution of Richards' equation requires the knowledge of the unsaturated conductivity and soil moisture vs. hydraulic head. The van Genuchten–Mualem model (van Genuchten, 1980) is used to describe these constitutive relationships:

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

$$K h = \begin{cases} K_s S_e^{1/2} \left[1 - \left(1 - S_e^{1/m} \right)^m \right]^2, & h < 0 \\ K_s, & h \geq 0 \end{cases} \quad (3)$$

$$m = 1 - \frac{1}{n}, \quad n > 1 \quad (4)$$

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

where θ_s [$L^3 L^{-3}$] and θ_r [$L^3 L^{-3}$] are the saturated and residual moisture content, respectively; α [L^{-1}] and n [–] are the shape parameters of the soil water characteristic curve; K_s [$L T^{-1}$] is the saturated hydraulic conductivity; and S_e is the effective saturation. It is noted that heat transfer equation is not included in this study, even though temperature data provide an auxiliary data source during data-worth assessment. The Richards' equation is solved by the Ross method (Zha, Shi, Ye, & Yang, 2013).

2.2 | The modified restart ensemble Kalman filter

The classical EnKF cannot guarantee the physical consistency in nonlinear system after updating (i.e., the updated soil moisture content and soil hydraulic parameters do not follow the Richards' equation). Such inconsistency may cause severe performance deterioration of data assimilation in a strongly nonlinear soil water problem. In this study, the modified restart EnKF (Song, Shi, Ye, Yang, & Navon, 2014) that has better physical consistency is used to implement the coupling of the GP regression and the EnKF, while retaining the computational cost at an acceptable level. Here, we give a brief introduction to this filter, and more details can be found in Sun, Wang, and Xu (2014).

The parameter vector of interest, \mathbf{p}_k (e.g., soil hydraulic parameters), is augmented with the state variable vector, \mathbf{s}_k (e.g., soil moisture), into a joint state vector, $\mathbf{y}_k = [\mathbf{p}_k, \mathbf{s}_k]^T$, at time t_k (k is the time step index). A collection of N_1 members of the state vector \mathbf{Y}_k can be written as

$$\mathbf{Y}_k = \left\{ \mathbf{y}_{1,k}, \mathbf{y}_{2,k}, \dots, \mathbf{y}_{N_1,k} \right\} \quad (6)$$

For a set of observations $\mathbf{d}_k^{\text{obs}}$ available at time $t = t_k$, their relationship with the true but unknown state variable $\mathbf{d}_k^{\text{true}}$ and the true augmented state vector $\mathbf{y}_k^{\text{true}}$ can be expressed as

$$\mathbf{d}_k^{\text{obs}} = \mathbf{d}_k^{\text{true}} + \boldsymbol{\varepsilon}_k = \mathbf{H} \mathbf{y}_k^{\text{true}} + \boldsymbol{\varepsilon}_k \quad (7)$$

where $\boldsymbol{\varepsilon}_k$ represents observation error, which is assumed to be zero-mean Gaussian with a covariance of $\mathbf{C}_{D_k} = E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k^T)$; the matrix \mathbf{H} represents the observation operator, which relates the state vector \mathbf{y} and observation vector \mathbf{d}^{obs} .

For any ensemble member i at time t_k , the state vector can be updated by combining model predictions and observations, using

$$\mathbf{y}_{i,k}^a = \mathbf{y}_{i,k}^f + \mathbf{K}_k (\mathbf{d}_{i,k}^{\text{obs}} - \mathbf{H}\mathbf{y}_{i,k}^f) \quad (8)$$

where $i = 1, 2, \dots, N_1$; the superscripts “f” and “a” refer to forecast and analysis, respectively; the subscript i is the ensemble member index; $\mathbf{y}_{i,k}^f$ represents the model forecast for realization i at time t_k based on information at time t_{k-1} ; $\mathbf{y}_{i,k}^a$ is the model analysis given by conditioning on the observations at time t_k ; and \mathbf{K}_k is the Kalman gain, which is given by

$$\mathbf{K}_k = \mathbf{C}_k^f \mathbf{H}^T (\mathbf{H} \mathbf{C}_k^f \mathbf{H}^T + \mathbf{C}_{D_k})^{-1} \quad (9)$$

where the covariance matrix at time t_k , and \mathbf{C}_k^f can be estimated by

$$\mathbf{C}_k^f \approx \frac{1}{N_1 - 1} \sum_{i=1}^{N_1} \left\{ \left[\mathbf{y}_{i,k}^f - \bar{\mathbf{y}}_k^f \right] \left[\mathbf{y}_{i,k}^f - \bar{\mathbf{y}}_k^f \right]^T \right\} \quad (10)$$

where $\bar{\mathbf{y}}_k^f$ refers to the ensemble mean of \mathbf{y}_k^f .

After assimilating the observation $\mathbf{d}_k^{\text{obs}}$, the modified Restart EnKF reruns the forward model from time 0 to the current time t_k with the ensemble mean of updated parameter \mathbf{p}_k at time t_k (Song et al., 2014). The new mean of state variables is given by

$$\mathbf{E}(\mathbf{s}_k^{\text{re}}) \approx G_{0 \rightarrow k}(\mathbf{p}_k^a) \quad (11)$$

where $G_{0 \rightarrow k}$ represents the model simulation from time 0 to t_k ; the superscript *re* refers to the rerun; and $\mathbf{E}(\cdot)$ denotes the expectation.

Then, a new ensemble of the state variables is rebuilt based on $\mathbf{E}(\mathbf{s}_k^{\text{re}})$ in two steps:

- Step 1. Calculate and save the fluctuations of ensemble realizations around their means:

$$\Delta \mathbf{s}_{i,k}^a = \mathbf{s}_{i,k}^a - \frac{1}{N_1} \sum_{i=1}^{N_1} \mathbf{s}_{i,k}^a \quad (12)$$

- Step 2. Construct the new ensemble members by imposing the updated mean on the fluctuations of the forecast ensemble:

$$\mathbf{s}_{i,k}^{\text{a,re}} = \mathbf{E}(\mathbf{s}_k^{\text{re}}) + \Delta \mathbf{s}_{i,k}^a \quad (13)$$

Similar procedures are repeated until the final time. The readers may refer to Sun et al. (2014) for further method details.

2.3 | The construction of a data-driven structural error model with Gaussian process training

On the basis of studies of Kennedy and O’Hagan (2001), Xu and Valocchi (2016), and Xu, Valocchi, Ye, and Liang (2017), the model structural error, $e(\mathbf{x}, \mathbf{p})$, can be expressed as an additive term:

$$\mathbf{d}^{\text{obs}} = G(\mathbf{p}) + e(\mathbf{x}, \boldsymbol{\varphi}) + \delta \quad (14)$$

where the error model input \mathbf{x} may consist of the physical-based model output $G(\mathbf{p})$ and other relevant information in addition to time and location of the quantity of interest. This allows for training data that are not directly used to construct the physical model G . \mathbf{x} is a $1 \times d$ vector where d is the input dimension. The value of d is determined by the number of data types we attempt to fuse. Moreover, $\boldsymbol{\varphi}$ is a vector of the hyperparameter of the error model. Here, δ is the error term (more details are presented below).

A GP is fully specified by its mean function $\mu = \mathbf{E}[e(\mathbf{x})]$ and covariance function $k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{E}\{[e(\mathbf{x}_i) - \mu(\mathbf{x}_i)][e(\mathbf{x}_j) - \mu(\mathbf{x}_j)]\}$. In this study, a linear mean function $\mu(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$ is selected somewhat arbitrarily. In this function, $\boldsymbol{\beta}$ is the linear coefficient. An isotropic squared exponential covariance function is given by

$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma^2 \exp \left[-\frac{1}{\lambda^2} (\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j) \right] \quad (15)$$

where σ^2 and λ are two hyperparameters: σ^2 controls the marginal variance of $e(\mathbf{x})$; and λ is the characteristic length scale hyperparameter.

Inferences about all of the hyperparameters can be made in the light of training data (Rasmussen & Williams, 2006). Let $\{\mathbf{X}, \mathbf{d}^{\text{obs}} - \mathbf{G}\} = \{[\mathbf{x}_1, \mathbf{d}_1^{\text{obs}} - G(\mathbf{p})_1], \dots, [\mathbf{x}_n, \mathbf{d}_n^{\text{obs}} - G(\mathbf{p})_n]\}$ denote a set of n training data. \mathbf{X} and \mathbf{G} represent, respectively, observations and physically based model outputs at different locations and times, and \mathbf{X} is a $n \times d$ matrix. Therein, $G(\mathbf{p})_q$ ($q = 1, 2, \dots, n$) is the simulated soil moisture of the q th training data. Since by assumption the distribution of the data is Gaussian, the log marginal likelihood is given by

$$\begin{aligned} L &= \log [p(\mathbf{d}^{\text{obs}} - \mathbf{G} | \mathbf{X})] \\ &= -\frac{1}{2} (\mathbf{d}^{\text{obs}} - \mathbf{G} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{d}^{\text{obs}} - \mathbf{G} - \boldsymbol{\mu}) - \frac{1}{2} \log (|\boldsymbol{\Sigma}|) \\ &\quad - \frac{n}{2} \log (2\pi) \end{aligned} \quad (16)$$

where the covariance matrix $\boldsymbol{\Sigma}$ is calculated using Equation 15, and its ij th entry is $\Sigma_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

Based on the above training points, the posterior distribution of the model structural error, \mathbf{e}^* , can therefore be derived for m arbitrary future input $\mathbf{X}^* = \{\mathbf{x}_1^*, \mathbf{x}_2^*, \dots, \mathbf{x}_m^*\}$:

$$\mathbf{e}^* | \mathbf{y}, \varphi \sim \mathcal{N} \left[\overline{\mathbf{e}^*}, \mathbf{C}_{ee}(\mathbf{e}^*) \right] \quad (17)$$

$$\overline{\mathbf{e}^*} = \boldsymbol{\mu}^* + \boldsymbol{\Sigma}^{*T} \boldsymbol{\Sigma}^{-1} (\mathbf{d}^{\text{obs}} - \mathbf{G} - \boldsymbol{\mu}) \quad (18)$$

$$\mathbf{C}_{ee}(\mathbf{e}^*) = \boldsymbol{\Sigma}^{**} - \boldsymbol{\Sigma}^{*T} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}^* \quad (19)$$

where $\boldsymbol{\Sigma}_{i,j}^* = k(\mathbf{x}_i, \mathbf{x}_j^*)$ and $\boldsymbol{\Sigma}_{i,j}^{**} = k(\mathbf{x}_i^*, \mathbf{x}_j^*)$; $\boldsymbol{\mu}^*$ represents the mean $\mu(\mathbf{X}^*, \boldsymbol{\varphi})$; $\overline{\mathbf{e}^*}$ and $\mathbf{C}_{ee}(\mathbf{e}^*)$ represent the posterior mean and covariance of \mathbf{e}^* , respectively.

Finally, the predictions of state variables of interest, \mathbf{y}^* , can be calculated using

$$\mathbf{y}^* = \mathbf{G}^*(\mathbf{p}) + \mathbf{e}^*(\mathbf{x}^*, \boldsymbol{\varphi}) + \boldsymbol{\delta} \quad (20)$$

where $\mathbf{G}^*(\mathbf{p})$ is the physical-based model output corresponding to \mathbf{X}^* . It is worth emphasizing that the error term $\boldsymbol{\delta}$ includes \mathbf{C}_{ee} in addition to the measurement error ($\boldsymbol{\epsilon}$) (Zhang, Li, Zeng, & Wu, 2016; i.e., $\boldsymbol{\delta} = \mathbf{C}_{ee} + \boldsymbol{\epsilon}$). We can see from Equation 20 that the variance of posterior state vector becomes larger due to the additional introduction of uncertainty from the GP error model. However, a larger variance is not necessarily adverse in sequential data assimilation. On the one hand, the inclusion of uncertainty from the GP error model can improve the underestimation of model error due to the ignorance of model structural uncertainty. On the other hand, the additional error can alleviate the filter inbreeding problem, which is commonly accounted for by inflating the state covariance (Hendricks Franssen & Kinzelbach, 2008; Yu et al., 2019).

In our study, the GPML MATLAB toolbox version 4.1 documented in Rasmussen and Williams (2006) was used to carry out all GP training.

2.4 | The hybrid sequential data-worth analysis method with gaussian process

The new hybrid data-worth analysis framework can be divided into three stages as follows.

2.4.1 | Prior stage

At the prior stage (from time zero to time T_p) the integration of the modified Restart EnKF and GP is implemented sequentially (i.e., like EnKF-GP2 in Zhang et al., 2019, as shown in Figure 1). We take the simulation at time t_k ($0 < t_k$

$\leq T_p$) as an example to show the calculation procedure:

1. Assimilate the current observations $\mathbf{d}_k^{\text{obs}}$ and update the posterior state vector $\mathbf{y}_k^a = [\mathbf{p}_k^a, \mathbf{s}_k^a]$ via Equations 8–10.
2. Rerun the forward model in the modified restart EnKF with the mean of the updated physical-based parameters, $\mathbf{E}(\mathbf{p}_k^a)$, to obtain new soil moisture values from t_1 to t_k , i.e. $\mathbf{E}(\mathbf{s}_1^{\text{a, re}}), \mathbf{E}(\mathbf{s}_2^{\text{a, re}}), \dots, \mathbf{E}(\mathbf{s}_k^{\text{a, re}})$ via Equation 11. Here $\mathbf{E}(\mathbf{s}^{\text{a, re}})$ is equivalent to $\mathbf{G}(\mathbf{p})$ in Section 2.3.
3. Construct GP error model based on n training data: $\{\mathbf{X}, \mathbf{d}^{\text{obs}} - \mathbf{G}\} = \{[\mathbf{x}_1, \mathbf{d}_1^{\text{obs}} - \mathbf{G}(\mathbf{p})_1], \dots, [\mathbf{x}_n, \mathbf{d}_n^{\text{obs}} - \mathbf{G}(\mathbf{p})_n]\}$

Here, $n = k^* N_{\text{obs}}$, where N_{obs} is the number of observations at each time (via Equations 15–16).

1. Use the trained GP model to predict the soil moisture bias at each node (here, m is the number of nodes) and compensate for the model structural error in the original predictions, $\mathbf{E}(\mathbf{s}_k^{\text{re}})$ via Equations 17–20 to obtain a new mean of the state variables, $\mathbf{E}(\mathbf{s}_k^{\text{re, gp}})$.
2. Reproduce sample $\mathbf{s}_{i,k}^{\text{re, gp}}$ via Equation 13 based on $\mathbf{E}(\mathbf{s}_k^{\text{re, gp}})$ and the saved fluctuations of ensemble realizations in Equation 12.

Repeat these procedures until time T_p . After sequentially assimilating all prior data $\mathbf{A} = \{\mathbf{d}_1^{\text{obs}}, \mathbf{d}_2^{\text{obs}}, \dots, \mathbf{d}_{T_p}^{\text{obs}}\}$, the updated parameter ensembles, $\mathbf{p}_{T_p}^a$, are used to rerun the forward model from 0 to T_t (T_t is the total simulation time, and $T_t > T_p$). Once $\mathbf{s}_{i,j}^{\text{re}} (i = 1, 2, \dots, T_t; j = 1, 2, \dots, N_1)$ is yielded again, N_1 new GP models based on $n = T_p N_{\text{obs}}$ training data can be built to make predictions of \mathbf{e}^* after time T_p . In other words, a set of N_1 hypothetical observations are generated and denoted as $\mathbf{B}_{i,j} = \mathbf{e}_i^* + \mathbf{E}(\mathbf{s}_i^{\text{re}}) + \boldsymbol{\delta} (i = T_p + 1, T_p + 2, \dots, T_t; j = 1, 2, \dots, N_1)$ via Equation 20. N_1 should be large enough to include sufficiently diverse potential observations. In this study, N_1 is set as 200 to compromise between the computational cost and accuracy, as suggested by Man et al. (2016).

The difference between the traditional and hybrid data-worth analysis approaches is the procedure for generating potential observations during this stage. The traditional method does not consider model structural error but directly implements EnKF to assimilate all available prior data \mathbf{A} , and then utilizes the updated physical parameters to generate potential observations \mathbf{B} . The proposed hybrid method constructs the GP error model with the returned open-loop results based on the updated parameter mean via the modified restart EnKF. The learned model error by GP regression is used to compensate the current state vector to avoid or alleviate its adverse effect on the updated physical parameters. In this case, more stable parameters are utilized for predictions to obtain \mathbf{B} .

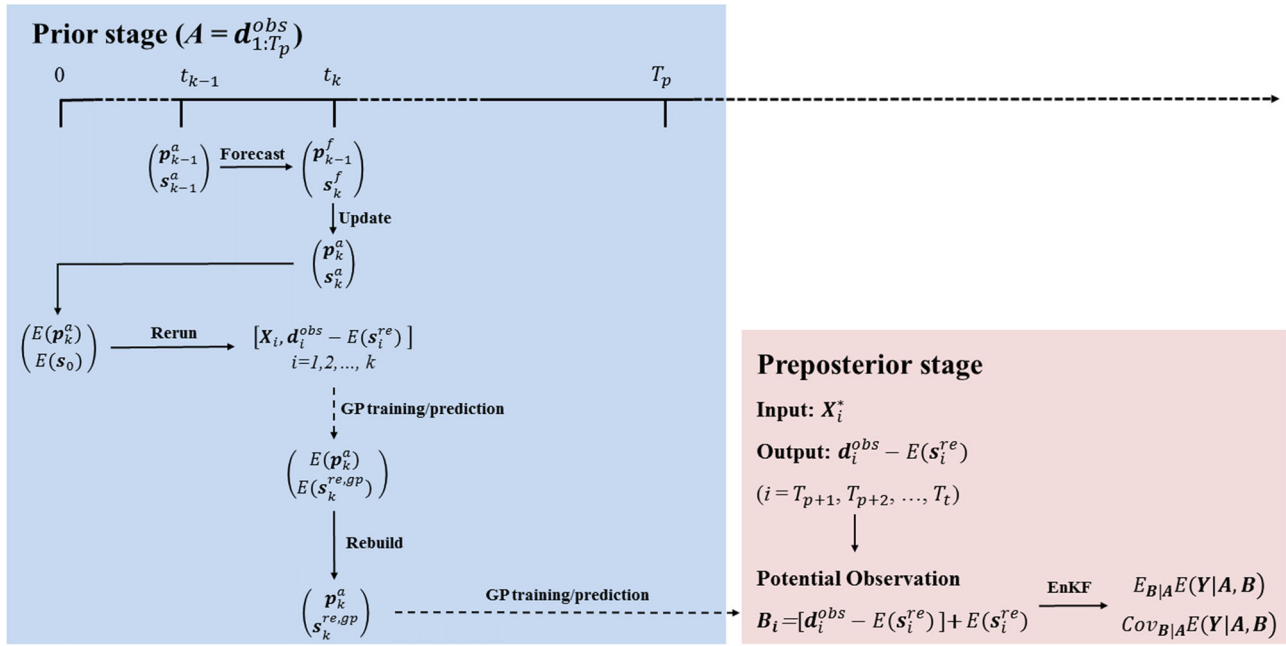


FIGURE 1 The workflows of the new hybrid data-worth analysis framework coupled with the modified restart ensemble Kalman filter and Gaussian process (GP). See Section 2 for definition of variables

2.4.2 | Preposterior stage

For each possible data \mathbf{B}_i ($i = 1, 2, \dots, N_1$) at any time t_k ($T_p < t_k \leq T_t$), firstly we generate N_2 realizations satisfying a Gaussian distribution with a mean of \mathbf{B}_i , whereas the variance is the measurement error. Next, the EnKF is implemented through a set of N_2 Monte Carlo realizations for each of the N_1 hypothetical observations. This allows us to calculate the updated ensemble mean $E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ and covariance $\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B})$, jointly conditioned on $\{\mathbf{A}, \mathbf{B}\}$. Quantities $E_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$, $E_{\mathbf{B}|\mathbf{A}}\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B})$, and $\text{Cov}_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ are then yielded by averaging over the collection of $N_1 \times N_2$ realizations.

2.4.3 | Posterior stage

At this stage, the real observations \mathbf{B}' corresponding to the potential observations \mathbf{B} have already been collected. Now, the real (i.e., reference or posterior) data worth can be evaluated in a sequential manner. The comparison of the expected and reference data worth reveals the effectiveness of the data-worth analysis framework.

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

$$E(\mathbf{Y}|\mathbf{A}) = E_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B}) \quad (21)$$

$$\text{Cov}(\mathbf{Y}|\mathbf{A}) = E_{\mathbf{B}|\mathbf{A}}\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B}) + \text{Cov}_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B}) \quad (22)$$

where $E(\mathbf{Y}|\mathbf{A})$ and $E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ are the expectations of state vector \mathbf{Y} conditioning on prior dataset $\{\mathbf{A}\}$ and augmented dataset $\{\mathbf{A}, \mathbf{B}\}$, respectively; $\text{Cov}(\mathbf{Y}|\mathbf{A})$ represents prior predictive uncertainty only based on $\{\mathbf{A}\}$; $E_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ and $E_{\mathbf{B}|\mathbf{A}}\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B})$, respectively, are the expectation of $E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ and $\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ over all \mathbf{B} vectors generated, conditional on $\{\mathbf{A}\}$. Therein, $E_{\mathbf{B}|\mathbf{A}}\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ represents the predictive uncertainty in the preposterior data-worth analysis. $\text{Cov}_{\mathbf{B}|\mathbf{A}}E(\mathbf{Y}|\mathbf{A}, \mathbf{B})$ is the reduction of uncertainty due to the addition of future possible data \mathbf{B} . For ease of presentation, these quantities can be denoted as

$$\begin{aligned} E(\mathbf{Y}|\mathbf{A}) &= \mathbf{E}_1 \\ \text{Cov}(\mathbf{Y}|\mathbf{A}) &= \mathbf{C}_1 \\ E(\mathbf{Y}|\mathbf{A}, \mathbf{B}) &= \mathbf{E}_2 \\ E_{\mathbf{B}|\mathbf{A}}\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B}) &= \mathbf{C}_2 \end{aligned} \quad (23)$$

Following Xu (2007), Zhang et al. (2016), and Man et al. (2016), a mean-covariance-type scalar measure, the relative entropy (RE), was introduced to quantify data worth in our study. As a signal-dispersion combined index, RE provides a measure of the information content of an analysis (posterior) probability density function (pdf) with respect to a background (prior) pdf. Considering that the background and the analysis pdfs are n -dimensional Gaussian functions, the relative entropy can be expressed as

$$\text{RE} = J_b + \text{DP} \quad (24)$$

$$J_b = (\mathbf{E}_2 - \mathbf{E}_1)^T \mathbf{C}_1^{-1} (\mathbf{E}_2 - \mathbf{E}_1) / 2 \quad (25)$$

$$DP = [\ln \det (\mathbf{C}_1 \mathbf{C}_2^{-1}) + T_r (\mathbf{C}_2 \mathbf{C}_1^{-1}) - n] / 2 \quad (26)$$

where J_b is the signal (mean) part of the RE, and DP is the dispersion (covariance) part of the RE.

In addition, when calculating the reference data worth, \mathbf{C}_2 and \mathbf{E}_2 refer to $\text{Cov}(\mathbf{Y}|\mathbf{A}, \mathbf{B}')$ and $\mathbf{E}(\mathbf{Y}|\mathbf{A}, \mathbf{B}')$, respectively.

2.5 | Description of experimental data and model setup

2.5.1 | Data source and site description

In situ soil moisture observations, precipitation data, and soil temperature data were obtained from the ISMN. The Falkenberg station (52.1669° N, 14.1241° E; as depicted in Figure 2) of the MOL-RAO network was selected as the study site. At this grassland site, TRIME-EZ TDR sensors (IMKO) were installed to measure soil moisture at six different depths (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m), as depicted in Figure 2a. Here, we used 700 d of observations from 1 Jan. 2004 to 30 Nov. 2005 as the study period. The daily precipitation and potential evapotranspiration data are presented in Figure 2b. The soil texture and saturated soil water content of the topsoil (0–30 cm) and the subsoil (30–100 cm) were also downloaded from the ISMN. In topsoil, the proportions of clay, sand, and silt are 6, 73, and 21%, respectively, while being 12, 61, and 27% in subsoil. The saturated soil water contents in topsoil and subsoil are 0.43 and 0.40 $\text{cm}^3 \text{cm}^{-3}$, respectively.

Meteorological data including air temperature at 2-m height, specific humidity at 2-m height, and all sky insolation incident on a horizontal surface downloaded from the NASA Prediction of Worldwide Energy Resources (NASA POWER) are shown in Figures 2d, 2e, and 2f, respectively.

2.5.2 | Model simulations and evaluation setup

The main parameters of the study are listed in Table 1. The soil column has a height of 1 m and is discretized with 53 grids (the grid size is mainly 2 cm, with a local refinement of 1 cm at observation depths of 0.15 and 0.45 m). The top and bottom boundaries are set as an atmospheric boundary and a constant water content boundary (being equal to 0.35), respectively.

The van Genuchten–Mualem model is used to describe the constitutive relationships of soil hydraulic parameters in Richards' equation. Therein, α , n , and K_s are assumed to be unknown. Based on the soil texture classification, the prior values of soil parameters are estimated using the Neural Network Prediction module of HYDRUS-1D (PC-Progress, <https://www.pc-progress.com/en/Default.aspx?hydrus-1d>),

TABLE 1 Summary of key model parameters

Parameter	Value
Description	
Soil column height, m	1.0
No. of nodes	53
No. of realizations	
N_1	200
N_2	200
Initial distribution of soil hydraulic parameters	
$\ln(\alpha)$, $\ln(\text{m}^{-1})$	$N(1.228, 0.485)$
$\ln(n)$	$N(0.404, 0.067)$
$\ln(K_s)$, $\ln(\text{m d}^{-1})$	$N(-1.309, 1.441)$
Prior hyperparameters	
λ	1
σ^2	0.5
β	0

whereas variances of α , n , and K_s are specified empirically. Considering that soil hydraulic parameters generally satisfy lognormal distributions, the given mean and variance were converted to logarithmic form, as shown in Table 1. Subsequently, N_1 samples are generated as the initial parameter ensemble. In addition, θ_r is assumed to be invariable ($\theta_r = 0.04$), which is also estimated with HYDRUS-1D. The soil moisture measurement error is set to be 0.03 $\text{cm}^3 \text{cm}^{-3}$ (Wang et al., 2018).

We deliberately introduced two categories of model structural error source: (a) we simplified the double-layered soil column to a homogeneous one, having properties that were consistent with those of the original topsoil; and (b) we neglected the surface cover of grass and assumed it was bare soil, when calculating evapotranspiration.

The GP error model is constructed to describe the model structural error in soil moisture modeling. The input \mathbf{x} of the GP includes observation depth, precipitation, evapotranspiration, and simulated soil moisture. Other candidate information, such as meteorological data and soil temperature, will also be considered for specific research purposes in different test cases. The output of the GP error model is the bias between the simulated and observed soil moisture. The prior values of GP hyperparameters are given in Table 1.

Here, to quantify the ability of capturing actual observations with potential observations, the RMSE of the two is calculated:

$$\text{RMSE} = \sqrt{\frac{1}{N_t} \sum_{k=1}^{N_t} (\bar{\theta}_k^f - \theta_k^{\text{Obs}})^2} \quad (27)$$

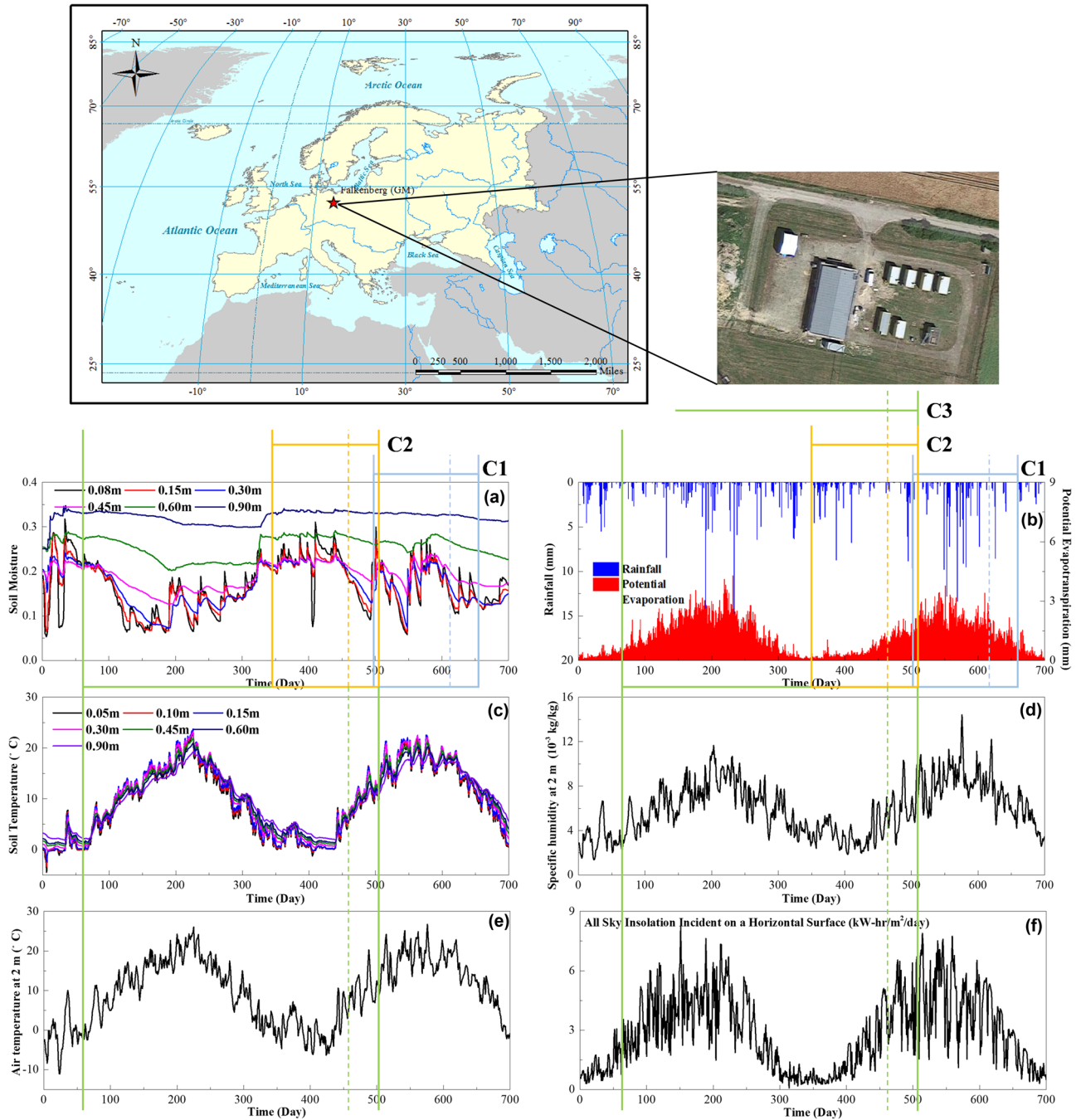


FIGURE 2 The spatial location of Falkenberg Station and the temporal change of (a) soil moisture at various depths (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m), (b) daily rainfall and potential evaporation, (c) soil temperature at various depths (0.05, 0.10, 0.15, 0.30, 0.45, 0.60, and 0.90 m), (d) specific humidity at 2-m height, (e) air temperature at 2-m height, and (f) all sky insolation incident on a horizontal surface from 1 Jan., 2004 to 30 Nov. 2005 (700 d in total) at this station. GM, Grenzsichtmessfeld

where $\bar{\theta}_i^f$ represents the mean of forecasted potential soil moisture samples at time t_k ; θ_k^{Obs} is the corresponding observed value; and N_t is the total duration of the simulation period at the preposterior or posterior stage (i.e., $N_t = T_t - T_p$).

Within our sequential data-worth analysis framework, the value of the reference data worth depends on \mathbf{C}_1 and \mathbf{E}_1 as

well as \mathbf{C}_2 and \mathbf{E}_2 . At the prior stage, the traditional method is to directly use the EnKF to sequentially assimilate the prior data to obtain \mathbf{C}_1 and \mathbf{E}_1 , whereas in the new data-worth analysis framework, \mathbf{C}_1 and \mathbf{E}_1 are calculated by hybridizing the modified restart EnKF and the GP. It is clear the produced \mathbf{C}_1 and \mathbf{E}_1 values from these two approaches are different. Even though the same observations are assimilated at the

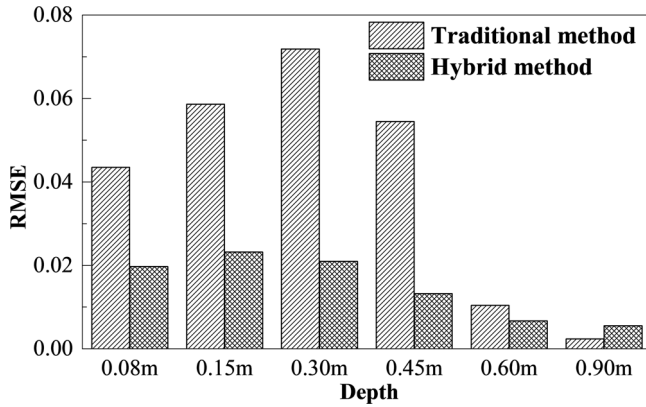


FIGURE 3 A comparison of RMSE between actual soil moisture observations at various depths (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m) and corresponding predictions with the traditional and new hybrid methods, respectively

posterior stage in both methods, they yield different values of reference data worth. To compare the relative differences of data-worth estimation accuracy using these methods, the mean absolute percentage error (MAPE) between expected and reference RE is defined:

$$\text{MAPE} = \frac{1}{N_t} \sum_{k=1}^{N_t} \left| \frac{\text{RE}_k^{\text{expect}} - \text{RE}_k^{\text{refer}}}{\text{RE}_k^{\text{refer}}} \right| \quad (28)$$

where $\text{RE}_i^{\text{expect}}$ and $\text{RE}_i^{\text{refer}}$ represent the expected and reference relative entropy at time t_k , respectively.

3 | RESULTS AND DISCUSSIONS

3.1 | Validation of the hybrid data-worth analysis framework

In the first case (denoted C1), data between 25 Apr. and 1 Oct. 2005 (highlighted by blue rectangle in Figure 2) were used to verify the validity of the hybrid method. The data of first 120 d (the left of the blue dotted line in Figure 2) serve as the prior data (also the training data for GP), whereas the remaining data of the following 40 d (the right of the blue dotted line) are used to test the accuracy of soil moisture predictions after GP training.

The potential soil moisture observations during the next 40 d can be predicted by including the GP correction. For ease of comparison, RMSE values between the observed and predicted soil moisture at 0.08 m during the whole preposterior stage (i.e. from the 120th to the 160th day) are presented in Figure 3. The significantly reduced RMSE values for the hybrid method indicate the greater efficiency of the hybrid method in reproducing the real observations.

A comparison of the matching degree of expected and reference data worth based on traditional and hybrid methods is shown in Figure 4. Only data worth of the soil moisture at 0.08 m is presented, and the data worth is quantified using RE. Because the values of expected and reference RE differ in some cases by a few orders of magnitude, the logarithmic RE $[\ln(\text{RE})]$ is shown. It is seen that the expected and reference data worth from the hybrid approach have a much better match than those of the conventional method, for both parameter identification and soil moisture profile estimation. Thus, the integration of the GP within our sequential data-worth analysis framework led to marked improvement of data-worth estimation accuracy.

3.2 | Scenario diversity of training (or prior) data

As defined in Zook et al. (2012), “scenario diversity is a measurement of variations among the content of scenarios.” The given scenario of the training data may reflect the temporal variability of daily precipitation, evapotranspiration, and other conditions. In this study, if a case experiences precipitation and evapotranspiration events of various intensities, we call it “diverse.” A rich scenario diversity may be critical for GP training since machine learning method (GP regression here) is apt to obtain high prediction ability when it is trained by diverse input data. The impact of scenario diversity at the prior stage on the estimation of the distribution of potential observations and corresponding expected data worth is explored in this section.

The success of data-worth estimation using the proposed hybrid approach in C1 should probably be attributed to the diverse scenario of the training data. We compared the performance of the hybrid method, when the prior stage and the preposterior stage had totally different scenarios (herein, rainfall). Case 2 (C2) was designed to test this effect: at the prior stage (training data), the first 120 d between 26 Nov. 2004 and 25 Mar. 2005 (the left side of yellow dotted line in the yellow rectangle in Figure 2) had frequent rainfall, whereas during the last 40 d between 25 Mar. 2005 and 4 May 2005 (the right side of yellow dotted line in the yellow rectangle), there was little rainfall and strong evapotranspiration.

Figures 5a and 5b present a comparison of actual soil moisture measurements and corresponding predictions based on the traditional method, as well as our hybrid method for case C2. Only the results for depths of 0.08 m and 0.30 m are presented. It is seen that the potential observations deviate from actual observations, even when using the hybrid method. The continuous drop in soil moisture content during the preposterior stage is not experienced during the training period, and the GP seems unable to make a proper correction based on the

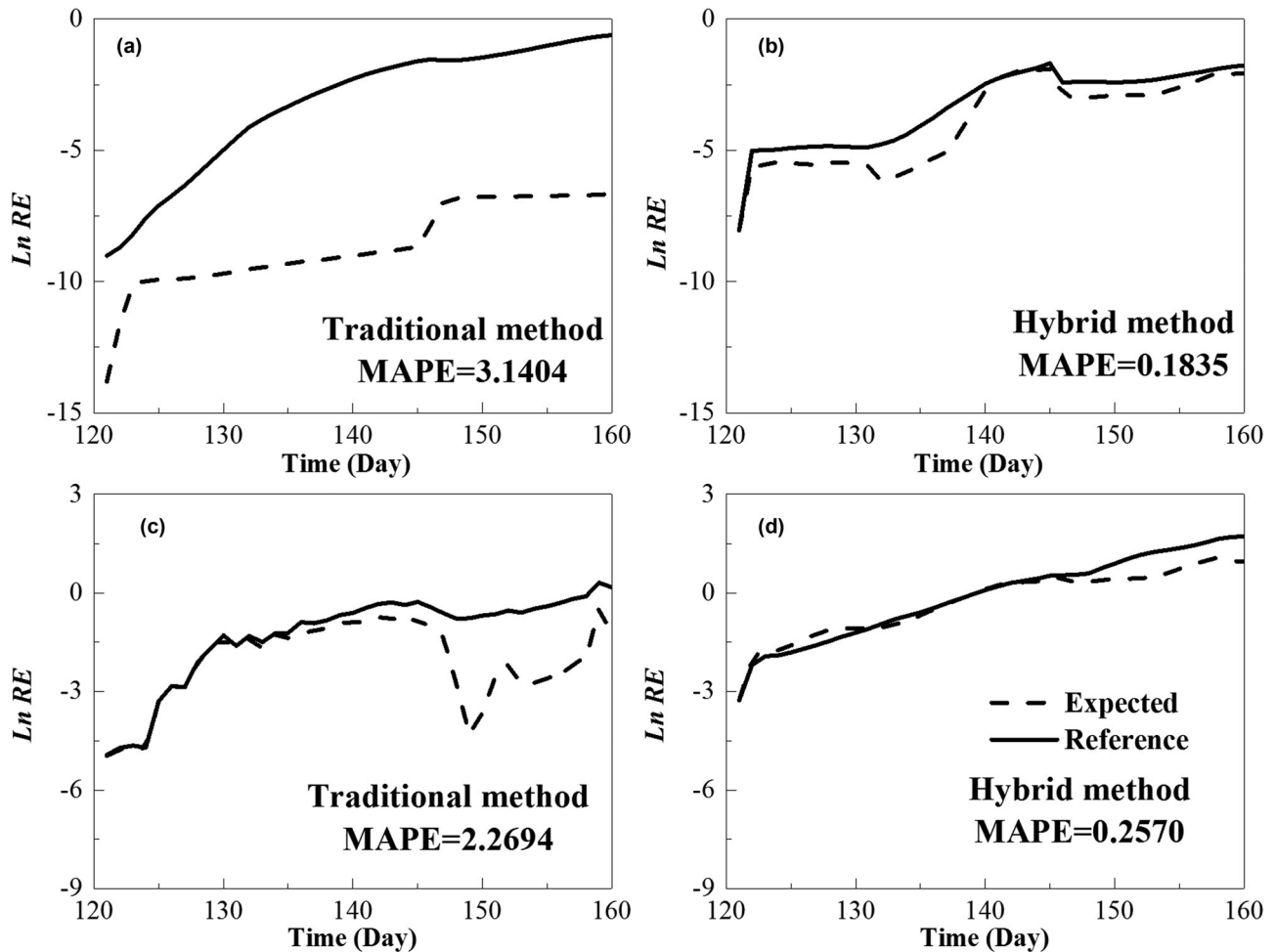


FIGURE 4 A comparison between the expected and reference data worth [the logarithmic relative entropy, $\ln(\text{RE})$] of soil moisture at 0.08 m regarding (a, b) parameter identification and (c, d) soil moisture profile predictions based on the traditional method and our hybrid approach, respectively. MAPE, mean absolute percentage error

historical scenarios. Subsequently, the accuracy of data-worth evaluation is not satisfactory, as shown in Figures 6a and 6b).

We extended the training period from 120 to 400 d between 20 Feb. 2004 and 25 Mar. 2005 (the left of the green dotted line in the green rectangle in Figure 2). This new case is labeled as C3. As shown in Figures 5c and 5d, the soil moisture predictions after GP training are in better agreement with the actual measurements, especially at depths of 0.08 m. The expected data worth also approached the reference counterpart (Figures 6c and d). The results demonstrate that inclusion of more prior data in the hybrid method can better reduce the negative influence from the model structural error. However, there is a potential failure of the hybrid method, if contrasting scenarios occur between the prior stage and the posterior or preposterior stages. From the perspective of model robustness, it is suggested to include datasets covering diverse scenarios during the training stage. Furthermore, the comparison between Figures 4b and 4d and

Figures 6c and 6d reveals that the use of data from 120 d prior (in C1) performed comparably in estimating data worth with using data from 400 d prior (in C3). This indicates that having scenario diversity in the prior training data is of more importance than having a large data volume.

3.3 | Augmentation of soil temperature data into the gaussian process training

In C3, although the prior (or training) data has been augmented from 120 to 400 d, there is still a large deviation between the predicted and observed soil moisture in deeper soils (e.g., at depths of 0.30 m), as shown in Figure 5d. In this section, we further improve the accuracy of data-worth estimation at these depths by using multiple data sources. One major advantage of the GP approach is that the inputs of the GP can take into account a variety of relevant data

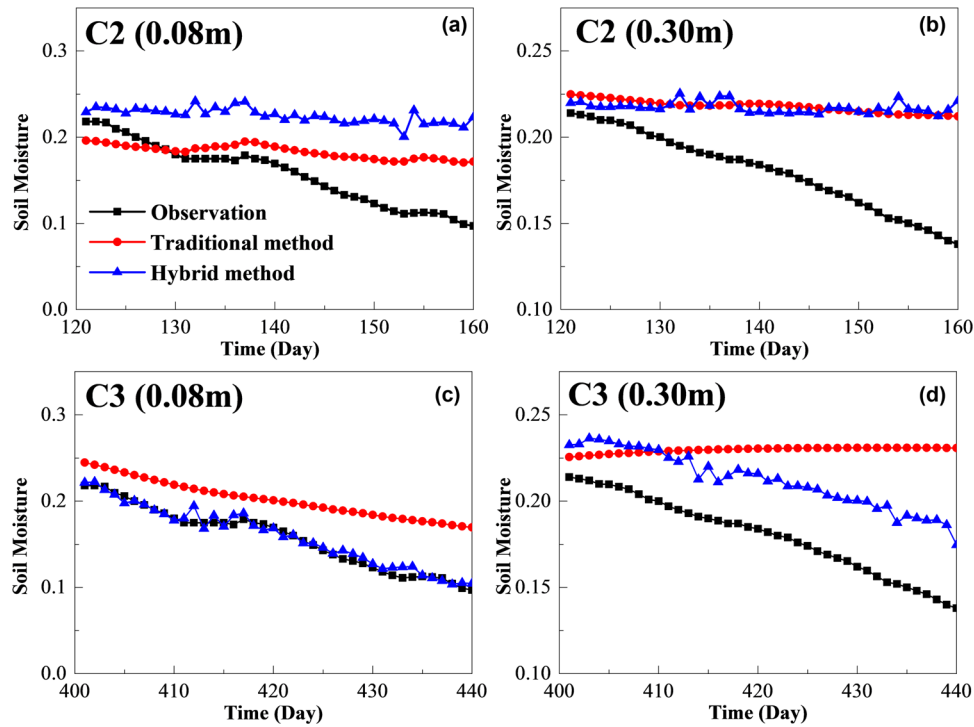


FIGURE 5 A comparison of actual soil moisture observations and predictions at the depths of 0.08 and 0.30 m with the traditional method, as well as with the hybrid method, from 26 Mar. to 4 May 2005 when (a, b) the prior data (or training data) is of 120-d length, and (c, d) the prior data is of 400-d length

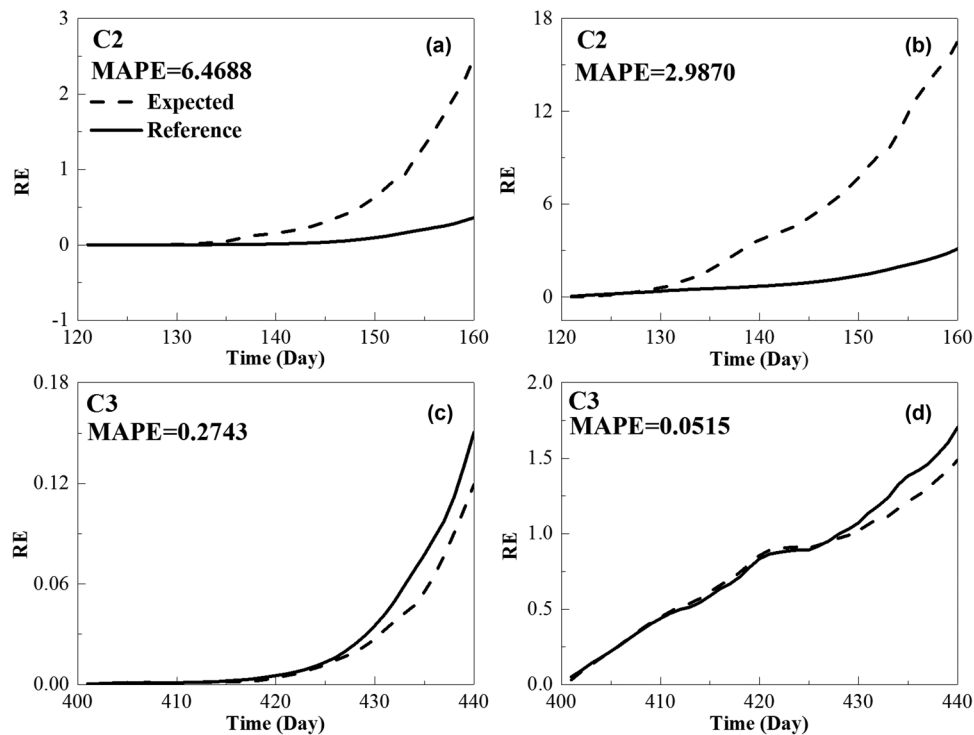


FIGURE 6 A comparison between the expected and reference data worth (relative entropy [RE]) of soil moisture at 0.08 m regarding (a, c) parameter identification and (b, d) soil moisture profile estimation based on our new hybrid method when (a, b) the prior data (or training data) of case C2 is of 120-d length, and (c, d) the prior data of C3 is of 400-d length. MAPE, mean absolute percentage error

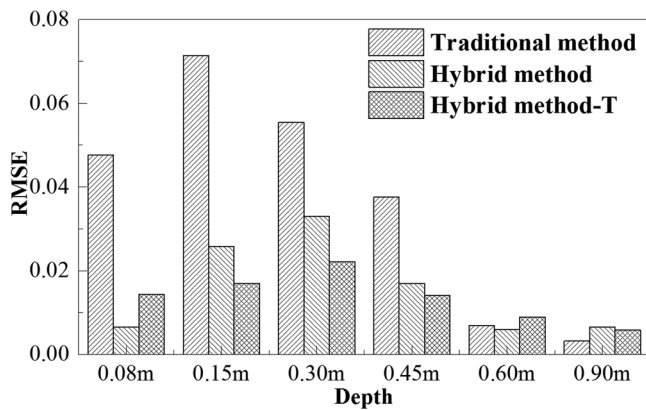


FIGURE 7 A comparison of RMSE between actual soil moisture observations and corresponding predictions with the traditional method, hybrid method, and the hybrid method-T where the Gaussian process error model input \mathbf{x} is augmented with soil temperature

that are not directly used to construct the physical model (Xu et al., 2017). Here, we incorporate soil temperature data that were not directly related to the soil moisture equation into the training input \mathbf{x} of Equation 14, and we investigate the benefit of training with additional soil temperature data on the prediction of potential observations and data-worth analysis. We then designed one more case, in which soil temperature (as shown in Figure 2c) was augmented into the GP input, whereas other settings were identical to case C3. Thus, the GP input included monitoring depth, daily precipitation, evapotranspiration, soil temperature, and simulated soil moisture.

The RMSE values between observed and predicted soil moisture from the traditional method, the hybrid method only trained with soil moisture data (labeled as the hybrid method), and the hybrid method trained with soil moisture and temperature data (labeled as the hybrid method-T) are compared in Figure 7. The RMSE values dropped significantly at most observation locations, especially for the soil moisture at depths of 0.30 and 0.45 m. Recalling results of Figure 5d, it seems that these additional temperature data are helpful to improve the prediction of potential observations. Figure 8 compared the data worth of soil moisture observations (at a depth of 0.30 m) in C3 and the new case that included soil temperature data. Here, only the data worth regarding parameter identification are presented. It was seen that after joining the temperature into the GP training, the data-worth estimation accuracy improved, yielding a reduction of the MAPE from 0.8312 to 0.7267.

The improvements in both potential observation reproduction and data-worth estimation demonstrate the feasibility of using the GP to mine the information hidden in soil temperature. The effectiveness of jointly training soil moisture and temperature data is due to the physical connection between soil moisture and soil temperature. As depicted in Figure 9a,

there is a strong correlation between soil temperature and soil moisture at each depth. The cross-correlation matrix between soil moisture and soil temperature in space (at six different observation depths) is also presented (Figure 9b). It is interesting to see that soil moisture and soil temperature have a strong correlation across the whole soil profile. The complementary role of soil temperature data in soil moisture dynamics and soil properties estimation was investigated by Dong, Steele-Dunne, Judge, and van de Giesen (2015) and Dong, Steele-Dunne, Ochsner, and van de Giesen (2015, 2016). The intrinsic reason for such strong correlation is that soil thermal properties are a function of soil moisture, and the soil heat transfer process can be monitored to estimate soil moisture (Dong, Steele-Dunne, Ochsner, & van de Giesen, 2016). To be specific, soil-specific heat and volumetric heat capacity increase with increasing moisture content due to the gradual substitution of the gas phase in the soil pores by the liquid phase (Abu-Hamdeh, 2003). Our results demonstrate that the hidden value of soil temperature for soil moisture and parameter estimation can be excavated by GP regression without resorting to the coupled soil water and heat transport equation.

The correlation between soil moisture and soil temperature varies with depth (Figure 9b), with the lowest correlations in very deep soil (at a depth of 0.9 m) and relatively low correlation in surface soil (0.08 m) and deep soil (0.6 m). The small correlation coefficient in surface soils may be because both soil moisture and temperature are susceptible to changes in external conditions. As the soil depth increases, the soil temperature is more influenced by the water temperature through direct heat transfer, whereas the soil moisture is less variable due to increasing difficulty of water vapor diffusion. It is noteworthy that the degree of correlation is in line with the degree of improvement in predicting potential observations by introducing soil temperature as additional training data. Thus, higher correlations at depths of 0.15, 0.30, and 0.45 m (Figure 9b) correspond to larger drops in RMSE values (Figure 7). The above analysis also explains the improvement in Figure 8b, suggesting that a universal value of soil temperature data can be expected. One important implication is that the data worth of one particular measurement program is not only determined by direct data described by the primary variables in the governing equation, but it also may be influenced by auxiliary data for variables that are not given in the equation.

3.4 | Replacement of evapotranspiration calculation with the Gaussian process regression

Considering the flexibility of the GP in merging multisource data, we attempted to directly replace the Penman–Monteith

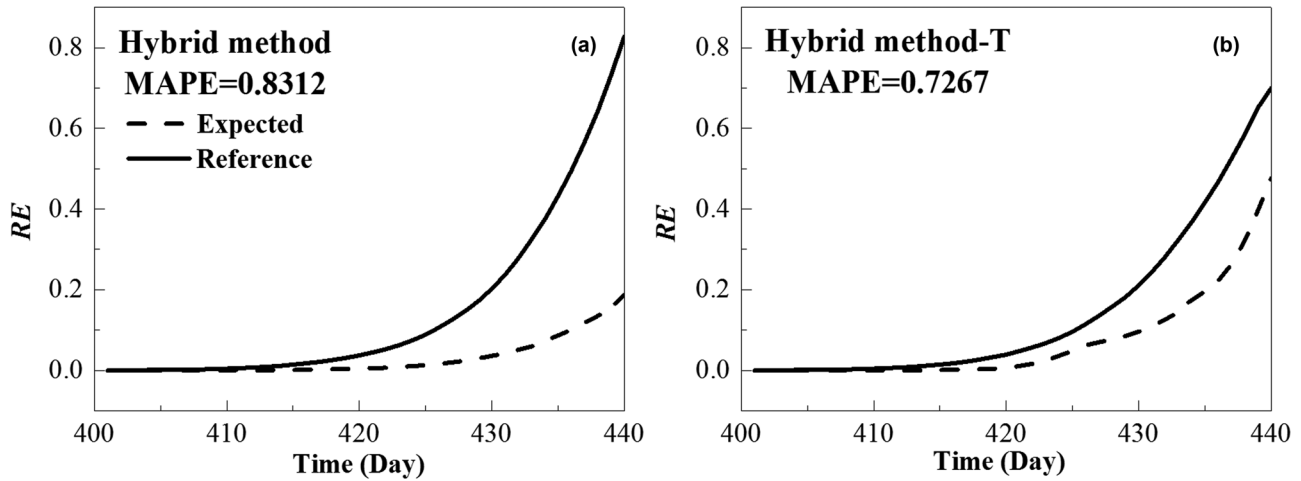


FIGURE 8 A comparison of expected (reference) data worth (relative entropy [RE]) of soil moisture at 0.30 m regarding parameter identification with (a) the hybrid method and (b) the hybrid method-T where Gaussian process error model input x is augmented with soil temperature. MAPE, mean absolute percentage error

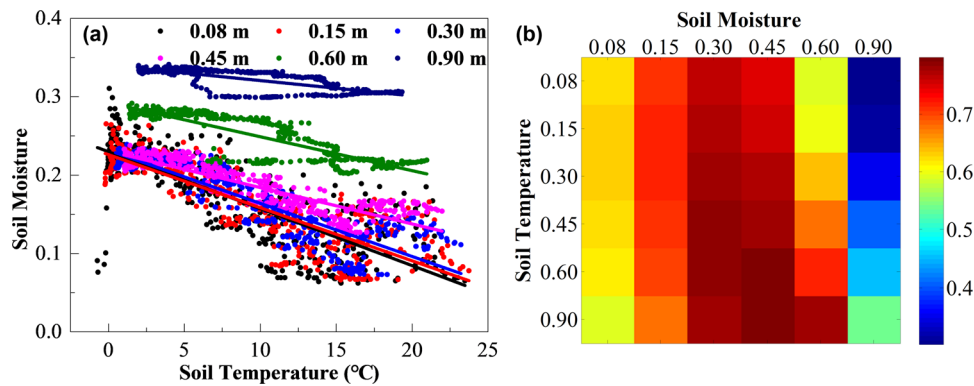


FIGURE 9 (a) Scatters between soil temperature and soil moisture at each depth (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m); (b) the cross-correlation matrix between soil moisture and soil temperature at each depth (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m)

equation with the GP, by introducing several easily available observations obtained from a regular weather station. Three new cases were designed, in which air temperature at 2-m height, specific humidity at 2-m height, and all sky insolation incident on a horizontal surface were respectively used to replace the evapotranspiration term in the GP input x . The approximate mapping from these meteorological data together with other basic information (e.g., monitoring depth, daily precipitation, etc.) was constructed by GP regression. For comparison, one additional case simultaneously considering all three meteorological observations was also designed. Other model settings in these four cases were identical to those of case C3.

Figures 10a–10d depict a comparison of expected and reference data worth with the above four alternative GP inputs. Only the data worth of soil moisture at 0.30-m depth regarding parameter identification is shown. Overall, the MAPE

values between expected and reference data worth in all four cases were smaller than that of of case C3 (having a MAPE value of 0.8312, as shown in Figure 8a). Training the GP with air temperature data brought about the most substantial improvement to the data-worth assessment, followed by training with radiation data, and then specific humidity data. By inspecting the correlation between the potential evapotranspiration (ET_0) and these three types of observations (with correlation coefficients of .53, .77, and .58, respectively), we found that a higher correlation corresponded to a larger MAPE reduction. These results surprisingly demonstrate that training a GP system with input data for the Penman–Monteith equation led to even better results than using a direct calculation of the Penman–Monteith equation. This may be because the grassland was oversimplified to bare soil, resulting in a large difference between the reality and modeling settings. To examine the modeling performance under

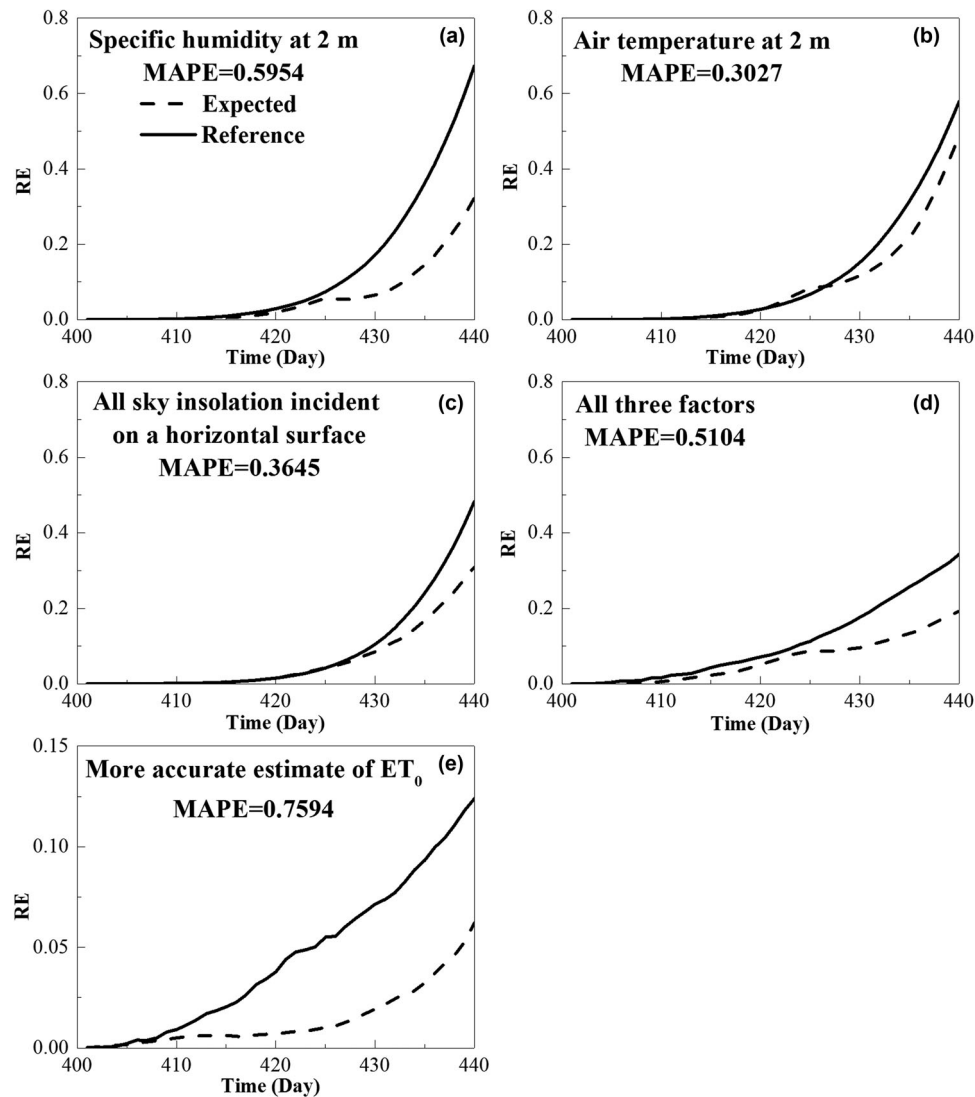


FIGURE 10 A comparison of expected (reference) data worth (relative entropy [RE]) of soil moisture at 0.30 m regarding parameter identification when the evapotranspiration term of Gaussian process input in C3 is replaced with (a) specific humidity at 2-m height, (b) air temperature at 2-m height, (c) all sky insolation incident on a horizontal surface, (d) simultaneously all three observations, and (e) a more accurate estimate of potential evapotranspiration (ET_0) with the Penman–Monteith equation by roughly considering vegetation information. MAPE, mean absolute percentage error

more plausible settings, another case was designed using the true vegetation cover (grassland), whose root depth is assumed to be 30 cm and root density is vertically distributed evenly. Other settings were identical to those in C3. A comparison of expected and reference data worth of soil moisture at 0.30 m regarding parameter identification is shown in Figure 10e. In this case, even when more plausible settings for calculating evapotranspiration are provided, training the GP model with a direct calculation of the Penman–Monteith equation (MAPE = 0.7594) results in a poorer accuracy than that obtained directly with input data of this equation (Figure 10d). These results also imply that the evapotranspiration calculated with the Penman–Monteith equation often

includes considerable uncertainties, which may arise from the equation form, as well as its associated model parameters. Meanwhile, limited by inadequate knowledge of the reality, we assume that soil evaporation only occurs within a depth of 0.4 m in this study. This excessive simplification is not in line with the fact that the depth range and vertical distribution of soil evaporation vary over time. Instead of directly calculating the Penman–Monteith method, we found that soil moisture at certain depths actually had a rather strong correlation with meteorological factors. For example, Figure 11 shows that soil moisture at the depths of 0.08, 0.15, 0.30, 0.45 m has a rather strong correlation with air temperature at 2-m height.

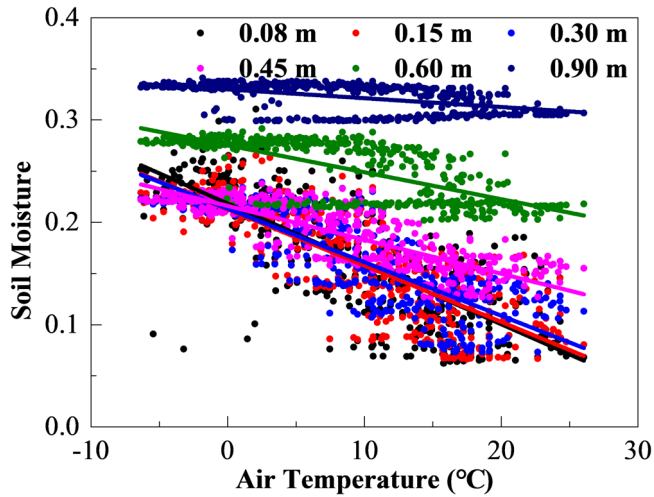


FIGURE 11 Scatters between air temperature at 2-m height and soil moisture at various depths (0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m)

In comparison with case C3 (Figure 8b), training the GP with all three types of data (Figure 10d) still produces better accuracy of data-worth estimation. The improvement was larger than that obtained by training with specific humidity data alone. However, compared with cases using air temperature or radiation data, a degraded data-worth estimation accuracy was obtained by jointly using all three types of data.

This result seems contradictory to our common sense that the inclusion of multisource data can provide more information to improve our understanding of unknown systems (Corcoran, Knight, & Gallant, 2013; Shi et al., 2015). Figure 12a–12d shows the ensemble (black line) and mean (red line) of the hyperparameter (β) of the linear mean function after training for each input entry, for the above four GP input augmentation schemes. A comparison of Figures 12a–12c and Figure 12d leads to two findings: (a) integrating multiple meteorological data simultaneously results in a significant decrease in the absolute value of β with respect to each input entry, with the variation range being equal to 0.00–0.25 (Figure 12d); and (b) depth information that is expected to have a high weight only has a very small weight, with a mean of 0.02, after all three types of meteorological data are merged. Similarly, rainfall is supposed to be an important factor for soil moisture estimation, whereas the weight of rainfall is close to zero in Figure 12d. These changes in hyperparameter values indicate that the excessive augmentation of input entry in the GP regression increases the difficulty of achieving global optimality. Under such circumstances, the performance of the trained GP system deteriorates. Since GP regression is a purely data-driven approach, it is still necessary to investigate criteria for choosing data for GP construction in the future.

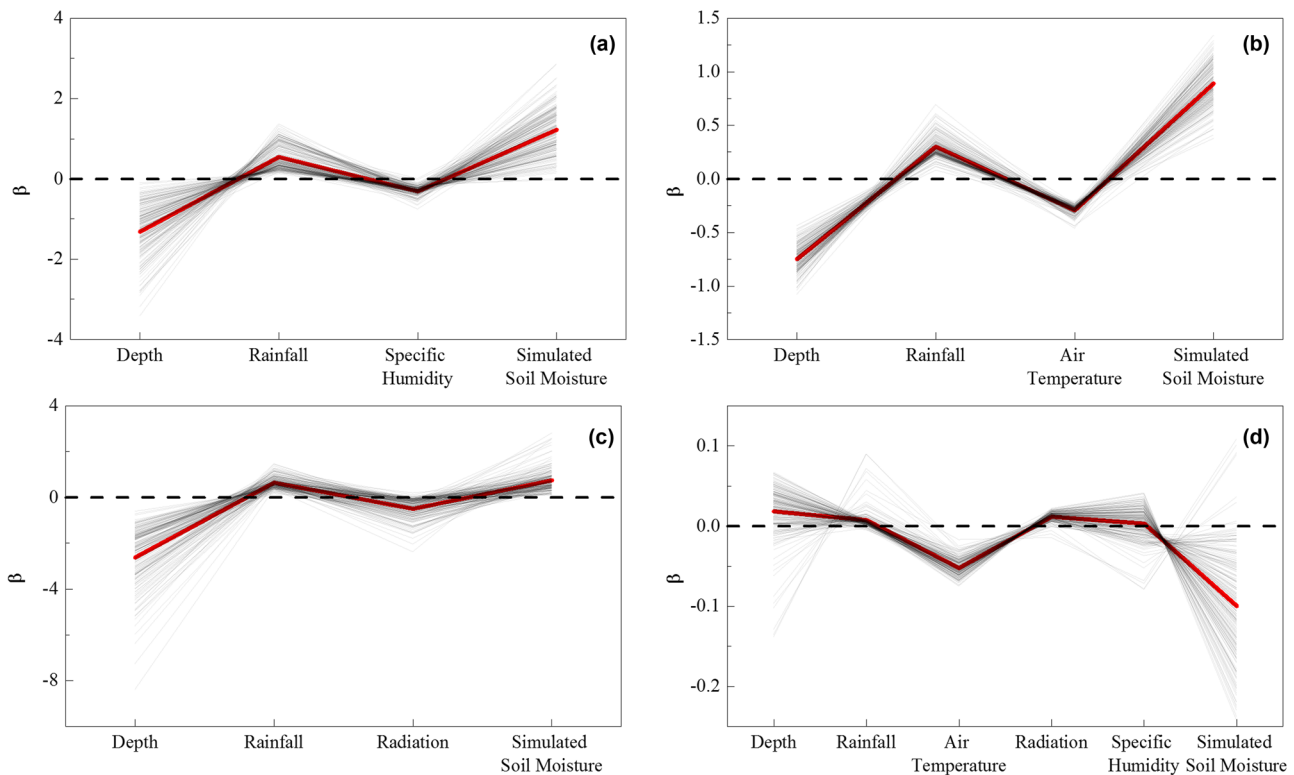


FIGURE 12 The ensemble (black line) and mean (red line) of the hyperparameter (β) of the linear mean function after training for each input entry, when the Gaussian process input is augmented respectively with (a) specific humidity, (b) air temperature, (c) all sky insolation incident on a horizontal surface (radiation), and (d) all three types of meteorological data, on the basis of monitoring depth, rainfall, and simulated soil moisture

4 | CONCLUSIONS

In this study, a sequential data-worth analysis framework based on a hybrid data assimilation approach (EnKF-GP) was investigated. Compared with the conventional data-worth analysis approach, this hybrid method had two major advantages: (a) by introducing a GP training, the hybrid method was able to learn the model structural error from the prior data; and (b) the hybrid method was able to utilize multiple sources of data without involving all of them in the governing equation. Our research led to the following major conclusions:

1. Data worth relies on the amount of information that can be provided to support the modeling for a given purpose. Besides data characteristics (data collection location, frequency, and accuracy), model quality also affects data-worth analysis. The hybrid approach loosens the requirement of constructing perfectly sound conceptual models for hydrologic realities during data-worth assessment.
2. The comparison of cases C1–C3 demonstrated that the scenario diversity at the prior stage plays a crucial role in establishing effective GP training. It is suggested to include either sufficiently long or enough diverse prior data to avoid potential failure of the hybrid approach.
3. The data worth of one given measurement program is influenced by direct data and indirect data. The integration of soil temperature into the GP training input in soil water problems can unravel hidden information and consequently improve data-worth estimation. It also seems feasible to replace the Penman–Monteith equation with a GP system trained by critical weather data. However, integration of more types of data does not necessarily further improve the accuracy of data-worth estimation.

Further work may extend this study to two- or three-dimensional variably saturated flow. Besides, only a few types of indirect data are considered in this paper. Our future interest is to integrate multiple different types of indirect data at various spatial scales.

ACKNOWLEDGMENTS

This study was supported by the National Natural Science Foundation of China Grants 51779180 and 51861125202. We gratefully acknowledge the help of Trudi Semeniuk, who polished the English language of the text.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Liangsheng Shi  <https://orcid.org/0000-0003-0446-0488>

REFERENCES

- Abebe, A. J., & Price, R. K. (2003). Managing uncertainty in hydrological models using complementary models. *Hydrological Sciences Journal*, 48, 679–692. <http://doi.org/10.1623/hysj.48.5.679.51450>
- Abu-Hamdeh, N. H. (2003). Thermal properties of soils as affected by density and water content. *Biosystems Engineering*, 86, 97–102. [https://doi.org/10.1016/S1537-5110\(03\)00112-0](https://doi.org/10.1016/S1537-5110(03)00112-0)
- Anderson, J. L., & Anderson, S. L. (1999). A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127, 2741–2758. [http://doi.org/10.1175/1520-0493\(1999\)127%3c2741:AMCIOT%3e2.0.CO;2](http://doi.org/10.1175/1520-0493(1999)127%3c2741:AMCIOT%3e2.0.CO;2)
- Ben-Zvi, M., Berkowitz, B., & Kesler, S. (1988). Pre-posterior analysis as a tool for data evaluation: Application to aquifer contamination. *Water Resources Management*, 2, 11–20. <http://doi.org/10.1007/BF00421927>
- Beven, K. (2005). On the concept of model structural error. *Water Science & Technology*, 52, 167–175.
- Corcoran, J. M., Knight, J. F., & Gallant, A. L. (2013). Influence of multi-source and multi-temporal remotely sensed and ancillary data on the accuracy of random forest classification of wetlands in northern Minnesota. *Remote Sensing*, 5, 3212–3238. <https://doi.org/10.3390/rs5073212>
- 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. *Advances in Water Resources*, 26, 137–149. [http://doi.org/10.1016/S0309-1708\(02\)00088-X](http://doi.org/10.1016/S0309-1708(02)00088-X)
- Dai, C., Xue, L., Zhang, D., & Guadagnini, A. (2016). Data-worth analysis through probabilistic collocation-based ensemble Kalman filter. *Journal of Hydrology*, 540, 488–503. <http://doi.org/10.1016/j.jhydrol.2016.06.037>
- Davis, D. R., and Kisiel, C. C., & Duckstein, L. (1972). Bayesian decision theory applied to design in hydrology. *Water Resources Research*, 8, 33–41. <http://doi.org/10.1029/WR008i001p00033>
- De Lannoy, G. J., Houser, P. R., Pauwels, V. R. N., & Verhoest, N. E. C. (2007). State and bias estimation for soil moisture profiles by an ensemble Kalman filter: Effect of assimilation depth and frequency. *Water Resources Research*, 43(6). <http://doi.org/10.1029/2006WR005100>
- Demissie, Y. K., Valocchi, A. J., Minsker, B. S., & Bailey, B. A. (2009). Integrating a calibrated groundwater flow model with error-correcting data-driven models to improve predictions. *Journal of Hydrology*, 364, 257–271. <http://doi.org/10.1016/j.jhydrol.2008.11.007>
- Dong, J., Steele-Dunne, S. C., Judge, J., & van de Giesen, N. (2015). A particle batch smoother for soil moisture estimation using soil temperature observations. *Advances in Water Resources*, 83, 111–122. <http://doi.org/10.1016/j.advwatres.2015.05.017>
- Dong, J., Steele-Dunne, S. C., Ochsner, T. E., & van de Giesen, N. (2015). Determining soil moisture by assimilating soil temperature measurements using the ensemble Kalman filter. *Advances in Water Resources*, 86, 340–353. <http://doi.org/10.1016/j.advwatres.2015.08.011>
- Dong, J., Steele-Dunne, S. C., Ochsner, T. E., & van de Giesen, N. (2016). Estimating soil moisture and soil thermal and hydraulic properties by assimilating soil temperatures using a particle batch

- smoother. *Advances in Water Resources*, 91, 104–116. <http://doi.org/10.1016/j.advwatres.2016.03.008>
- Drécourt, J., Madsen, H., & Rosbjerg, D. (2006). Bias aware Kalman filters: Comparison and improvements. *Advances in Water Resources*, 29, 707–718. <http://doi.org/10.1016/j.advwatres.2005.07.006>
- Geiges, A., Rubin, Y., & Nowak, W. (2015). Interactive design of experiments: A priori global versus sequential optimization, revised under changing states of knowledge. *Water Resources Research*, 51, 7915–7936. <http://doi.org/10.1002/2015WR017193>
- Hamill, T. M., & Whitaker, J. S. (2005). Accounting for the error due to unresolved scales in ensemble data assimilation: A comparison of different approaches. *Monthly Weather Review*, 133, 3132–3147. <http://doi.org/10.1175/MWR3020.1>
- Hendricks Franssen, H. J., & Kinzelbach, W. (2008). Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research*, 44(9). <http://doi.org/10.1029/2007WR006505>
- Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of The Royal Statistical Society, Series B (Statistical Methodology)*, 63, 425–464. <http://doi.org/10.1111/1467-9868.00294>
- 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. *Stochastic Environmental Research And Risk Assessment*, 32, 2477–2493. <http://doi.org/10.1007/s00477-018-1541-1>
- Man, J., Zhang, J., Li, W., Zeng, L., & Wu, L. (2016). Sequential ensemble-based optimal design for parameter estimation. *Water Resources Research*, 52, 7577–7592. <http://doi.org/10.1002/2016WR018736>
- Neuman, S. P., Xue, L., Ye, M., & Lu, D. (2012). Bayesian analysis of data-worth considering model and parameter uncertainties. *Advances in Water Resources*, 36, 75–85. <http://doi.org/10.1016/j.advwatres.2011.02.007>
- Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263, 641–646. <https://doi.org/10.1126/science.263.5147.641>
- Pathiraja, S., Moradkhani, H., Marshall, L., Sharma, A., & Geenens, G. (2018). Data-driven model uncertainty estimation in hydrologic data assimilation. *Water Resources Research*, 54, 1252–1280. <http://doi.org/10.1002/2018WR022627>
- Pauwels, V. R., & De Lannoy, G. J. (2015). Error covariance calculation for forecast bias estimation in hydrologic data assimilation. *Advances in Water Resources*, 86, 284–296. <http://doi.org/10.1016/j.advwatres.2015.05.013>
- Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
- Richards, L. A. (1931). Capillary conduction of liquids through porous mediums. *Physics*, 1, 318–333. <http://doi.org/10.1063/1.1745010>
- Shi, L., Song, X., Tong, J., Zhu, Y., & Zhang, Q. (2015). Impacts of different types of measurements on estimating unsaturated flow parameters. *Journal of Hydrology*, 524, 549–561. <http://doi.org/10.1016/j.jhydrol.2015.01.078>
- Song, X., Shi, L., Ye, M., Yang, J., & Navon, M. (2014). Numerical comparison of iterative ensemble Kalman filters for unsaturated flow inverse modeling. *Vadose Zone Journal*, 13(2). <http://doi.org/10.2136/vzj2013.05.0083>
- Sun, A. Y., and Wang, D., & Xu, X. (2014). Monthly streamflow forecasting using gaussian process regression. *Journal of Hydrology*, 511, 72–81. <http://doi.org/10.1016/j.jhydrol.2014.01.023>
- van Genuchten, M. Th. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44, 892–898. <http://doi.org/10.2136/sssaj1980.03615995004400050002x>
- Wagner, W., and Lemoine, G., & Rott, H. (1999). A method for estimating soil moisture from ERS scatterometer and soil data. *Remote Sensing of Environment*, 70, 191–207. [http://doi.org/10.1016/S0034-4257\(99\)00036-X](http://doi.org/10.1016/S0034-4257(99)00036-X)
- Walker, J. P., and Willgoose, G. R., & Kalma, J. T. (2001). One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application. *Journal of Hydrometeorology*, 2, 356–373. [http://doi.org/10.1175/1525-7541\(2001\)002%3c0356:ODSMR%3e2.0.CO;2](http://doi.org/10.1175/1525-7541(2001)002%3c0356:ODSMR%3e2.0.CO;2)
- Wang, Y., Shi, L., Zha, Y., Li, X., Zhang, Q., & Ye, M. (2018). Sequential data-worth analysis coupled with ensemble kalman filter for soil water flow: A real-world case study. *Journal of Hydrology*, 564, 76–88. <http://doi.org/10.1016/j.jhydrol.2018.06.059>
- Xu, Q. (2007). Measuring information content from observations for data assimilation: Relative entropy versus Shannon entropy difference. *Tellus A: Dynamic Meteorology and Oceanography*, 59, 198–209. <http://doi.org/10.1111/j.1600-0870.2006.00222.x>
- Xu, T., Valocchi, A. J., Choi, J., & Amir, E. (2014). Use of machine learning methods to reduce predictive error of groundwater models. *Groundwater*, 52, 448–460. <http://doi.org/10.1111/gwat.12061>
- Xu, T., & Valocchi, A. J. (2016). A Bayesian approach to improved calibration and prediction of groundwater models with structural error. *Water Resources Research*, 51, 9290–9311. <http://doi.org/10.1002/2015WR017912>
- 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 Resources Research*, 53, 4084–4105. <http://doi.org/10.1002/2016WR019831>
- Yu, D., Yang, J., Shi, L., Zhang, Q., Huang, K., Fang, Y., & Zha, Y. (2019). On the uncertainty of initial condition and initialization approaches in variably saturated flow modeling. *Hydrology and Earth System Sciences*, 23, 2897–2914. <http://doi.org/10.5194/hess-23-2897-2019>
- Zha, Y., Shi, L., Ye, M., & Yang, J. (2013). A generalized Ross method for two-and three-dimensional variably saturated flow. *Advances in Water Resources*, 54, 67–77. <http://doi.org/10.1016/j.advwatres.2013.01.002>
- Zha, Y., Zhu, P., Zhang, Q., Mao, W., & Shi, L. (2019). Investigation of data assimilation methods for soil parameter estimation with different types of data. *Vadose Zone Journal*, 18(1). <http://doi.org/10.2136/vzj2019.01.0013>
- Zhang, J., Li, W., Zeng, L., & Wu, L. (2016). An adaptive Gaussian process-based method for efficient Bayesian experimental design in groundwater contaminant source identification problems. *Water Resources Research*, 52, 5971–5984. <http://doi.org/10.1002/2016WR018598>
- Zhang, Q., Shi, L., Holzman, M., Ye, M., Wang, Y., Carmona, F., & Zha, Y. (2019). A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation. *Advances in Water Resources*, 132. <http://doi.org/10.1016/j.advwatres.2019.103407>

- Zook, A., Lee-Urban, S., Riedl, M. O., Holden, H. K., Sottolare, R. A., & Brawner, K. W. (2012). Automated scenario generation: toward tailored and optimized military training in virtual environments. In *Proceedings of the International Conference on the Foundations of Digital Games* (pp. 164–171). <https://doi.org/10.1145/2282338.2282371>
- Zupanski, D., & Zupanski, M. (2006). Model error estimation employing an ensemble data assimilation approach. *Monthly Weather Review*, 134, 1337–1354. <http://doi.org/10.1175/MWR3125.1>

How to cite this article: Wang Y, Shi L, Lin L, Holzman M, Carmona F, Zhang Q. A robust data-worth analysis framework for soil moisture flow by hybridizing sequential data assimilation and machine learning. *Vadose Zone J.* 2020;19:e20026. <https://doi.org/10.1002/vzj2.20026>