Data worth analysis within a model-free data assimilation framework for soil moisture flow

Yakun Wang[1], Xiaolong Hu[2], Lijun Wang[2], Jinmin Li[2], Lin Lin[2], Kai Huang[3], Liangsheng Shi[2*]

[1]Key Laboratory of Agricultural Soil and Water Engineering of in Arid and Semiarid Areas, Ministry of Education, Northwest A & F University, Yangling, Shaanxi 712100, China

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

[3]Guangxi Key Laboratory of Water Engineering Materials and Structures, Guangxi Institute of Water Resources Research, Nanning530023, China

Correspondence to: Liangsheng Shi (liangshs@whu.edu.cn)

Abstract. Conventional data-worth (DW) analysis for soil water problems depends on physical dynamic models. The widespread occurrence of model structural errors and the strong nonlinearity of soil water flow may lead to biased or wrong worth assessment. By introducing the nonparametric data-worth analysis (NP-DWA) framework coupled with the ensemble Kalman filter (EnKF), this real-world case study attempts to assess the worth of potential soil moisture observations regarding the reconstruction of fully data-driven soil water flow models prior to data gathering. The DW of real-time soil moisture observations after Gaussian process training and Kalman update was quantified with three representative information metrics, including the trace, Shannon entropy difference, and relative entropy. The sequential NP-DWA framework was examined by a number of cases in terms of the variable of interest, spatial location, observation error, and prior data content. Our results indicated that similar to the traditional DW analysis based on physical models, the overall increasing trend of the DW from the sequential augmentation of additional observations within the NP-DWA framework was also susceptible to interruptions by localized surges due to never-experienced atmospheric conditions (i.e., rainfall events). The difference is that this biased DW in the former is caused by model structural errors triggered by contrasting scenarios, which is difficult to be compensated by assimilating more prior data, while this performance degradation in the NP-DWA can be effectively alleviated by enriching training scenarios or the appropriate amplification of observational noise under extreme meteorological conditions. Nevertheless, a substantial expansion of the prior data content may cause an unexpected increase in DW of future potential observations due to the possible introduction of ensuing observation noises. Hence, high-quality and representative “small” data may be a better choice than unfiltered “big” data. Compared with the observations in the surface layer with the strongest time-variability, the soil water content in the middle layer robustly exhibited remarkable superiority in the construction of model-free soil moisture models. We also demonstrated that the DW assessment performance was jointly determined by ‘3C’, i.e., capacity of potential observation realizations to “capture” actual observations, correlation of potential observations with the variables of interest, and choice of DW indicators. Direct mapping from regular meteorological data to soil water content within the NP-DWA mitigated the adverse effects of nonlinearity-related interference, which thus facilitated the identification of the soil moisture covariance matrix, especially the cross-covariance.

Keywords: Data worth; Nonparametric data assimilation; Soil moisture; Gaussian process

1 Introduction

As one of the few directly observable hydrological variables, soil water content (SWC) exhibits critical importance in optimal water resource management, irrigation and drainage schemes, fertilizer application, and crop production in agriculture (Liu et al., 2011; Akhtar et al., 2019; Dobriyal et al., 2012; Gu et al., 2021). Various data assimilation (DA) approaches (Dunne and Entekhabi, 2005; Li and Ren, 2011; Reichle et al., 2008; Song et al., 2014) have been established to reconstruct the spatiotemporal dynamics of SWC from noisy or partial observations. The core of these traditional parametric filters is their reliance on repeated forward integrations of an explicitly known physical model of unsaturated flow, such as the HYDRUS (Šimůnek et al., 2006), Soil and Water Assessment Tool (SWAT) (Van Dam and Feddes, 2000), and Ross models (Ross, 2003; Zha et al., 2013).

Currently, the ever-increasing availability of multi-source data from remote sensing (Montzka et al., 2011; Shi et al., 2011), ground-based measurements (Li et al., 2018; Shuwen et al., 2005; Yang et al., 2000), and numerical modeling has paved the way for the development of fully data-driven techniques within the DA framework. In particular, recent advances in machine learning-based DA schemes (Brajard et al., 2020; Brajard et al., 2021; Yamanaka et al., 2019) offer exciting new opportunities for extracting patterns and insights of soil moisture dynamics from data (Ju et al., 2018; Li et al., 2020; Liu et al., 2020; Wang et al., 2021a). For instance, Kashif Gill et al. (2007) proposed a hybrid DA methodology that combined support vector machines and the ensemble Kalman filter (EnKF) for soil moisture dynamics. Li et al. (2020) compared the performance of a physical-based model with DA and machine learning methods in terms of the simulation of soil water dynamics under synthetic and real-world conditions. Wang et al. (2021a) and Wang et al. (2021b) further attempted to learn unknown relationships between SWC as well as its spatio-temporal gradients and highly accessible data via the Gaussian process (GP) regression.

Notwithstanding the success of these model-free DA schemes built on machine learning for unsaturated flow, essential caveats and limitations have hampered their further adoption and impact. First, the amount of data required to infer nonlinear relationships in unsaturated flow problems may be overwhelming (Hughes, 1968), thus greatly increasing the data collection budget. Subsequently, addressing the abovementioned explosive data growth is also a challenging and time-demanding task requiring an extensive computational infrastructure. Second, the performance and quality of the knowledge extracted by machine learning algorithms are highly dependent on the quality and suitability of data (García-Gil et al., 2019). Unfortunately, data gathering is rarely perfect, and data corruption often occurs (Wang et al., 2018). The identification of the multi-source SWC data quality or measurement error is not an easy task. This limitation instead can create extra uncertainties in DA systems (Kisekka et al., 2015). Third, it is the diversity of scenarios contained in prior data rather than its volume that is more decisive for the generalization ability of machine learning methods (Wang et al., 2020). Direct data fusion without screening may instead induce accidental correlations in learning algorithms, thereby diminishing their generalization ability (García et al., 2016). To avoid the overloaded monitoring cost due to redundant data, it is essential to develop a framework to assess the worth of alternative sampling strategies prior to data collection.

Data worth, sometimes called data information content or data impact, of a design is often defined as its individual capacity to reduce uncertainty associated with a prediction goal, or to maximize some related measure of data utility. Over the past decades, two main types of sophisticated DW analysis frameworks have been proposed to identify the most informative monitoring strategy in hydrology, namely, one type based on sensitivity analysis (Dausman et al., 2010; Fienen et al., 2010; Hill and Tiedeman, 2006) and the other within a fully Bayesian framework (Dai et al., 2016; Neuman et al., 2012; Nowak et al., 2012). The former approaches are computationally fast, but these methods require model calibration and assume linear models (Finsterle, 2015). The latter methods are derived based on the law of the total possibility, without assumptions of the model and of the distributions of observations and model parameters. Nevertheless, both well-established frameworks are predicated on the availability of the underlying physical models. For example, Man et al. (2016) evaluated the expected value of alternative SWC sampling strategies with respect to the estimation of soil hydraulic parameters in the Hydrus-1D model, while Finsterle (2015) examined the worth of datasets potentially applicable for the calibration of geothermal reservoir models. Within such parametric data-worth analysis frameworks, however, the strong nonlinearity of soil water problems (De Lannoy et al., 2006; Leube et al., 2012; Yeh et al., 1985) and the prevalence of model structural errors (Zhang et al., 2019) are highly likely to lead to biased data worth (Wang et al., 2018; Wang et al., 2020). Ultimately the reliability of the optimal design of monitoring networks based on such evaluations is greatly compromised.

Fortunately, the superiority of data-driven algorithms in handling nonlinearities and structural errors in unsaturated flow has been well demonstrated in our previous studies (Wang et al., 2021a; Wang et al., 2021b). With the explosive growth of big data, how to evaluate the worth of multi-source data in this new data mining approach is becoming a critical issue. Several recent works in the field of statistical learning have bloomed in identifying and removing irrelevant and redundant information from big data, such as feature selection (Chandrashekar and Sahin, 2014; Hall, 1999) and instance reduction (Al-Akhras et al., 2021; Olvera-López et al., 2010). To the best of our knowledge, few studies have systematically evaluated the worth of future observations regarding the construction of fully data-driven models prior to data gathering. As a follow-up study of Wang et al. (2018) and Wang et al. (2020), one major contribution of this study is the first embedding of a purely data-driven model into the Bayesian data-worth analysis framework, referring to as the nonparametric data-worth analysis (NP-DWA). Similar to traditional DW analysis, the proposed NP-DWA consists of prior, posterior, and preposterior stages (Dai et al., 2016). The preposterior analysis evaluates the anticipated worth of future observations regarding the construction of purely data-driven models, for which possible distributions are predicted in advance by conditioning on prior data.

There is a consensus in the field of statistics that “the highest accuracy results that an inductive learning system can achieve depend on the quality of data and the appropriate selection of a learning algorithm for the data” (Pechenizkiy et al., 2006). In other words, once the algorithm specified, the significance of data noise on learning accuracy as almost the only factor should not be overlooked. Considering the powerful ability of dealing with observational noises of the ensemble Kalman filter (EnKF) (Hamilton et al., 2017; Li et al., 2018), another innovation of this study is the introduction of

EnKF into our NP-DWA framework. In conventional DW analysis, the worth of data is primarily embodied in its ability to be utilized or calibrated to adjust physical parameters (Dai et al., 2016; Finsterle, 2015; Man et al., 2016). In the proposed NP-DWA, nevertheless, future observations are first used to construct data-driven models in the forecast step and then sequentially assimilated with the Kalman update in the analysis step. Ultimately, its combined capacity to reduce system uncertainties in these two ways is defined as its worth. Furthermore, as a typical sequential DA scheme, the EnKF facilitates the dynamic models as well as its hyperparameters to be updated in real-time, so the data utility to modeling system can be detected instantaneously. Eventually, the sampling scheme can be dynamically adjusted to save the monitoring and analysis costs.

Most previous studies are based on synthetic cases, and data-worth analysis in the context of dynamically evolving soil moisture profiles was still poorly studied in a real-world case. For nonlinear problems, nevertheless, the estimation variance and more sophisticated measures of data utility depend on the actual values of measurements, which are still unknown prior to collection (Leube et al., 2012). It will be more convincing to investigate the data worth regarding the reconstruction of fully data-driven models under real-world cases for unsaturated flow. With the aid of observed data retrieved from three typical stations with different climate regimes, we aim to shed light on the following questions: (1) as opposed to the traditional way of utilizing data (to calibrate physical parameters), is the worth of observations capable of being accurately quantified by NP-DWA in this new, purely data-driven approach? (2) Given multiple prediction objectives, how does the DW (in the form of various indices) evolve under different hydrometeorological conditions in the determination of fully data-driven soil moisture dynamics? (3) How does the proposed NP-DWA respond to the presence of multiple levels of data noise? It is strived that this study can provide guidance in the design of future monitoring strategies within the fully data-driven soil water flow models for real-world problems.

The remainder of this paper is organized as follows: Sect. 2 first summarizes the experimental data and methods. Thereinto, the principles of Bayesian DW analysis, nonparametric DA, and the hybrid framework are detailed. Sect. 3 presents the results, and a discussion is contained in Sect. 4. Finally, conclusions are outlined in Sect. 5.

2 Methodology

In Wang et al. (2021a), a nonparametric sequential data assimilation scheme (Kalman-GP) has been proposed based on the filtering equations of EnKF and data-driven modeling with GP. On top of that, this paper further develops a nonparametric data-worth analysis framework to assess the potential worth of future observations in the reconstruction of dynamical soil water flow models prior to data collection. In our proposed NP-DWA framework, the mapping from the input vector $\boldsymbol{x}$ and state variable of interest *y* is approximated by the GP regression model. In this study, the entries of $\boldsymbol{x}$ included the observation time, depth, daily precipitation, and air temperature, while the output *y* is the corresponding soil moisture. Thus, here the dimension of input $\boldsymbol{x}$ is *d*=4. Similar to the Kalman-GP in Wang et al. (2021a), EnKF is implemented to update the forecast yield by GP models by assimilating real-time observations of soil moisture. On the one hand, this fusion can effectively reduce the risk of unreasonable spatio-temporal interpolation in GP models, ultimately enhancing the robustness of such purely data-driven models. On the other hand, by combining with the Kalman update, the forecast cross-covariance between the state and the predictions corresponding to available observations constrained the otherwise high error covariances of state variables at unobserved locations, which resulted in a significantly reduced uncertainty for this hybrid method relative to GP alone. More details can be found in Wang et al. (2021a). In addition, to demonstrate the generalizability of the NP-DWA, other machine learning algorithms (including support vector machine and random forest) and DA scheme (i.e., particle filter) are also employed to replace GP and EnKF, respectively. The specific details of these algorithms are not presented here, but the corresponding results are provided in the Supplementary.

2.1 Construction of GP Dynamic Models

As stated in Wang et al. (2021a), *N* GP models are constructed independently at each time step in a sequential manner as new data are recorded. *N* is the ensemble size. Thereinto, the observed time series is corrupted by the prescribed observation noises satisfying Gaussian distribution to obtain *N* sets of training data. Here we only take the procedure of building the *m*th GP model at any time *t*=*k* as an example, i.e., GP_k^m (*m*=1, 2, …, *N*). The superscript *m* will be omitted below for ease of expression.

At *t*=*k*, the input $\boldsymbol{X}_{1:(k-1)}$ and output $\boldsymbol{y}_{1:(k-1)}$ of training data in GP_k can be expressed as:

$$\boldsymbol{X}_{1:(k-1)} = \left[\boldsymbol{x}_1^1, \boldsymbol{x}_1^2, \dots, \boldsymbol{x}_1^{q_1}, \dots, \boldsymbol{x}_p^1, \boldsymbol{x}_p^2, \dots, \boldsymbol{x}_p^{q_p}, \dots., \boldsymbol{x}_{k-1}^1, \boldsymbol{x}_{k-1}^2, \dots, \boldsymbol{x}_{k-1}^{q_{k-1}}\right]^T \tag{1}$$

$$\boldsymbol{y}_{1:(k-1)} = \left[y_1^1, y_1^2, \dots, y_1^{q_1}, \dots, y_p^1, y_p^2, \dots, y_p^{q_p}, \dots., y_{k-1}^1, y_{k-1}^2, \dots, y_{k-1}^{q_{k-1}}\right]^T \tag{2}$$

where $\boldsymbol{X}_{1:(k-1)}$ and $\boldsymbol{y}_{1:(k-1)}$ denote a collection of all available $\boldsymbol{x}$ and *y* from *t*=1 to (*k*-1), respectively; q_p denotes the number of available observations at *t*=*p* (*p*=1, 2, …, *k*-1). In this paper we assume that the number of available observations at each time step is identical, i.e., $q_1 = q_2 = \cdots = q_{k-1} = q$. Hence the dimensions of matrix $\boldsymbol{X}_{k-1}$ and vector $\boldsymbol{y}_{k-1}$ are $q(k-1) \times d$ and $q(k-1)$, respectively.

As defined in Rasmussen (2003) and Williams and Rasmussen (2006), a GP model can be fully specified by a mean function $\mu(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$, i.e., $G(\boldsymbol{x}) \sim N(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}'))$. In this study, a linear mean function and an anisotropic squared exponent covariance function are specified (Zhang et al., 2019) as:

$$\mu(\boldsymbol{x}) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} \tag{3}$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 exp\left[-\sum_{l=1}^{d}\frac{(x_l - x_l')^2}{\tau_l^2}\right] \tag{4}$$

where $\boldsymbol{\beta}$ is vector containing d linear coefficients, i.e., $\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_d\}$; σ^2 controls the marginal variance in the output; and $\boldsymbol{\tau} = \{\tau_1, \tau_2, \ldots, \tau_d\}$ determines the dependence strength in each of the component directions of $\boldsymbol{x}$.

Then, the hyperparameters of the GP_k, $\boldsymbol{\emptyset} = \{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\tau}\}$, can be inferred from the training datasets $\{\boldsymbol{X}_{1:(k-1)}, \boldsymbol{y}_{1:(k-1)}\}$ via log marginal likelihood maximization:

$$\begin{aligned} L &= \log p(\boldsymbol{y}_{1:(k-1)} | \boldsymbol{X}_{1:(k-1)}, \boldsymbol{\emptyset}) \\ &= -\frac{1}{2}\left(\boldsymbol{y}_{1:(k-1)} - \boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_{1:(k-1)} - \boldsymbol{\mu}\right) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log 2\pi \end{aligned} \tag{5}$$

where $\boldsymbol{\mu}$ denotes the prior mean vector with the dimension of $q(k-1)$, and $\boldsymbol{\Sigma}$ denotes the covariance matrix whose elements in the *i*th row and *j*th column constitute $\Sigma_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ (*i*=1, 2, …, $q(k-1)$; *j*=1, 2, …, $q(k-1)$). The GPML MATLAB toolbox (version 4.2), as documented in Williams and Rasmussen (2006), was adopted for GP inference in this study (http://www.gaussianprocess.org/gpml/code/matlab/doc/).

In this study, the entire soil moisture profile at t=k is expected to be forecasted. Assuming that the total number of nodes of the vertical one-dimensional soil profile is N_n, then the input at the current time step is $\boldsymbol{X}_k{}^* = \left[\boldsymbol{x}_k^1, \boldsymbol{x}_k^2, \ldots, \boldsymbol{x}_k^{N_n}\right]^T$ with the dimension of $N_n \times d$. The corresponding output vector $\boldsymbol{y}_k^f$ with the dimension of N_n can be calculated as:

$$\boldsymbol{y}_k^f = \boldsymbol{\mu}^* + \boldsymbol{\Sigma}^{*\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_{1:(k-1)} - \boldsymbol{\mu}) \tag{6}$$

where $\boldsymbol{\mu}^*$ denotes the prior mean of $\mu(\boldsymbol{X}_k{}^*)$ with the dimension of N_n and $\boldsymbol{\Sigma}^*$ is calculated as $\boldsymbol{\Sigma}_i^* = k(\boldsymbol{x}_k^i, \boldsymbol{x}_j)$ (*i*=1, 2, …, N_n; *j*=1, 2, …, $q(k-1)$).

As a collection of $\boldsymbol{y}_k^f$ from *N* GP models, the resultant forecasted state vector $\boldsymbol{Y}_k^f$ at t=k can be represented as:

$$\boldsymbol{Y}_k^f = \left[\boldsymbol{y}_{k,1}^f, \boldsymbol{y}_{k,2}^f, \dots, \boldsymbol{y}_{k,m}^f, \dots, \boldsymbol{y}_{k,N}^f\right]^T \tag{7}$$

where $\boldsymbol{y}_{k,m}^f$ denotes the forecasted state vector of interest for GP_k^m (*m*=1, 2, …, *N*); the dimension of $\boldsymbol{Y}_k^f$ is $N \times N_n$; superscripts *f* denotes forecast.

2.2 The Kalman Update in Nonparametric Data Assimilation Scheme

In the analysis step of the EnKF, for any ensemble member *m* at *t*=*k*, the state vector can be updated by combing GP model predictions and observations $\boldsymbol{d}_k^{obs}$:

$$\boldsymbol{y}_{k,m}^a = \boldsymbol{y}_{k,m}^f + \boldsymbol{K}_k(\boldsymbol{d}_{k,m}^{obs} - \boldsymbol{H}\boldsymbol{y}_{k,m}^f) \tag{8}$$

where $\boldsymbol{y}_{k,m}^a$ denotes the improved estimates for realization *m* by conditioning on the observed information at *t*=*k*; $\boldsymbol{H}$ is the observation operator with the dimension of $q \times N_n$, which represents the relationship between the state and observation vectors; superscript *a* indicates analysis. and $\boldsymbol{d}_{k,m}^{obs}$ with the dimension of *q* denotes the observation vector at *t*=*k* for the *m*th ensemble member of $\boldsymbol{d}_k^{obs}$. It should be emphasized that the relationship between observations at *t*=*k*, $\boldsymbol{d}_k^{obs}$ and their true values $\boldsymbol{y}_k = \left[y_k^1, y_k^2, \dots, y_k^q\right]^T$ can be expressed as follows,

$$\boldsymbol{d}_k^{obs} = \boldsymbol{y}_k + \boldsymbol{\varepsilon}_k \tag{9}$$

where $\boldsymbol{\varepsilon}_k$ with the dimension of $N \times q$ represents measurement error vector which is assumed to be zero-mean Gaussian with $\boldsymbol{R}_k$; $\boldsymbol{R}_k$ denotes the error covariance matrix of the observations with the dimension of $q \times q$.

The Kalman gain at *t*=*k*, $\boldsymbol{K}_k$, with the dimension of $N_n \times q$ can be defined as:

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

where $\boldsymbol{C}_k^f$ with the dimension of $N_n \times N_n$ is the covariance matrix of the state vector at *t*=*k*, which can be approximated as:

$$\boldsymbol{C}_k^f \approx \frac{1}{N-1} \sum_{m=1}^{N} \left\{\left[\boldsymbol{y}_{k,m}^f - \langle \boldsymbol{Y}_k^f \rangle\right]\left[\boldsymbol{y}_{k,m}^f - \langle \boldsymbol{Y}_k^f \rangle\right]^{\mathrm{T}}\right\} \tag{11}$$

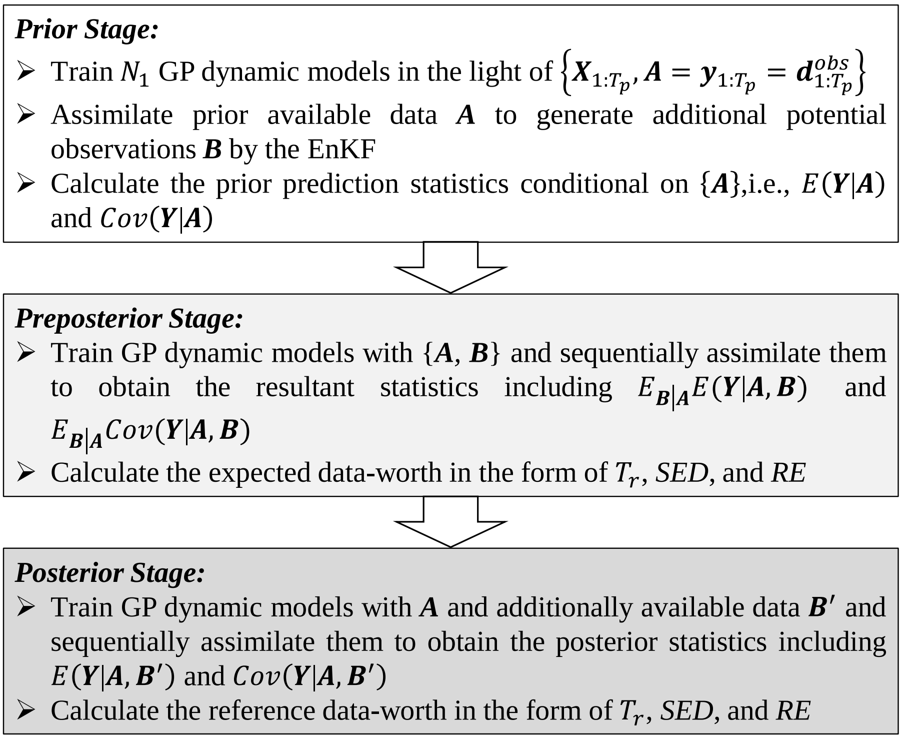
where $\langle \boldsymbol{Y}_k^f \rangle$ denotes the ensemble mean of $\boldsymbol{Y}_k^f$.

2.3 Nonparametric Data-worth Analysis Framework

Following the methodologies of Neuman et al. (2012) and Dai et al. (2016), data-worth analysis of future monitoring networks within the aforementioned NP-DWA framework also consists of three stages. The whole workflow of the NP-DWA framework is depicted in Fig. 1.

2.3.1 Prior Stage

At the prior stage ($0 < t \leq T_p$), the integration of GP dynamic models and EnKF with an ensemble size of $N = N_1$ is implemented to sequentially train and assimilate the prior data via Eqs. 1–11. Here, all available prior datasets from t=0 to t=T_p are denoted as vector $\boldsymbol{A} = \boldsymbol{y}_{1:T_p} = \boldsymbol{d}_{1:T_p}^{obs}$ with the dimension of qT_p, while the corresponding GP input is denoted as matrix $\boldsymbol{X}_{1:T_p}$ with the dimension of $qT_p \times d$. Then, a set of N_1 hypothetical observations can be generated, denoted as $\boldsymbol{B}_{k,i} = \boldsymbol{H}_k \boldsymbol{y}_{k,i}^f$ ($k = T_p + 1, T_p + 2, \dots, T_t$; $i = 1, 2, \dots, N_1$), via Eq. 6. T_t is the total simulation time. Moreover, prior prediction statistics (mean and covariance) of posterior vector $\boldsymbol{Y}_k$, i.e., $E(\boldsymbol{Y}|\boldsymbol{A})$ and $Cov(\boldsymbol{Y}|\boldsymbol{A})$, can be yielded conditional on $\{\boldsymbol{A}\}$, which can be denoted as $\boldsymbol{E}_1$ and $\boldsymbol{C}_1$, respectively, for the sake of simplicity.



Prior Stage:

- Train N_1 GP dynamic models in the light of $\left\{\boldsymbol{X}_{1:T_p}, \boldsymbol{A} = \boldsymbol{y}_{1:T_p} = \boldsymbol{d}_{1:T_p}^{obs}\right\}$
- Assimilate prior available data $\boldsymbol{A}$ to generate additional potential observations $\boldsymbol{B}$ by the EnKF
- Calculate the prior prediction statistics conditional on $\{\boldsymbol{A}\}$,i.e., $E(\boldsymbol{Y}|\boldsymbol{A})$ and $Cov(\boldsymbol{Y}|\boldsymbol{A})$

Preposterior Stage:

- Train GP dynamic models with $\{\boldsymbol{A}, \boldsymbol{B}\}$ and sequentially assimilate them to obtain the resultant statistics including $E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$ and $E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$
- Calculate the expected data-worth in the form of T_r, *SED*, and *RE*

Posterior Stage:

- Train GP dynamic models with $\boldsymbol{A}$ and additionally available data $\boldsymbol{B'}$ and sequentially assimilate them to obtain the posterior statistics including $E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B'})$ and $Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B'})$
- Calculate the reference data-worth in the form of T_r, *SED*, and *RE*

Figure 1. The workflow of nonparametric data-worth analysis framework coupled with Ensemble Kalman Filter (EnKF)

2.3.2 Preposterior Stage

At the preposterior stage ($T_p + 1 < t \leq T_t$), for each possible data $\boldsymbol{B}_{k,i}$ at t=k, N_2 realizations satisfying a Gaussian distribution are further generated. The ensemble mean is the value of $\boldsymbol{B}_{k,i}$, while the variance is the measurement error. Since this method is recursive, the time index k is omitted in the following equations. Then, the integration of GP models and EnKF is again implemented through a set of N_2 Monte Carlo realizations for each of the N_1 hypothetical observations. This allows us to calculate prediction statistics of the posterior state vector $\boldsymbol{Y}_{ij}$ $(i = 1, 2, \ldots, N_1;\ j = 1, 2, \ldots, N_2)$, i.e., $E(\boldsymbol{Y}_i|\boldsymbol{A}, \boldsymbol{B}_i)$ and $Cov(\boldsymbol{Y}_i|\boldsymbol{A}, \boldsymbol{B}_i)$, conditional on $\{\boldsymbol{A}, \boldsymbol{B}_i\}$. Finally, quantities $E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$, $E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$, and $Cov_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$ can be yielded by averaging over the collection of $N_1 \times N_2$ realizations. It should be emphasized that $E_{\boldsymbol{B}|\boldsymbol{A}}E(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$ and $E_{\boldsymbol{B}|\boldsymbol{A}}Cov(\boldsymbol{Y}|\boldsymbol{A}, \boldsymbol{B})$ represent the preposterior prediction mean and uncertainty after the addition of future possible data $\boldsymbol{B}$, which can be denoted as $\boldsymbol{E}_2$ and $\boldsymbol{C}_2$, respectively.

To quantify the expected data-worth of potential measurements, three commonly considered information metrics, including the trace (T_r), Shannon entropy difference (*SED*), and relative entropy (*RE*), are introduced in this study. T_r and *SED* offer scalar indices to measure the decrease in variance and covariance, respectively, while the *RE* comprehensively quantifies both mean and covariance effects.

(1) Trace

As a scalar indicator (Dai et al., 2016), T_r quantifies the DW in terms of variance reduction as follows:

$$T_r = T_r(\boldsymbol{C}_1) - T_r(\boldsymbol{C}_2) \tag{12}$$

where $T_r(*)$ denotes the trace (sum of the diagonal entries) of a matrix.

(2) Shannon entropy difference

According to Shannon (1949), the Shannon entropy (*SE*) of *PDF p*(*x*) can be defined as:

$$SE(p) = -\int p(x) \ln p(x)\, dx, x \in R \tag{13}$$

The *SED* between the prior and preposterior PDFs can also be considered to quantify the information content extracted from additional observations. Assuming that these two PDFs are both Gaussian in the EnKF model, the *SED* can be expressed in terms of covariance reduction (Xu, 2007) as:

$$SED = \frac{\ln \det(\boldsymbol{C}_1)}{2} - \frac{\ln \det(\boldsymbol{C}_2)}{2} = \frac{\ln \det(\boldsymbol{C}_1 \boldsymbol{C}_2^{-1})}{2} \tag{14}$$

where det(∗) is the determinant of a matrix.

(3) Relative entropy

Similar to the *SED*, the *RE* also provides a measure of the information content of the preposterior PDF with respect to the prior PDF. In addition to uncertainty reduction, the influence of future data on the mean behavior of PDFs is considered (Singh et al., 2013; Zhang et al., 2015). Considering that the prior and preposterior PDFs are *n*-dimensional Gaussian functions, the *RE* can be defined as:

$$RE = \frac{1}{2}(\boldsymbol{E}_2 - \boldsymbol{E}_1)^T \boldsymbol{C}_1^{-1}(\boldsymbol{E}_2 - \boldsymbol{E}_1) + \frac{1}{2}[\ln \det (\boldsymbol{C}_1 \boldsymbol{C}_2^{-1}) + T_r(\boldsymbol{C}_2 \boldsymbol{C}_1^{-1}) - n] \quad (15)$$

Finally, the expected DW of $\boldsymbol{B}_k$ can be estimated in the form of the above three indices prior to data gathering. Similar procedures are repeated until the final time *t*=T_t is reached.

2.3.3 Posterior Stage

At the posterior stage ($T_p + 1 < t \le T_t$), the available actual dataset $\boldsymbol{B}'$ is incorporated into the GP training datasets and assimilated in a sequential manner. The actual mean and covariance of posterior state vector $\boldsymbol{Y}$, i.e., $E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')$ and $Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')$, respectively, are obtained conditional on $\{\boldsymbol{A},\boldsymbol{B}'\}$. The reference data-worth in the form of the various indices can be calculated via Eqs. 12–15, where $\boldsymbol{E}_2$ and $\boldsymbol{C}_2$ are replaced with $E(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')$ and $Cov(\boldsymbol{Y}|\boldsymbol{A},\boldsymbol{B}')$, respectively.

3 Description of experimental data and model setup

3.1 Data Sources and Site Description

Three typical sites, including Falkenberg (52.1669 N, 14.1241 E), Cape_Charles_5_ENE (37.2907 N, 75.9270 W, hereafter referred to as Cape), and DAHRA (15.4035 N, 15.4320 W) were selected from the International Soil Moisture Network (ISMN, https://ismn.geo.tuwien.ac.at/en/) to evaluate the performance of the proposed NP-DWA framework under different soil types and climatic regimes. According to the dominant fraction of clay, silt, and sand for two layers (topsoil: 0.0–0.3 m, subsoil: 0.3–1.0 m) provided by ISMN, we use the USDA soil texture classification and classified the soil at three sites. The soil at Falkenberg is sandy loam, and the DAHRA soil is loamy sand. The topsoil and subsoil at Cape are clay loam and loamy clay, respectively. At these three sites, the in situ volumetric SWC was operationally measured with TRIME-EZ (IMKO), Stevens Hydraprobe II Sdi-12 (Stevens Water Inc.), and ThetaProbe ML2X (Delta-T Devices) instruments, respectively. The measurement depths were (1) 0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m at the Falkenberg site, (2) 0.05, 0.10, 0.20, 0.50, and 1.00 m at the Cape site, and (3) 0.05, 0.10, 0.50, and 1.00 m at the DAHRA site. The measurement error was artificially specified to be 0.02 cm^3/cm^3 unless otherwise specified.

Apart from soil water measurements at different depths, the daily precipitation and air temperature at the height of 2 meters were obtained from the ISMN. At each site, 200-day time series (from January 15 and August 2 in 2005 at the Falkenberg site, from April 24 to November 9 in 2004 at the Cape site, and from April 9 to October 25 in 2011 at the DAHRA site) were collected in this study, as shown in Fig. 2. Having a continental climate, the Falkenberg receives frequent but less intense precipitation during the simulation period. The Cape has a humid subtropical climate with the highest rates of rainfall among the three sites, and there were a few rainstorm events during the study period (e.g., up to 150 mm/d on September 8, 2014). The region of DAHRA has a tropical climate with well-defined dry and humid seasons. The early stage of the simulation is in its dry season, with little to no rainfall. The late stage is in its humid season when frequent but less intense rainfall events occur and the daily average air temperature is about 30 ◦C.

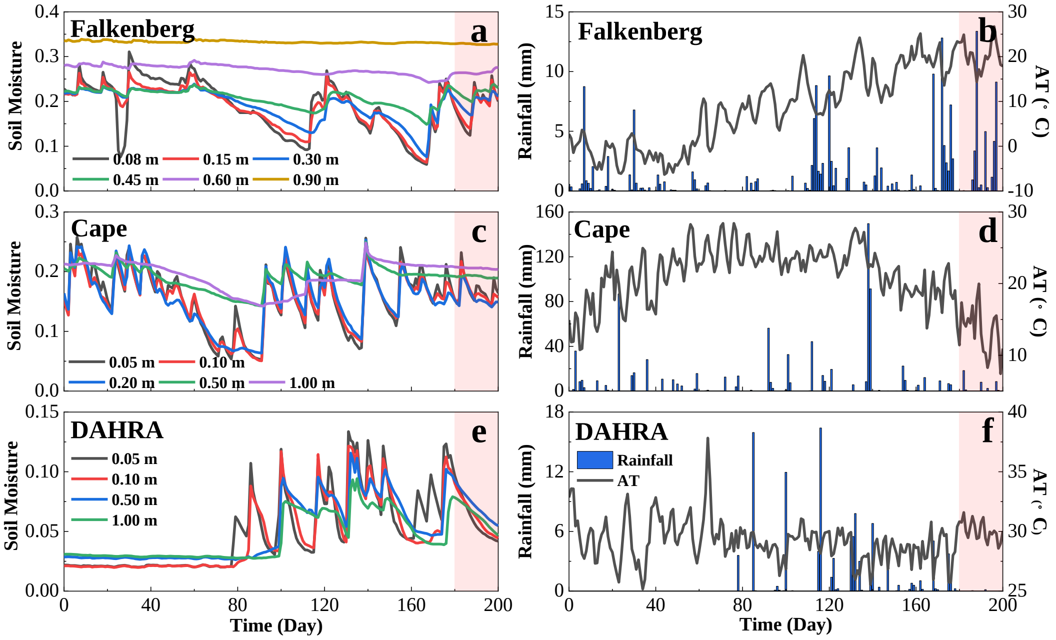


Figure 2. The temporal evolutions of soil moisture at various depths, daily rainfall, and mean daily air temperature (AT) at 2 meters height at Falkenberg, Cape, and DAHRA, respectively. Note that the red area indicates the preposterior or posterior stage

3.2 Model Simulation Setup and Case Design

The key parameters of this study are summarized in Table 1. Each site is represented by a one-dimensional soil column with a height of 1 m, which is discretized into 2cm grids with the local refinement of 1-cm monitoring depth intervals, i.e., z=0.15 m and 0.45 m at the Falkenberg site and z=0.05 m at the Cape and DAHRA sites. At each time step, $N_1 = 50$ GP-based dynamic models of unsaturated flow are constructed. The GP model input $\boldsymbol{x}$ includes the observation time, depth, daily precipitation, and air temperature, while the output is the corresponding soil moisture. The state vector

$\boldsymbol{y}$ comprises the soil moisture for all nodes at each site, and the trained and assimilated observations $\boldsymbol{d}^{obs}$ refer to the available soil moisture at all observed depths (as described in Sect. 3.1).

We illustrate our approach based on a set of real-world test cases, as listed in Table 2. The performance of the three indices, namely, T_r, *SED*, and *RE*, in data-worth quantification are compared at all three sites. In all test cases of this study, prior data for training GP includes the soil water content at all observed depths during the prior stage (from *t*=1 to *t*=T_p), i.e., *z*=0.08, 0.15, 0.30, 0.45, 0.60, and 0.90 m at Falkenberg, *z*=0.05, 0.10, 0.20, 0.50, and 1.00 m at Cape, and *z*=0.05, 0.10, 0.50, and 1.00 m at the DAHRA. At the preposterior or posterior stage in the NP-DWA, the worth of potential observations regarding the retrieval of three quantities of interest, including $\theta_{1.00}^{ave}$, $\theta_{0.60}^{ave}$, and $\theta_{0.30}^{ave}$, is evaluated. Here, $\theta_{1.00}^{ave}$, $\theta_{0.60}^{ave}$, and $\theta_{0.30}^{ave}$ represent the average soil moisture in the top 1.00 m, 0.60 m, and 0.30 m, respectively. A comparison among cases TC1, TC2, and TC3, is designed to investigate the data-worth response of surface (θ_S), middle (θ_M), and deep (θ_D) SWC regarding the above different prediction objectives. Specifically, θ_S refers to soil moisture at *z*=0.08 m at Falkenberg, and *z*=0.05 m at Cape and DAHRA, respectively. θ_M refers to soil moisture at *z*=0.45 m at Falkenberg, and *z*=0.50 m at Cape and DAHRA, respectively. θ_D refers to soil moisture at *z*=0.90 m at Falkenberg, and *z*=1.00 m at Cape and DAHRA, respectively. The prior datasets entering these cases comprise SWC, daily precipitation, and air temperature over the first 80 days, as shown in the gray areas of Fig. 2. The subsequent 20-day data (red areas in Fig. 2) are augmented as additional data for reference DW assessment in the posterior stage.

Table 1. The summary of key parameters

Parameter	Value
Description of soil column	
Soil column height [m]	1.00
No. of Nodes, N_n	53 (Falkenberg)/52 (Cape&DAHRA)
Number of realizations	
N_1	50
N_2	50
Prior values of GP hyperparameters	
τ_1, τ_2, τ_3, τ_4	1
σ^2	0.5
β_1, β_2, β_3, β_4	0

Table 2. The summary of designed test cases and main characteristics

Case Name		Potential Observation	Observation Error	Prior Data (d)	Variable of Interest
TC1	TC1-1	θ_S	0.02^2	80	$\theta_{1.00}^{ave}$
	TC1-2	θ_M	0.02^2	80	$\theta_{1.00}^{ave}$
	TC1-3	θ_D	0.02^2	80	$\theta_{1.00}^{ave}$
TC2	TC2-1	θ_S	0.02^2	80	$\theta_{0.60}^{ave}$
	TC2-2	θ_M	0.02^2	80	$\theta_{0.60}^{ave}$
	TC2-3	θ_D	0.02^2	80	$\theta_{0.60}^{ave}$
TC3	TC3-1	θ_S	0.02^2	80	$\theta_{0.30}^{ave}$
	TC3-2	θ_M	0.02^2	80	$\theta_{0.30}^{ave}$
	TC3-3	θ_D	0.02^2	80	$\theta_{0.30}^{ave}$
TC4		θ_S	0.01^2	80	$\theta_{1.00}^{ave}$
TC5		θ_S	0.04^2	80	$\theta_{1.00}^{ave}$
TC6		θ_S	0.02^2	40	$\theta_{1.00}^{ave}$
TC7		θ_S	0.02^2	180	$\theta_{1.00}^{ave}$
TC8		$\theta_S,\ \theta_M$	0.02^2	80	$\theta_{1.00}^{ave}$
TC9		$\theta_S,\ \theta_M,\ \theta_D$	0.02^2	80	$\theta_{1.00}^{ave}$

As stated in Pechenizkiy et al. (2006) and Zhu and Wu (2004), the maximum accuracy of statistical learning algorithms mainly depends on the quality of training data, in addition to the inherent bias in the algorithm itself. In other words, the magnitude and accuracy of the expected worth of driving data in machine learning-based DA may be closely related to the noise level. Thus, two additional test cases (TC4 and TC5) are considered to evaluate the performance of the proposed NP-DWA framework under different measurement errors. The soil moisture measurement error variance values of 0.01^2 and 0.04^2 are artificially specified in TC4 and TC5, respectively, to be compared to a value of 0.02^2 in TC1-1.

Moreover, test cases TC6 and TC7 differ from test case TC1-1. These test cases are designed to investigate the influence of the prior data content on data-worth analysis, which facilitates the determination of the required prior information content to ensure the accuracy of data-worth assessment. The 80-day prior data in test case TC1-1 are reduced backward in time to 40 days in test case TC6 and augmented forward to 180 days in test case TC7. In addition, test cases TC1-1, TC8, and TC9 consider the composite DW of different combinations of monitoring schemes. The comprehensive contributions of the surface SWC jointly with the middle and/or deep ones are compared with its individual contribution.

3.3 Evaluation setup

To compare the relative differences in data-worth estimation accuracy under the various test scenarios, the mean absolute percentage error (MAPE) between the expected and reference data-worth in the form of T_r, *SED*, and *RE* is defined as:

$$MAPE = \frac{1}{T_t - T_p} \sum_{k=T_p+1}^{T_t} \left| \frac{DW_k^{Expect} - DW_k^{Refer}}{DW_k^{Refer}} \right| \tag{16}$$

where DW_k^{Expect} and DW_k^{Refer} denote the expected and reference DW values, respectively, at time step *t*=*k*.

4 Results and discussions

4.1 Optimal Monitoring Location for the Multiple Predictive Objectives (TC1/TC2/TC3)

Fig. 3 shows the probability distributions of the generated potential observation realizations as well as their ensemble mean and the corresponding actual observations of the surface (θ_S), middle (θ_M), and deep (θ_D) soil moisture at three sites. Only the results on the 81st, 90th, and 99th days are presented here. Overall, the $N_1 = 50$ potential realizations could "capture" the actual SWC observations with acceptable accuracy. Specifically, the forecasted middle SWC exhibited a considerably more robust 'capturing' performance with sustained better proximity of potential and actual θ_M throughout the simulation period. This occurred especially pronounced at the Cape site. For example, both surface and deep layers at Cape may be at risk of poor fit of potential observations to measurements (Fig. 3d and Fig. 3w), while the generated middle SWC is always fairly well approximated to the corresponding actual values in Fig. 3(m-o).

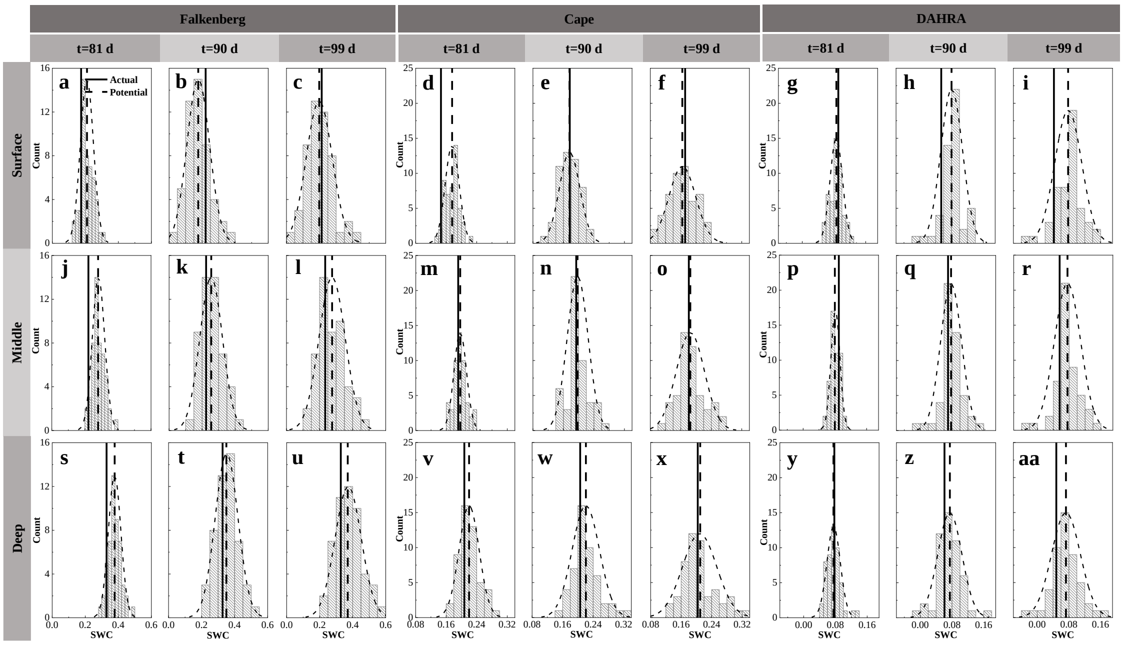


Figure 3. The probability distributions (dotted curved line) of potential observation realizations as well as their mean (dotted vertical line) and the corresponding actual soil water content (SWC) observation (solid line) in the surface, middle, and deep layers on the 81st, 90th, and 99th day at Falkenberg, Cape, and DAHRA, respectively

Based on the above potential observations, their expected data-worth regarding the retrieval of $\theta^{ave}_{1.00}$, $\theta^{ave}_{0.60}$, and $\theta^{ave}_{0.30}$ can be quantified in the form of T_r, *SED*, and *RE*, as depicted in Fig. 4. Meanwhile, for ease of analysis, Fig. 5 compares the covariance matrices of entire soil moisture profile in the prior stage, preposterior stage, and posterior stage. Only the results from the 81st day to 90th day at Falkenberg are revealed here. It can be observed that despite an overall increasing trend over time, the values of expected DW were prone to local spikes due to changes in the atmospheric boundary conditions such as rainfall. First of all, this general trend of increase should be attributed to the sequential augmentation of potential observations based on existing prior data, resulting in the cumulative values of DW over time. However, abrupt changes in external forcing, such as unexperienced rainfall events on the 88th day at the Falkenberg, could trigger temporal extrapolation of statistical learning (Li et al., 2020; Minns and Hall, 1996; Xu and Valocchi, 2015), which in turn led to a surge in prior predictive uncertainty, i.e., $\boldsymbol{C}_1 = Cov(\boldsymbol{Y}|\boldsymbol{A})$ (the 1st column of Fig. 5). Fortunately, joint GP training and sequential assimilation of real-time potential observations can effectively lower the risk of such irrational extrapolation (Wang et al., 2021a; Wang et al., 2021b), allowing these temporal mutations to be substantially attenuated at the preposterior stage [i.e., $C_2 = E_{B|A}Cov(Y|A,B)$]

(the 2nd-4th column of Fig. 5). This uncertainty reduction brought about the fusion of additional data became significantly larger when external forcing encountered mutations, which ultimately led to the localized surge in DW during rainfall events. We recall that such DW surges induced by contrasting scenarios also occurred in traditional DW analysis based on physically motivated models. As stated in Wang et al. (2018), "*If one model undergoes different and contrasting scenarios, the model structural error is likely to appear since model parameters updated or calibrated under one scenario have not been examined under another scenario.*" Although in different ways, the expected DW of future monitoring strategies within both the traditional parametric and proposed nonparametric DA frameworks heavily depends on the coverage of additional data by available prior scenarios (Li et al., 2020).

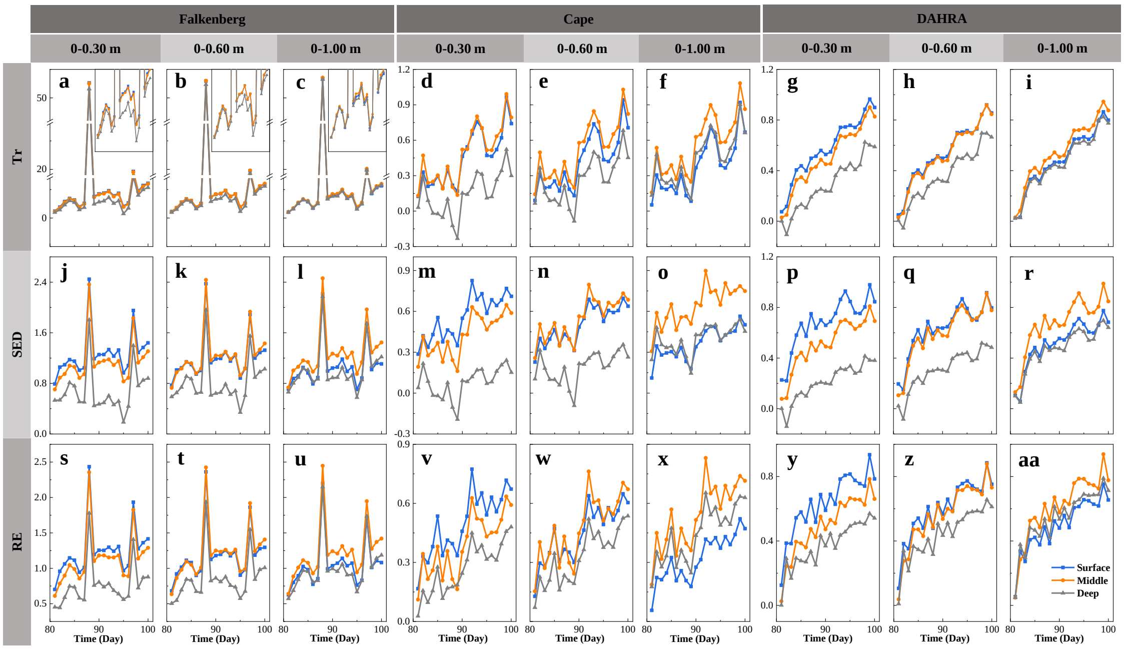


Figure 4. The expected data-worth of potential soil moisture observations in the surface, middle, and deep layers in the form of trace (T_r), Shannon entropy difference (*SED*), and relative entropy (*RE*), respectively, regarding the retrieval of average soil moisture in the top 0.30 m, 0.60 m, and 1.00 m at three sites

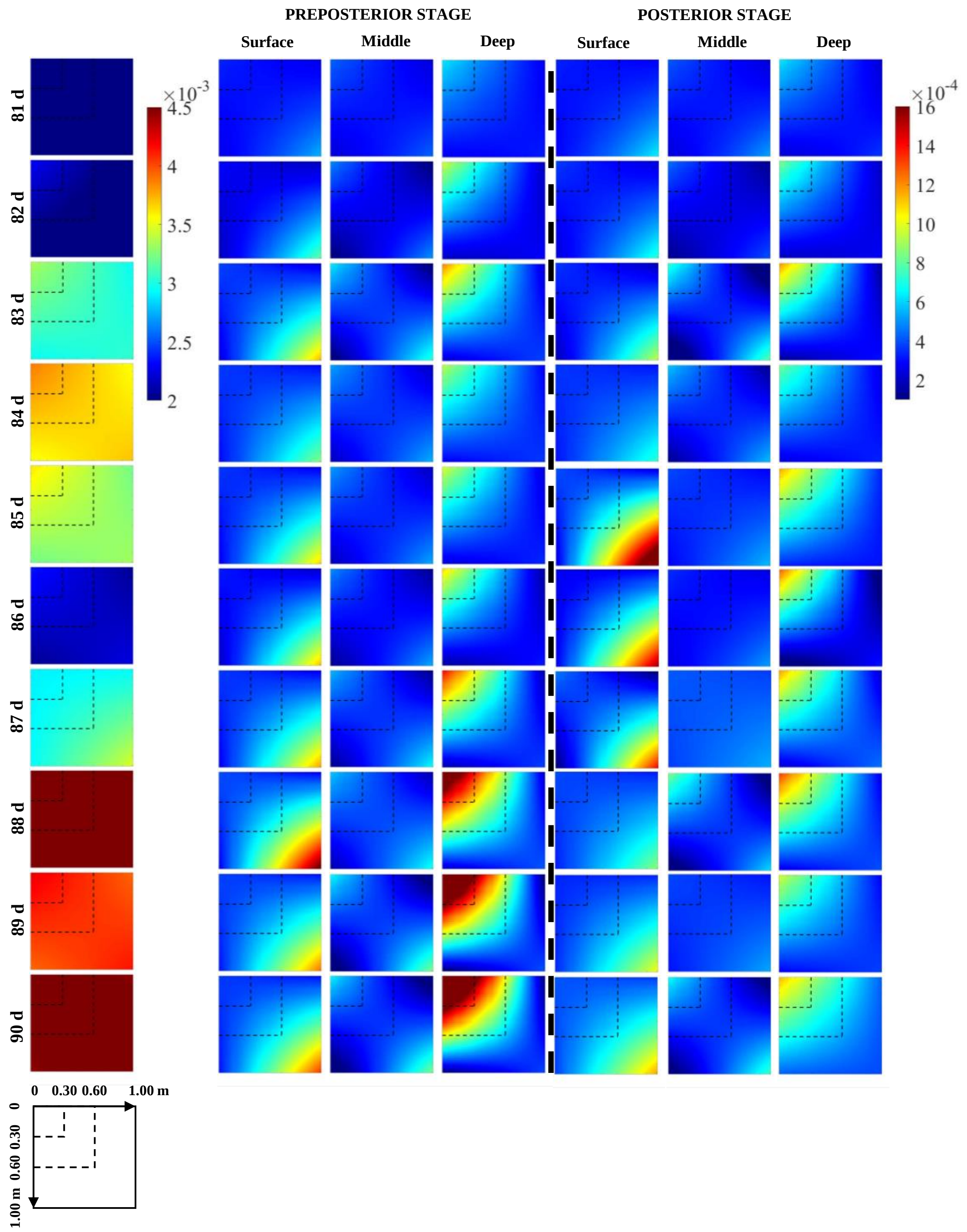


Figure 5. The covariance matrices of soil moisture profiles from the 81st to 90th day at Falkenberg before (column 1) and after potential (columns 2-4) and corresponding actual (columns 5-7) soil moisture observations in the surface, middle, and deep layers were fused, respectively

Moreover, Fig. 4 also suggests that the optimal observation depth shifted as the prediction target changed. As expected, the surface SWC θ_S produced higher T_r, *SED*, and *RE* values regarding the estimation of $\theta_{0.30}^{ave}$. As the depth range of the average SWC to be estimated was extended downward, the data-worth advantages of θ_M and θ_D began to emerge. Surprisingly, the potential middle SWC still exhibited a considerably higher superiority even in $\theta_{1.00}^{ave}$ estimation. In other words, the soil moisture in the middle layer has the most robust advantage in data-worth. This may be due to the fact that the integration of surface or deep SWC only reduced the uncertainty within the corresponding depth ranges (the 2nd and 4th column of Fig. 5), whereas the augmentation of θ_M significantly decrease the covariance matrices of the entire SWC profiles (the 3rd column of Fig. 5). This selection result of the optimal monitoring location seemingly contradicts previous findings within the traditional parametric DW analysis where the surface observations with the largest temporal variation always produced the greatest data worth, as reported in Dai et al. (2016) and Wang et al. (2018). This discrepancy is likely to depend on the different mechanisms that characterize soil moisture dynamics in the vertical direction between the two approaches. The traditional parametric unsaturated flow model follows the law of mass conservation-based physical governing equations (i.e., the Richardson–Richards equation, (Richards, 1931; Richardson, 1922)) as well as their physical properties to simulate the soil water infiltration process. The strongest time-varying nature of surface SWC was conducive to the effective updating of the physical parameters in EnKF, eventually generating the maximum data-worth (Wang et al., 2018). However, the spatial prediction performance of data-driven methods substantially hinged on the similarity of data between different depths. Theoretically, there occurs an inherent delayed response of soil moisture profiles to rainfall events, which has been well-documented experimentally (Wierenga et al., 1986; Bresler et al., 1971; Vauclin et al., 1979). This causes the temporal changes in surface and deep SWC to be naturally asynchronous, thus rendering their representativeness in characterizing the whole soil moisture profile somewhat limited. Ultimately, the complete reliance on statistical and information-theoretic measures allowed the most representative middle SWC to establish the most robust superiority in DW.

It can also be seen from Fig.4 that when using different information indices (i.e., T_r, *SED*, and *RE*) to quantify the data-worth, the optimal observation location selected is identical, regardless of soil textures and climatic regimes. This conclusion is generally in line with Wang et al. (2018) and Man et al. (2016). Furthermore, to quantify the data-worth assessment accuracy, Fig. 6 depicts the MAPE between the expected and reference data-worth in the form of T_r, *SED*, and *RE* of alternative monitoring schemes at different depths. It can be observed that the surface SWC yielded the smallest MAPE when retrieving $\theta_{0.30}^{ave}$, regardless of the metric type. During the estimation of $\theta_{0.60}^{ave}$ and $\theta_{1.00}^{ave}$, nevertheless, the expected data-worth of θ_M more accurately and robustly approached the reference counterparts with overall smaller MAPEs. We recall that this ranking of DW estimation accuracy was exactly in line with the ranking of the magnitude of their expected DW in Fig. 4. To be specific, a comparison of Fig. 4 and Fig. 6 reveals that potential observations with a larger expected DW are prone to a higher DW estimation accuracy due to its more robust ability of "imitating" the actual observations (Fig. 3).

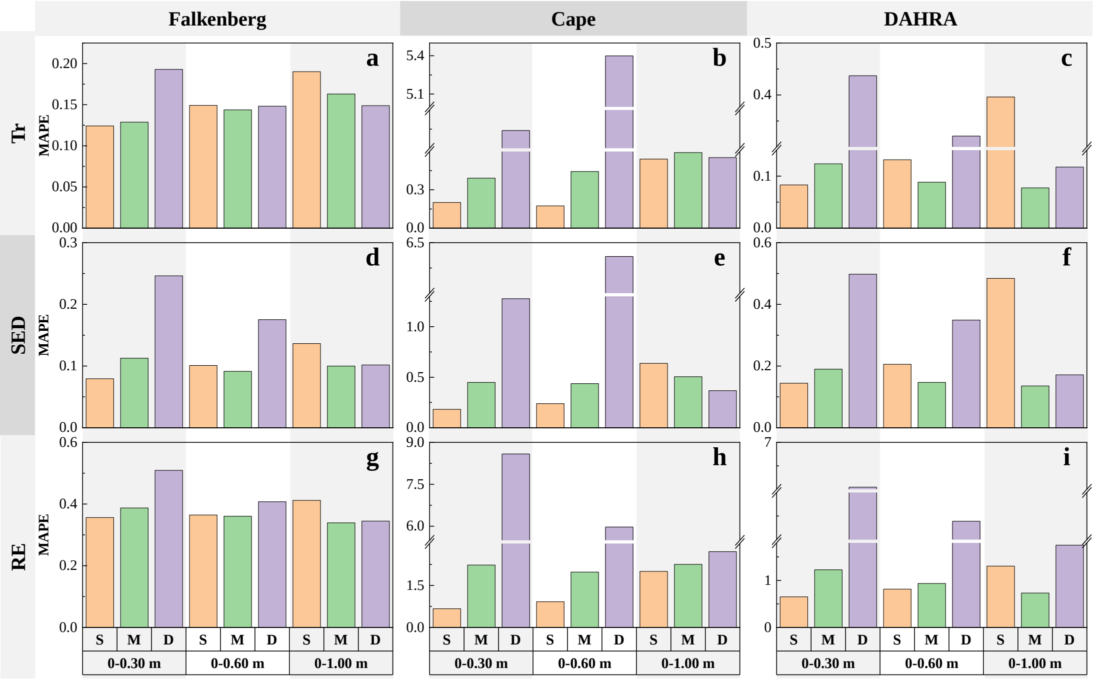


Figure 6. The MAPEs between expected and reference data-worth in the form of T_r, *SED*, and *RE* of potential soil moisture observations in the surface (S), middle (M), and deep (D) layers, respectively, regarding the retrieval of average soil moisture in the top 0.30 m, 0.60 m, and 1.00 m at three sites

4.2 Effects of Observation Noise (TC1-1/TC4/TC5)

Fig. 7 shows the probability distributions of the potential observation ensemble as well as their mean and the corresponding actual observations of the surface SWC under different SWC noise levels. Similarly, only the results on the 81st, 90th, and 99th days are displayed. It can be observed that a higher noise level was not always detrimental but rather instead expanded the distribution width along the SWC-axis and produced a flatter curve. The risk of failure of the generated realizations to "capture" the real observations was thus reduced. Even on the 81st day at Falkenberg, for example, the increase in SWC error variance from 0.01^2 to 0.04^2 facilitated a better agreement between the potential and actual surface soil moisture, as revealed in Fig. 7 (a j s). Similar phenomena can also be found via a comparison of Fig. 7e and Fig. 7n.

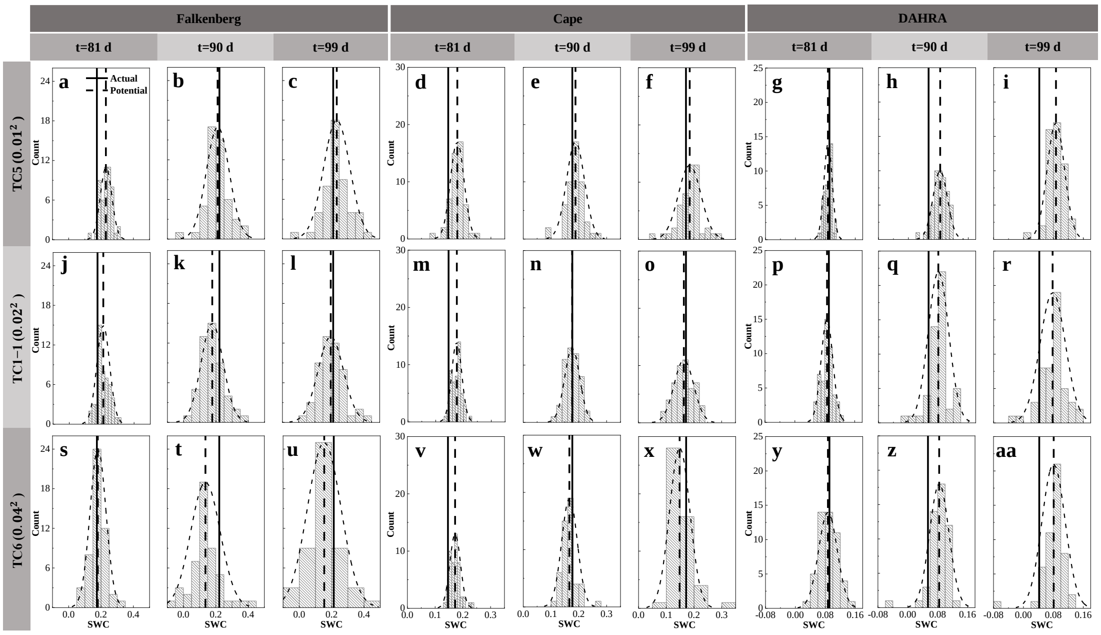


Figure 7. The probability distributions (dotted curved line) of potential observation realizations as well as their mean (dotted vertical line) and the corresponding actual soil water content (SWC) observations (solid line) in the surface layer on the 81st, 90th, and 99th day at three sites under different measurement error variances being 0.01^2 (TC4), 0.02^2 (TC1-1), and 0.04^2 (TC5), respectively

Fig. 8 shows the temporal evolution and time-averaged MAPE of the expected and reference data-worth in the form of three information indices, respectively, under various noise levels. Some interesting findings can be obtained: (1) Overall, the potential SWC data corrupted by a lower noise level yielded larger data-worth with higher accuracy. (2) Nevertheless, the occurrence of rainfall events triggered a futile DW increase while also rendering the potential observations with appropriately magnified observation errors more valuable. For instance, a properly inflated observation error of 0.02^2 on the 88th day at the Falkenberg site resulted in a notably higher data-worth than that of 0.01^2, as highlighted by the dashed ellipse boxes in Fig. 8a and Fig. 8d. Furthermore, this increase in data-worth resulting from noise amplification was particularly evident in the form of T_r over the other two metrics, as depicted in Fig. 8(a-c) and Fig. 8(d-i). At DAHRA, potential observations with an observation error of 0.02^2 even produced a significantly higher T_r value than that of 0.01^2 throughout almost the entire simulation period (Fig. 8c). (3) As opposed to T_r and *SED* indices focusing only on the system uncertainty (variance or covariance), the expected *RE*, as a comprehensive mean-covariance-type metric, was often more challenging to approach its reference counterparts with the largest MAPE at all sites, as shown in Fig. 8(j-l).

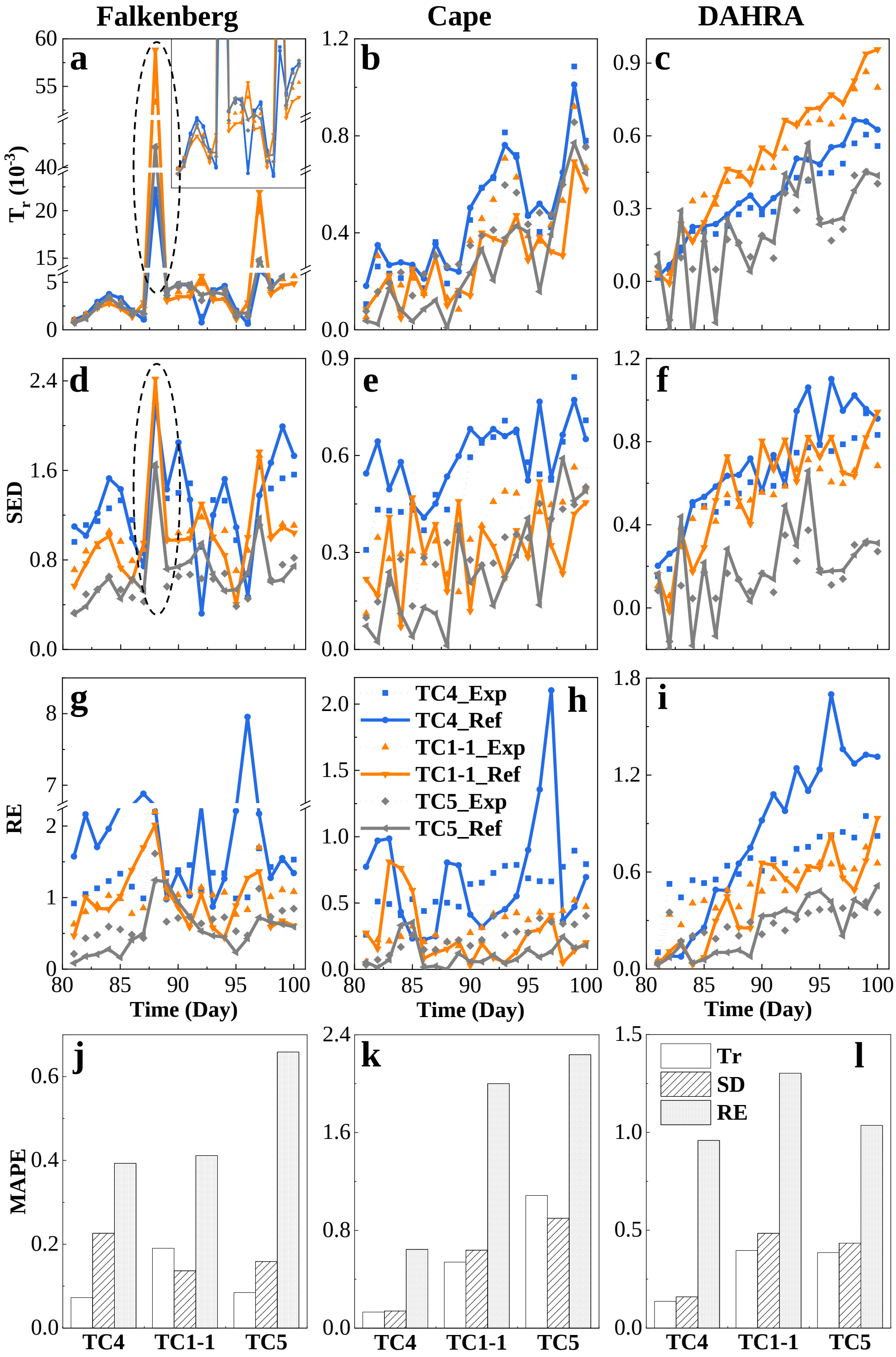
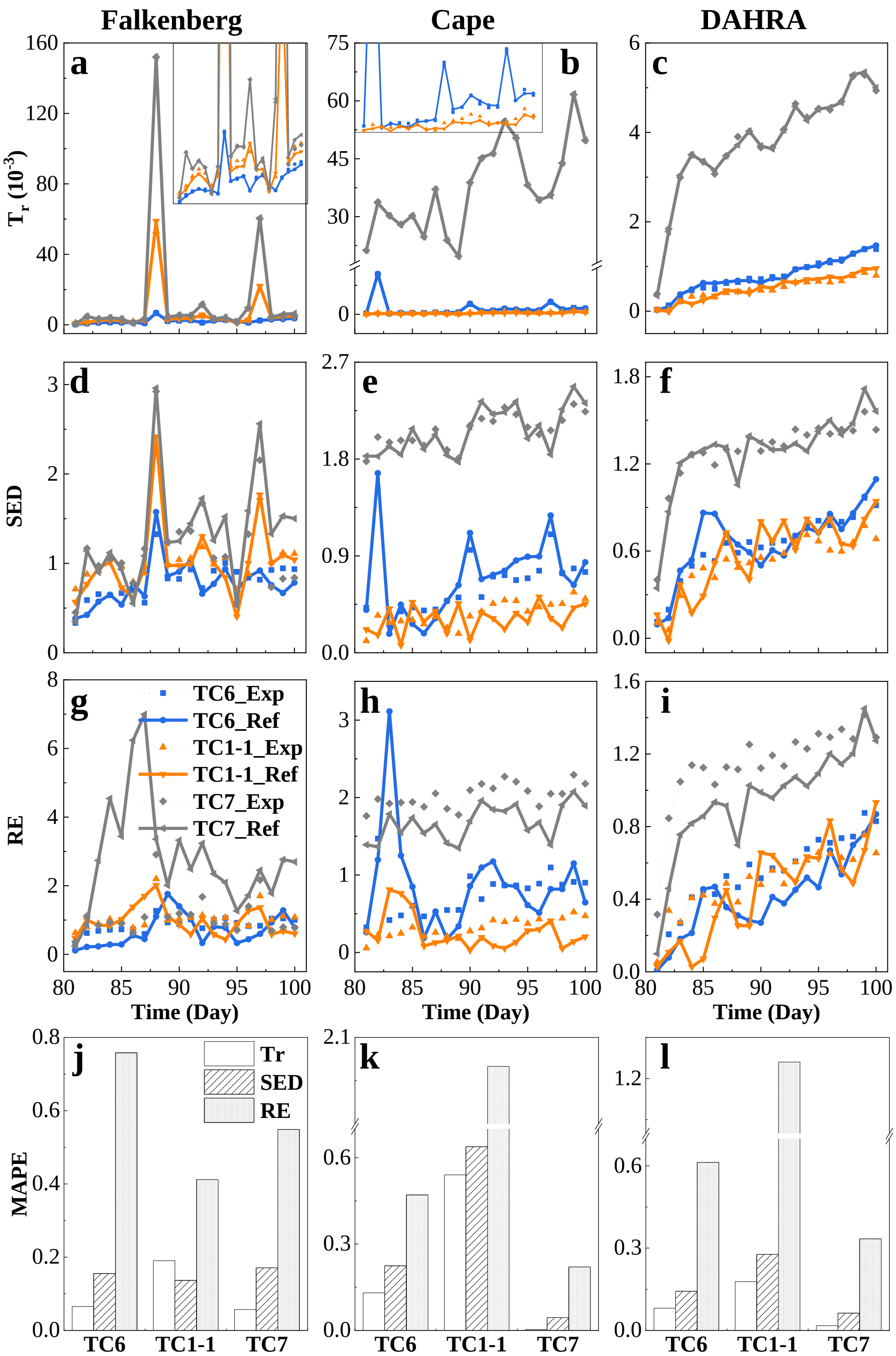


Figure 8. The temporal evolution (a-i) and time-averaged MAPEs (j-l) of the expected and reference data-worth in the form of T_r, *SED*, and *RE* at three sites, respectively, under different measurement error variances being 0.01^2 (TC4), 0.02^2 (TC1-1), and 0.04^2 (TC5), respectively

4.3 Effects of Prior Data Content (TC1-1/TC6/TC7)

Fig. 9(a-i) depicts the temporal evolution of the expected and reference data-worth of the surface SWC in cases TC6, TC1-1, and TC7 with the 40-day, 80-day, and 180-day prior data content, respectively. Under normal circumstances, an increase in the available prior data content inevitably entails a shrinkage in the DW of subsequent data due to the possibility of information redundancy. However, this seems to be valid only for a modest increase in prior data (from 40-day to 80-day) within our NP-DWA framework. The substantial augmentation in available data content from 80-day to 180-day instead resulted in a notably higher DW of the additional data (Fig. 9). Even more unexpectedly, this DW growth was prevalent across sites, regardless of the soil types and climatic regimes. To clarify this anomaly, Fig. 10 further shows the predicted covariance matrices of soil moisture profiles conditional on $\{\boldsymbol{A}\}$ in prior stage and $\{\boldsymbol{A}, \boldsymbol{B}\}$ in preposterior stage in cases TC6, TC1-1, and TC7, respectively. Only the results from the 81st day to 90th day at Falkenberg are presented here. Our previous studies have demonstrated that although the mean values of potential samples can approach actual observations well in fully (Wang et al., 2021a) or partially (Zhang et al., 2019) data-driven dynamical systems, their ensemble was apt to suffer from considerable uncertainty (Wang et al., 2021b). Unfortunately, augmented prior data, despite its potential to enrich available GP training scenarios, failed to prevent the non-convergence of $N_e = 50$ GP samples. In contrast, the additional noise associated with prior data supplementation could exacerbate the increase in the prior prediction uncertainty (i.e., $\boldsymbol{C}_1$), as illustrated by a comparison between the first three columns of Fig. 10. It should be highlighted that the fusion of $\boldsymbol{B}$ enabled a notable reduction in the preposterior uncertainties (i.e., $\boldsymbol{C}_2$) in the data assimilation system to a comparable level (the last three columns of Fig. 10), even with different prior data content. The gradual widening of the gap between $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ eventually yielded the highest data-worth with the maximum amount of prior data in test case TC7 (Fig. 9). This seems to alarm us that uncontrolled expansion of big data within fully data-driven systems may not be beneficial. The adverse effects of extra noise may overshadow its original superiority in generalization capability. Access to high-quality and representative “small” data may constitute the key to the successful application of fully data-driven algorithms for reshaping soil moisture dynamics.



520 **Figure 9.** The temporal evolution (a-i) and time-averaged MAPEs (j-l) of the expected and reference data-worth in the form of T_r, *SED*, and *RE* at three sites for cases TC6, TC1-1, and TC7 with 40-day, 80-day, and 180-day prior data content, respectively

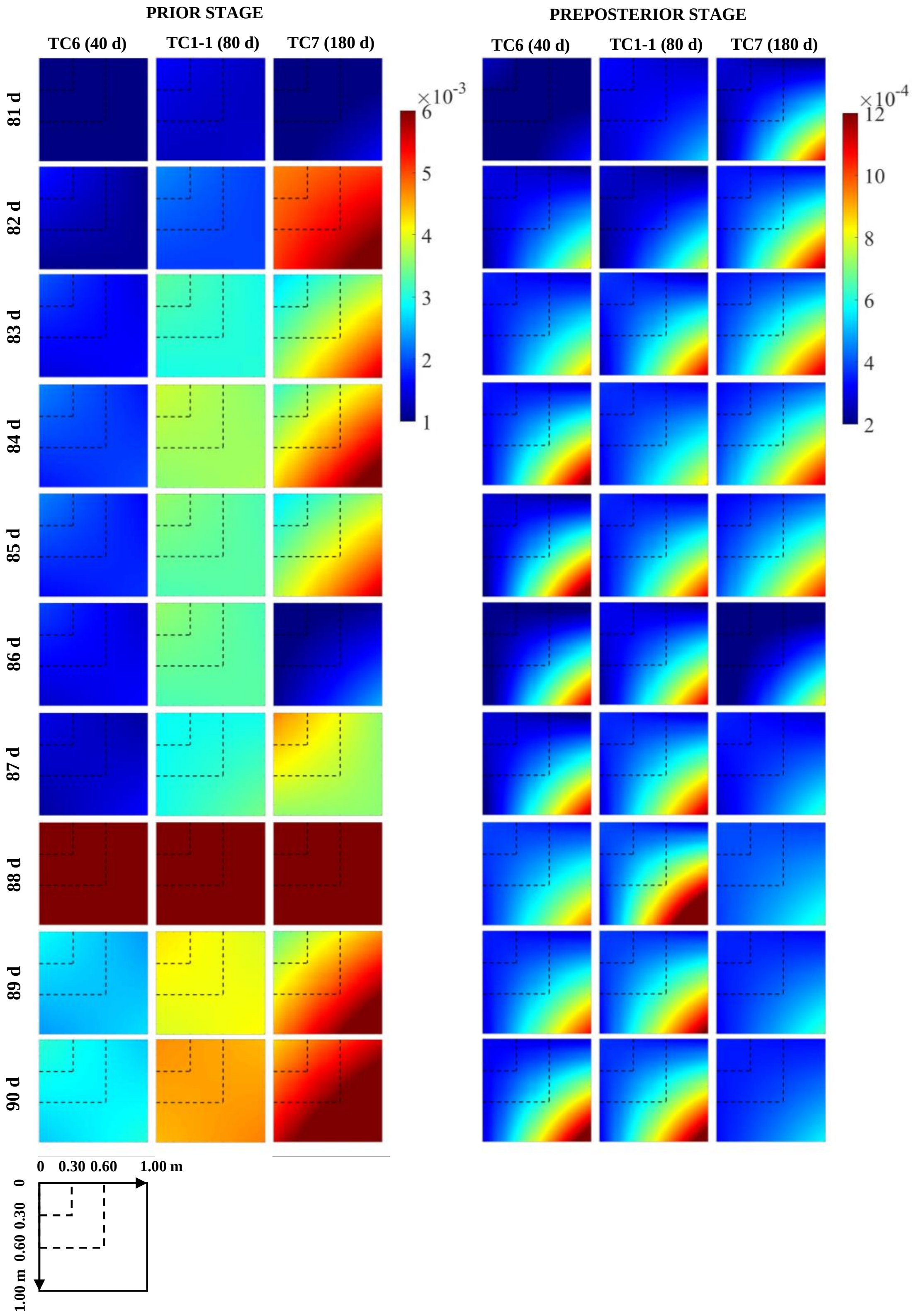


Figure 10. The covariance matrices of soil moisture profiles from the 81st day to 90th day at Falkenberg in the prior and preposterior stage for cases TC6, TC1-1, and TC7 with 40-day, 80-day, and 180-day prior data content, respectively

Furthermore, Fig. 9(j-l) depicts the time-averaged MAPE in the expected and reference data-worth in cases TC6, TC1-1, and TC7, respectively. A comparison of Fig. 9(a-i) and Fig. 9(j-l) reveals some interesting findings: (1) similar to the results in Sect. 4.1, the potential measurements with the largest expected (or reference) data-worth in TC7 are apt to possess the highest estimation accuracy of data-worth. (2) Local variations in data-worth at different sites respond slightly differently to the augmentation of prior data content. For instance, even with 180 days of available historical data, the DW spike induced by the unexperienced rainfall event on the 88th day at the Falkenberg has not been eliminated or diminished (Fig. 9(a d g)). However, similar DW surges on the 82nd day at Cape were successfully mitigated as the amount of prior data content increased from 40-day (TC6) to 80-day (TC1-1) (Fig. 9(e h)). This is because the prior data at Falkenberg, even if augmented to 180-day, did not cover the rainfall event on the 88th day (Fig. 2b), whereas the 80-day training data at Cape already included the scenario on the 82nd (Fig. 2d). These results agree with the conclusions reported in Wang et al. (2020) that the diversity of scenarios in the training data is more decisive than the data volume regarding the performance of data-driven methods. However, it is worth noting that the response of our nonparametric framework to the prior data augmentation is not in line with with that of the physical model-based DW analysis framework Wang et al. (2018). In the latter, insufficient prior scenarios within the traditional framework may trigger unresolved model structural errors whose resultant deterioration in the DW assessment performance cannot be compensated by assimilating more prior data. On the contrary, the elimination of dependence on physical governing equations shields our model-free DA schemes from model structural errors. Continuous enrichment of prior scenarios in the NP-DWA directly assures effective coverage of potential future scenarios, yielding the mitigation of DW local surges and an improved DW estimation accuracy. (3) Although inferior to T_r and *SED*, the estimation accuracy of *RE* is generally acceptable, especially when prior data is expanded to 180 days. This is certainly a remarkable improvement over the rather poor performance of *RE* in traditional parametric data-worth analysis (Wang et al., 2018; Wang et al., 2020). This progress should be attributed to the radical abandonment of physical models in the NP-DWA, which prevented adverse effects of the high nonlinearity of soil water flow in the propagation of uncertainties from input to output (i.e., soil moisture in this study). Direct mapping from regular meteorological data to SWC facilitated the identification of the soil moisture covariance matrix from potential observations.

4.4 Effects of potential observational combinations (TC1-1/TC8/TC9)

Fig. 11(a-i) compares the expected (and reference) data-worth of three combinations of potential observations at different depths at the three sites. It can be seen that the composite data-worth of the alternative monitoring schemes exhibited an increasing pattern as the depth range of the observed SWC continues to expand downward. Nevertheless, the response of the different data-worth indicators and study sites to this vertical expansion of potential observations varied slightly. Further integration of θ_D in TC9 did not cause a marked increase in T_r but yielded notably greater *SED* and *RE* values, especially at the DAHRA site (Fig. 11(c f d)). This is undoubtedly due to the extra consideration of the latter two indicators for the non-diagonal elements of the covariance matrix or/and the behavior of the mean. Moreover, the joint fusion of potential θ_S and θ_M failed to result in a sustained increase in T_r and *RE* at DAHRA, while creating a significant increase in composite DW at the other two sites. This could be attributed to the sandy soil texture at DAHRA (with the fraction of sand up to 90%, and K_s= 3.22 m/d), resulting in the almost synchronous responses of the SWC at *z*=0.05 m and 0.50 m to the atmospheric boundary conditions (Fig. 2e) and thus triggering possible data redundancy.

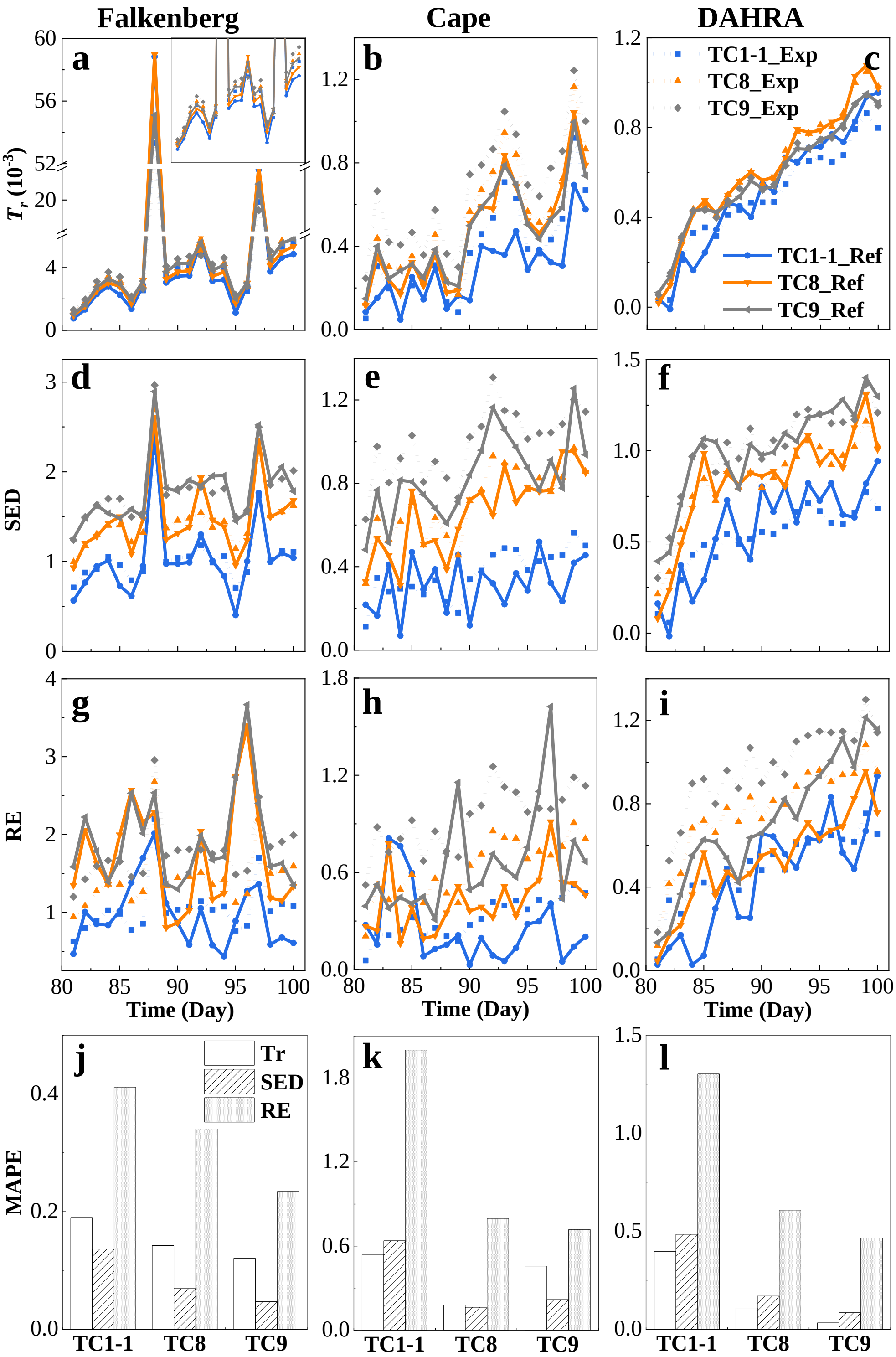


Figure 11. The temporal evolution (a-i) and time-averaged MAPEs (j-l) of the expected and reference data-worth in the form of T_r, *SED*, and *RE* at three sites for cases TC1-1 (surface soil moisture), TC8 (surface & middle soil moisture), and TC9 (surface & middle & deep soil moisture), respectively

575

Fig. 11(j-l) further shows the estimation accuracy of expected data-worth for the above three potential observation combinations. Surprisingly, the increase in the number of potential observations, while making it more difficult to “capture” actual SWC data, ends up significantly improving the accuracy of the data-worth assessment. As shown in Fig. 11(j-l), as more potential observations along the vertical direction were evaluated, the MAPEs between expected data-worth and its reference counterparts decreased continuously. This phenomenon actually breaks the misconceptions about the data-worth assessment accuracy in previous studies, i.e., that an excellent fit of potential observations is equivalent to high-precision estimates of the corresponding data-worth. For the sake of explanation, Fig. 12 shows the predicted covariance matrices of soil moisture profiles in cases TC1-1, TC8, and TC9 from the 81st day to 90th day at Falkenberg conditional on $\{\boldsymbol{A}\}$, $\{\boldsymbol{A},\boldsymbol{B}\}$, and $\{\boldsymbol{A},\boldsymbol{B}'\}$, respectively. It can be found that compared to TC1-1, which only reduces the uncertainties in the surface SWC, the integration of observations at multiple depths clearly reduces the uncertainties in the entire SWC profiles to a considerably lower level. This ultimately facilitates better proximity between expected and reference covariance matrices, as revealed in the 4th and 7th columns in Fig. 12. The above results suggest that the accuracy of data-worth assessment of potential observations does not only depend on their capacity to “capture” actual measurements, but is also closely related to their correlation with the variable of interest. We recall that similar phenomena also exist in the preceding test cases. For example, the weaker correlation between surface SWC observations and $\theta_{1.00}^{ave}$ led to deterioration in the DW estimation performance with the largest MAPE values (Fig.6(a d g)) even if the actual surface observations could be suitably reproduced (Fig. 3(a-c)). Therefore, to enhance the reliability of data-worth assessment, a strategy wherein potential observations at multiple depths were simultaneously incorporated into existing DA systems was recommended in this study.

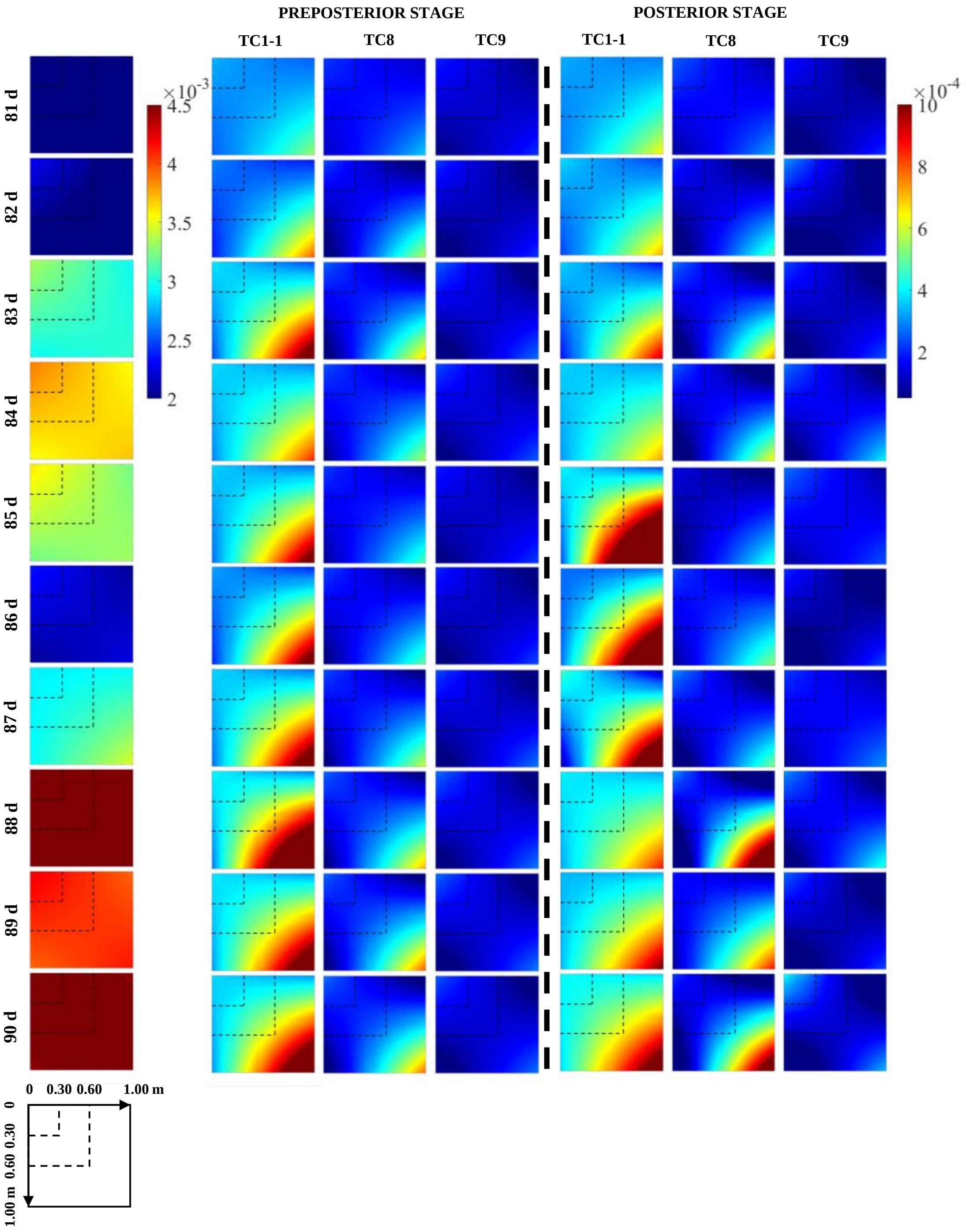


Figure 12. The covariance matrices of soil moisture profiles from the 81st day to 90th day at Falkenberg before (column 1) and after potential (columns 2-4) and actual observations (columns 5-7) for cases TC1-1 (surface soil moisture), TC8 (surface & middle soil moisture), and TC9 (surface & middle & deep soil moisture) were fused, respectively

5 Conclusions

Conventional data-worth analysis for soil water problems depends on physical dynamic models.

Due to the widespread occurrence of model structural errors, it may lead to biased or wrong worth assessment. The strong nonlinearity of unsaturated flow further deteriorates the DW assessment performance in the retrieval of soil moisture profiles. This study proposed a nonparametric data-worth analysis method within a fully data-driven modeling framework. The information extracted from real-time soil moisture data after GP training and Kalman update was quantified with three representative types of indicators, i.e., variance- (T_r), covariance- (*SED*), and mean-covariance-type (*RE*) indicators. With the aid of a series of real-world cases, the ability and challenge of the NP-DWA in terms of the variables of interest, spatial location, observation error, and prior data content were assessed. The following conclusions were drawn:

(1) The proposed NP-DWA framework enabled an accurate assessment of the data-worth of potential observations regarding the reconstruction of purely data-driven soil water flow models prior to data collection. Similar to traditional DW analysis based on physical models, the overall increasing trend of the DW from the sequential augmentation of additional observations within the proposed NP-DWA framework was also susceptible to interruptions by localized surges due to never-experienced atmospheric conditions. The difference is that this biased DW in the traditional parametric method is caused by model structural errors triggered by contrasting scenarios, which is difficult to be compensated by assimilating more prior data, while the adverse effects of anomalous GP extrapolation in our NP-DWA could be suitably avoided by the enrichment of training scenarios in prior data. Moreover, the appropriate amplification of observational noise under extreme meteorological conditions also facilitated the alleviation of these biased estimates by enhancing the generalization capacity of dynamic models.

(2) The optimal observation depth shifted as the prediction target varied. In contrast to the notably higher DW of surface SWC observations within the conventional DW analysis framework based on physical models, middle SWC observations tended to exhibit considerably higher robustness in the construction of model-free soil moisture dynamic models. This should be attributed to the ability of the SWC in the middle layer to effectively reduce the predictive uncertainty of the entire soil moisture profiles due to its optimal representativeness. The inherent delayed response of soil moisture profiles to rainfall events allowed this advantage of middle SWC prevalent across sites, even becoming increasingly pronounced with increasing delay effect.

(3) Although the addition of prior data content could greatly improve the estimation accuracy of

the expected DW, the ensuing observation noise could substantially increase the uncertainty in a purely data-driven DA system, leading to potentially higher data-worth of subsequent observations. Hence, high-quality and representative small data may be regarded as a better choice than unfiltered big data.

(4) The performance of data-worth assessment was jointly determined by ‘3Cs’, i.e., capacity of potential observation realizations to “capture” actual observations, correlation of potential observations with the predicted variables of interest, and choice of DW quantitative indicators. Furthermore, the direct mapping from regular meteorological data to SWC in our nonparametric method facilitated the identification of the soil moisture covariance matrix (especially the cross-covariance) due to its alleviation of high nonlinearity of soil water flow problems. Hence, satisfactory estimation accuracy could also be achieved even with covariance-related data-worth metrics (i.e., the *SED* and *RE*).

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China Grants U2243235 and 51979200, and the Open Research Fund of Guangxi Key Laboratory of Water Engineering Materials and Structures Grant GXHRI-WEMS-2020-06.

CODE/DATA AVAILABILITY

The code/data that support the findings of this study are available from the corresponding author upon reasonable request.

AUTHOR CONTRIBUTION

Yakun Wang: Conceptualization, Methodology, Software, Writing–original draft. **Xiaolong Hu**: Conceptualization, Software. **Lijun Wang**: Methodology. **Jinmin Li**: Data curation, Methodology. **Lin Lin**: Supervision. **Kai Huang**: Data curation. **Liangsheng Shi**: Writing – review & editing, Supervision.

COMPETING INTERESTS

The contact author has declared that none of the authors has any competing interests.

REFERENCES:

Akhtar, K. et al., 2019. Wheat straw mulching offset soil moisture deficient for improving physiological and growth performance of summer sown soybean. Agricultural water management, 211: 16-25.

Al-Akhras, M., El Hindi, K., Habib, M. and Shawar, B.A., 2021. Instance reduction for avoiding overfitting in decision trees. Journal of Intelligent Systems, 30(1): 438-459.

Brajard, J., Carrassi, A., Bocquet, M. and Bertino, L., 2020. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. Journal of Computational Science, 44: 101171.

Brajard, J., Carrassi, A., Bocquet, M. and Bertino, L., 2021. Combining data assimilation and machine learning to infer unresolved scale parametrization. Philosophical Transactions of the Royal Society A, 379(2194): 20200086.

Bresler, E. et al., 1971. Infiltration from a trickle source: II. Experimental data and theoretical predictions. Soil Science Society of America Journal, 35(5): 683-689.

Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1): 16-28.

Dai, C., Xue, L., Zhang, D. and Guadagnini, A., 2016. Data-worth analysis through probabilistic collocation-based Ensemble Kalman Filter. Journal of Hydrology, 540: 488-503.

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

De Lannoy, G.J., Verhoest, N.E., Houser, P.R., Gish, T.J. and Van Meirvenne, M., 2006. Spatial and temporal characteristics of soil moisture in an intensively monitored agricultural field (OPE3). Journal of Hydrology, 331(3-4): 719-730.

Dobriyal, P., Qureshi, A., Badola, R. and Hussain, S.A., 2012. A review of the methods available for estimating soil moisture and its implications for water resource management. Journal of Hydrology, 458: 110-117.

Dunne, S. and Entekhabi, D., 2005. An ensemble ‐ based reanalysis approach to land data assimilation. Water resources research, 41(2).

Evensen, G., 2003. The ensemble Kalman filter: Theoretical formulation and practical implementation. Ocean dynamics, 53(4): 343-367.

Fienen, M.N., Doherty, J.E., Hunt, R.J. and Reeves, H.W., 2010. Using prediction uncertainty analysis to design hydrologic monitoring networks: example applications from the Great Lakes water availability pilot project. U. S. Geological Survey.

Finsterle, S., 2015. Practical notes on local data‐worth analysis. Water Resources Research, 51(12): 9904-9924.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1): 1-22.

García-Gil, D., Luengo, J., García, S. and Herrera, F., 2019. Enabling smart data: noise filtering in big data classification. Information Sciences, 479: 135-152.

Gu, H., Lin, Z., Guo, W. and Deb, S., 2021. Retrieving surface soil water content using a soil texture adjusted vegetation index and unmanned aerial system images. Remote Sensing, 13(1): 145.

Hall, M.A., 1999. Correlation-based feature selection for machine learning. The University of Waikato.

Hamilton, F., Berry, T. and Sauer, T., 2017. Kalman-Takens filtering in the presence of dynamical noise. The European Physical Journal Special Topics, 226(15): 3239-3250.

Hill, M.C. and Tiedeman, C.R., 2006. Effective groundwater model calibration: with analysis of data, sensitivities, predictions, and uncertainty. John Wiley & Sons.

Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. IEEE transactions on information theory, 14(1): 55-63.

Ju, L., Zhang, J., Meng, L., Wu, L. and Zeng, L., 2018. An adaptive Gaussian process-based iterative ensemble smoother for data assimilation. Advances in water resources, 115: 125-135.

Kashif Gill, M., Kemblowski, M.W. and McKee, M., 2007. Soil moisture data assimilation using support vector machines and ensemble Kalman filter 1. JAWRA Journal of the American Water Resources Association, 43(4): 1004-1015.

Kisekka, I., Migliaccio, K.W., Muñoz Carpena, R., Schaffer, B. and Khare, Y., 2015. Modelling soil water dynamics considering measurement uncertainty. Hydrological Processes, 29(5): 692-711.

Leube, P.C., Geiges, A. and Nowak, W., 2012. Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design. Water Resources Research, 48(2).

Li, C. and Ren, L., 2011. Estimation of unsaturated soil hydraulic parameters using the ensemble Kalman filter. Vadose Zone Journal, 10(4): 1205-1227.

Li, P. et al., 2020. Comparison of the use of a physical-based model with data assimilation and machine learning methods for simulating soil water dynamics. Journal of Hydrology, 584: 124692.

Li, X., Shi, L., Zha, Y., Wang, Y. and 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(9): 2477-2493.

Liu, H.L. et al., 2011. Simulating water content, crop yield and nitrate-N loss under free and controlled tile drainage with subsurface irrigation using the DSSAT model. Agricultural Water Management, 98(6): 1105-1111.

Liu, K. et al., 2020. A gaussian process-based iterative Ensemble Kalman Filter for parameter estimation of unsaturated flow. Journal of Hydrology, 589: 125210.

Man, J., Zhang, J., Li, W., Zeng, L. and Wu, L., 2016. Sequential ensemble‐based optimal design for parameter estimation. Water Resources Research, 52(10): 7577-7592.

Minns, A.W. and Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrological sciences journal, 41(3): 399-417.

Montzka, C. et al., 2011. Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter. Journal of hydrology, 399(3-4): 410-421.

Neuman, S.P., Xue, L., Ye, M. and Lu, D., 2012. Bayesian analysis of data-worth considering model and parameter uncertainties. Advances in Water Resources, 36: 75-85.

Nowak, W., Rubin, Y. and de Barros, F.P., 2012. A hypothesis‐driven approach to optimize field campaigns. Water Resources Research, 48(6).

Olvera-López, J.A., Carrasco-Ochoa, J.A., Martínez-Trinidad, J. and Kittler, J., 2010. A review of instance selection methods. Artificial Intelligence Review, 34(2): 133-143.

Pechenizkiy, M., Tsymbal, A., Puuronen, S. and Pechenizkiy, O., 2006. Class noise and supervised learning in medical domains: The effect of feature extraction. IEEE, pp. 708-713.

Rasmussen, C.E., 2003. Gaussian processes in machine learning. Springer, pp. 63-71.

Reichle, R.H., Crow, W.T. and Keppenne, C.L., 2008. An adaptive ensemble Kalman filter for soil moisture data assimilation. Water resources research, 44(3).

Richards, L.A., 1931. Capillary conduction of liquids through porous mediums. Physics, 1(5): 318-333.

Richardson, L.F., 1922. Weather prediction by numerical process. University Press.

Ross, P.J., 2003. Modeling soil water and solute transport—Fast, simplified numerical solutions. Agronomy journal, 95(6): 1352-1361.

Shannon, C.E., 1949. Communication in the presence of noise. Proceedings of the IRE, 37(1): 10-21.

Shi, C., Xie, Z., Qian, H., Liang, M. and Yang, X., 2011. China land soil moisture EnKF data assimilation based on satellite remote sensing data. Science China Earth Sciences, 54(9): 1430-1440.

Shuwen, Z., Haorui, L., Weidong, Z., Chongjian, Q. and Xin, L.I., 2005. Estimating the soil moisture profile by assimilating near-surface observations with the ensemble Kaiman filter (EnKF). Advances in Atmospheric Sciences, 22(6): 936-945.

Šimůnek, J., Van Genuchten, M.T. and Šejna, M., 2006. The HYDRUS software package for simulating two-and three-dimensional movement of water, heat, and multiple solutes in variably-saturated media. Technical manual, version, 1: 241.

Singh, K., Sandu, A., Jardak, M., Bowman, K.W. and Lee, M., 2013. A practical method to estimate information content in the context of 4D-Var data assimilation. SIAM/ASA Journal on Uncertainty Quantification, 1(1): 106-138.

Song, X., Shi, L., Ye, M., Yang, J. and Navon, I.M., 2014. Numerical comparison of iterative ensemble Kalman filters for unsaturated flow inverse modeling. Vadose Zone Journal, 13(2): 1-12.

Van Dam, J.C. and Feddes, R.A., 2000. Numerical simulation of infiltration, evaporation and shallow groundwater levels with the Richards equation. Journal of Hydrology, 233(1-4): 72-85.

Vauclin, M., Khanji, D. and Vachaud, G., 1979. Experimental and numerical study of a transient, two ‐ dimensional unsaturated ‐ saturated water table recharge problem. Water Resources Research, 15(5): 1089-1101.

Wang, Y. et al., 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.

Wang, Y. et al., 2020. A robust data ‐ worth analysis framework for soil moisture flow by hybridizing sequential data assimilation and machine learning. Vadose Zone Journal, 19(1): e20026.

Wang, Y. et al., 2021. A nonparametric sequential data assimilation scheme for soil moisture flow. Journal of Hydrology, 593: 125865.

Wang, Y., Shi, L., Zhang, Q. and Qiao, H., 2021. A gradient-enhanced sequential nonparametric data assimilation framework for soil moisture flow. Journal of Hydrology, 603: 126857.

Wierenga, P.J., Gelhar, L.W., Simmons, C.S., Gee, G.W. and Nicholson, T.J., 1986. Validation of stochastic flow and transport models for unsaturated soils: A comprehensive field study. Pacific Northwest Lab., Richland, WA (USA); New Mexico State Univ., Las ….

Williams, C.K. and Rasmussen, C.E., 2006. Gaussian processes for machine learning, 2. MIT press Cambridge, MA.

Xu, Q., 2007. Measuring information content from observations for data assimilation: Relative entropy versus Shannon entropy difference. Tellus A: Dynamic Meteorology and Oceanography, 59(2): 198-209.

Xu, T. and Valocchi, A.J., 2015. Data-driven methods to improve baseflow prediction of a regional groundwater model. Computers & Geosciences, 85: 124-136.

Yamanaka, A., Maeda, Y. and Sasaki, K., 2019. Ensemble Kalman filter-based data assimilation for three-dimensional multi-phase-field model: Estimation of anisotropic grain boundary properties. Materials & Design, 165: 107577.

Yang, J., Li, B. and Shiping, L., 2000. A large weighing lysimeter for evapotranspiration and soil‐water‐groundwater exchange studies. Hydrological processes, 14(10): 1887-1897.

Yeh, T.C.J., Gelhar, L.W. and Gutjahr, A.L., 1985. Stochastic analysis of unsaturated flow in heterogeneous soils: 1. Statistically isotropic media. Water Resources Research, 21(4): 447-456.

Zha, Y., Shi, L., Ye, M. and Yang, J., 2013. A generalized Ross method for two-and three-dimensional variably saturated flow. Advances in Water Resources, 54: 67-77.

Zhang, J., Zeng, L., Chen, C., Chen, D. and Wu, L., 2015. Efficient B ayesian experimental design for contaminant source identification. Water Resources Research, 51(1): 576-598.

Zhang, Q. et al., 2019. A dynamic data-driven method for dealing with model structural error in soil moisture data assimilation. Advances in Water Resources, 132: 103407.

Zhu, X. and Wu, X., 2004. Class noise vs. attribute noise: A quantitative study. Artificial intelligence review, 22(3): 177-210.