

Anonymous referee 1

General comments: This study proposes a comprehensive data-driven framework for selecting the optimal observing operations (data-worth analysis) and updating the predictions for soil moisture dynamics. The fully data-driven approach provides a complement to physics-based models, especially for complex real-world scenarios. While the quality of the manuscript is good, there are still some issues that require clarification.

Specific comments:

1. A major concern is the conclusions drawn from applying the Gaussian processes and EnKF assimilation techniques. While efficient and simple to implement, these methods have inherent limitations such as excessively smooth predictions (GP) and optimality only for Gaussian linear problems (EnKF). As the soil moisture dynamics are not fully met by these assumptions, the proposed method may experience difficulties, such as the mentioned localized surges. Therefore, some conclusions "high-quality and small data may be better than unfiltered big data" and "the soil water content in the middle layer exhibits remarkable superiority in comparison to the surface with its highest-level variability" may be case-specific rather than generalizable. It is important to consider other data-driven and assimilation methods, such as deep neural networks, particle filtering, and MCMC, leading to potentially different outcomes. I would like to see some clarifications regarding this issue.

Answer:

Thank you for your constructive comments. We have accepted your suggestions and evaluated a new NP-DWA framework where EnKF is replaced by particle filtering (PF). **Fig. S1** depicts the expected data-worth of potential observations of θ_S , θ_M , and θ_D regarding the retrieval of $\theta_{0.30}^{ave}$, $\theta_{0.60}^{ave}$, and $\theta_{1.00}^{ave}$, respectively. A comparison of **Fig. S1** and **Fig. 4** reveals that the spatio-temporal changes of expected data-worth under these two assimilation methods are remarkably similar. This demonstrates the generalizability of our proposed framework and related conclusions under different data assimilation schemes. To avoid duplication of research, we are sorry that we finally decided not to add the results of PF in the main text, but rather to include them as supplementary material in the revised manuscript (please see Lines 165-172 and Supplementary).

In addition, we also tested two other NP-DWA frameworks where GP was replaced by support vector machines (SVM) and random forests (RF), respectively. The temporal changes of expected data-worth metrics are depicted in **Fig. S2**. Only the results at DAHRA are presented here. A comparison of **Fig. S2** and **Fig. 4** indicates that although the magnitude and trends of data-worth vary slightly across different machine learning methods, the selection of the optimal monitoring depths for specific targets is quite consistent. For example, the optimal observation depth shifted as the prediction target varied, and soil water content in the middle layer robustly exhibited remarkable superiority in the construction of model-free soil moisture models. Moreover, the performance comparison of various machine learning algorithms in reproducing soil moisture dynamics has been widely discussed in previous studies (Dubois et al., 2021; Liu et al., 2020; Prakash et al., 2018). In particular, the ability of GP to reproduce the nonlinearity of soil water problems has also been demonstrated in (He et al., 2023; Ju et al., 2018; Wang et al., 2021). Therefore, we finally decided only to include these results as supplementary material as well in the revised manuscript.

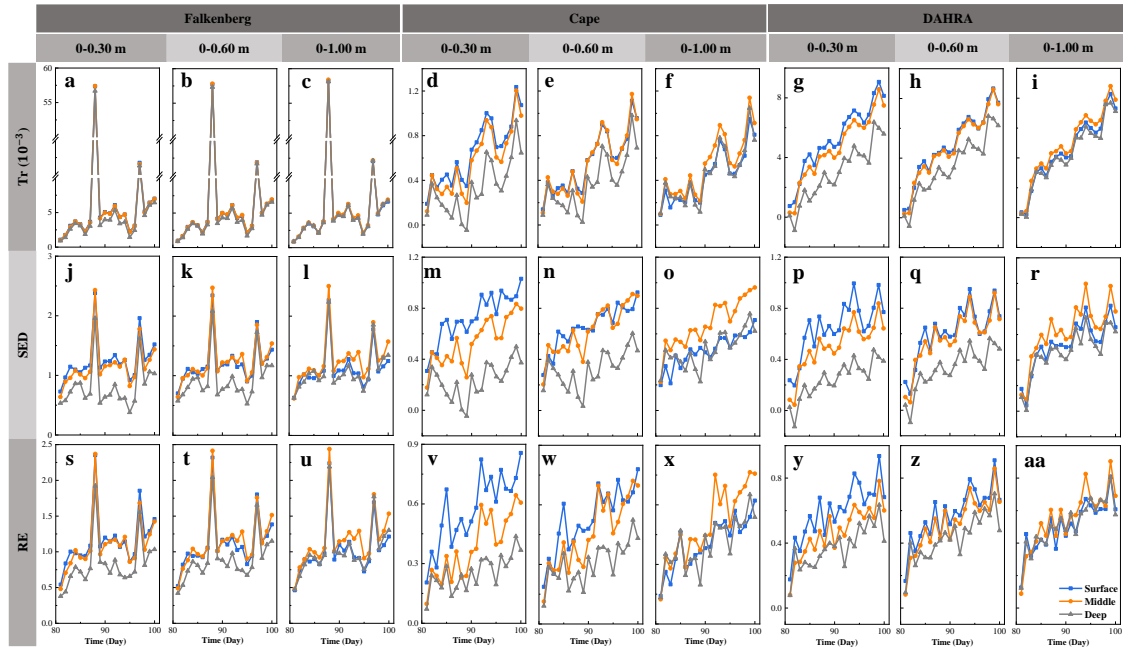


Figure S1. The expected data-worth of potential soil moisture observations in the surface, middle, and deep layers in the form of trace (T_r), Shannon entropy difference (SED), and relative entropy (RE), respectively, regarding the retrieval of average soil moisture in the top 0.30 m, 0.60 m, and 1.00 m at three sites, when EnKF is replaced by particle filtering (PF) in the proposed NP-DWA framework

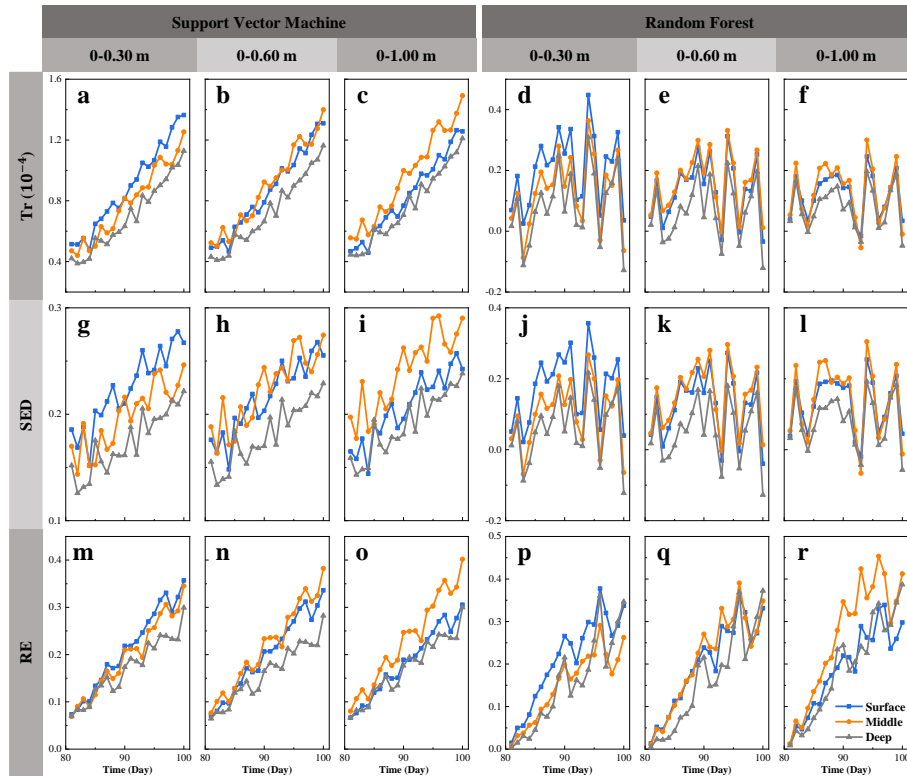


Figure S2. The expected data-worth of potential soil moisture observations in the surface, middle, and deep layers in the form of T_r , SED , and RE regarding the retrieval of average soil moisture in the top 0.30 m, 0.60 m, and 1.00 m at DAHRA site, when GP is replaced by support vector machine (SVM) and random forest (RF) in the proposed NP-DWA framework, respectively

References:

- Dubois, A., Teytaud, F. and Verel, S., 2021. Short term soil moisture forecasts for potato crop farming: A machine learning approach. *Computers and Electronics in Agriculture*, 180: 105902.
- He, L. et al., 2023. Physics-constrained Gaussian process regression for soil moisture dynamics. *Journal of Hydrology*, 616: 128779.
- 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.
- Liu, Y., Jing, W., Wang, Q. and Xia, X., 2020. Generating high-resolution daily soil moisture by using spatial downscaling techniques: A comparison of six machine learning algorithms. *Advances in Water Resources*, 141: 103601.
- Prakash, S., Sharma, A. and Sahu, S.S., 2018. Soil moisture prediction using machine learning. *IEEE*, pp. 1-6.
- Wang, Y. et al., 2021. A nonparametric sequential data assimilation scheme for soil moisture flow. *Journal of Hydrology*, 593: 125865.

2. It is recommended that the methodology section of this paper be better presented. Specifically, the problem setup for moisture prediction, an explicit list of the contents of vectors X and y should be provided prior to section 2.1. This will enable the reader to better understand the proposed data-driven framework.

Answer:

Thank you for your valuable suggestions. We have revised the methodology section. A clearer description of vectors X and y has also been added in section 2.1 of the revised manuscript (please see Lines 155-160 and 175-185).

3. Some techniques have been proposed for better performance in nonlinear problems, e.g., restart, iterations. How will these techniques perform in NP-DWA?

Answer:

We thank the reviewer for the constructive comments. In fact, the procedure of constructing GP models in a sequential manner in our NP-DWA framework resembles a restart operation. At any time step $t=k$, the construction of the GP model does not solely rely on the information from the previous time step, instead, its training data includes all available soil moisture data from $t=1$ to $t=(k-1)$. This restart-like operation ensures that the training database is sequentially augmented to include more diverse training scenarios, so that actual observations can be accurately “captured” by the generated potential observation samples. Ultimately, the accuracy (or reliability) of our NP-DWA framework for data-worth assessment can be guaranteed. Related descriptions have been added in the revised manuscript (please see Lines 175-185). The performance improvements of these techniques such as restart and iterations for our NP-DWA will be explored in our future study.

4. L31: "An alternative monitoring strategy with a larger data-worth was prone to a higher DW assessment accuracy within the proposed NP-DWA framework" This sentence is meaningless and should be removed.

Answer:

We have accepted the reviewer’s suggestion and deleted this sentence (please see Lines 35 and 665-670).

5. Please provide the dimensionality for all the involved vectors and matrices.

Answer:

We have added the dimensionality for all the involved vectors and matrices in Section 2 of the revised manuscript.

Anonymous referee 2

The manuscript by Wang et al. presents a framework for determining data worth of soil moisture measurements. The framework uses Gaussian process regression (GP) to replace unsaturated flow models; GP is combined with EnKF to evaluate the prior, post-posterior, and posterior (regarding potential new data) distributions of variables of interest (i.e., soil moisture averaged for varying portions of the soil column). The change of distributions was summarized using three indices to determine data worth. The framework was demonstrated using three soil columns from ISMN using several test cases to illuminate the roles of prior data length, observation noise, and combinations of potential new data.

Overall the manuscript is clearly organized and results are thoroughly described and discussed. Data worth analysis in a model-free framework (using machine learning) is novel. I therefore recommend it to be published in HESS after minor revisions. Below are detailed comments, most of which are intended to improve clarity and generalizability of the presented framework.

Some discussion is needed to support the conjunctive use of GP and EnKF. EnKF is very commonly used with deterministic models such as Hydrus for data assimilation and propagation of uncertainty in time. GP is capable of assimilating newly available data by simply training the model again once new data is available and calculating the mean and covariance of a variable of interest. Given this, it seems that the data-worth framework can be done using GP alone, without EnKF. Some discussion is needed to help readers understand the framework design. For example, what is the role of EnKF in improving accuracy of data worth estimation or enabling the framework to be used for machine learning algorithms other than GP? Will EnKF result in covariance matrices different from those calculated by GP?

Answer:

We thank the reviewer for the constructive suggestions. In fact, the necessity of the conjunctive use of GP and EnKF has been discussed in detail in our previous studies (Wang et al., 2021; Wang et al., 2021). As stated in (Wang et al., 2021), on the one hand, the fusion of EnKF 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 Kalman update, the forecast cross-covariance ($\mathbf{C}_k^f \mathbf{H}^T$) between the state (\mathbf{Y}_k^f) and the predictions corresponding to available observations (\mathbf{HY}_k^f) in Eqs (6-7) constrained the otherwise high error covariances of state variables at unobserved depths, which resulted in a significantly reduced uncertainty for this hybrid method relative to GP alone. To keep this manuscript more focused, we finally decided to only add a brief explanation of the conjunctive use of GP and EnKF (please see Lines 160-167), without adding extra cases with GP alone in the revised manuscript.

References:

- 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.

It is unclear from my reading (1) what depth(s) are used as prior data for training GP, (2) what

specific depth(s) are considered for potential data (θ_s , m, d), and (3) do the depths for prior data and potential new data overlap?

Answer:

We are sorry for the confusion caused by our unclear description. (1) 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. (2) The depth of the potential soil moisture data is different in different test cases. For example, as listed in **Table 2**, the potential data in TC1-1, TC2-1, TC3-1, TC4, TC5, TC6, and TC7 refers to soil moisture in the surface layer (θ_s), i.e., $z=0.08$ m at Falkenberg and $z=0.05$ m at Cape and DAHRA. (3) The depths of the prior and potential new data in this study partially overlapped due to the limited depths of observations under real-world circumstances. For example, TC1-1 at Falkenberg used soil moisture observations taken at six depths ($z=0.08, 0.15, 0.30, 0.45, 0.60,$ and 0.90 m) over the first 80 d as prior data, and the generated soil moisture at $z=0.08$ m over the last 20 d as potential data. We have added the relevant descriptions in the revised manuscript (please see Lines 345-355).

Line 305 - how is the noise level used in the computations? For GP or for KF? Is the noise level specified or estimated when training GP?

Answer:

Thank you for your carefully reading. Noises from soil moisture observations are considered in both GP and EnKF in this study. At any time step $t=k$ during GP modelling, the observed time series from $t=1$ to $(k-1)$ are corrupted by the prescribed observation noises satisfying Gaussian distribution to obtain N sets of training data. Subsequently N sets of GP models are constructed independently, to generate $\mathbf{Y}_k^f = [\mathbf{y}_{k,1}^f, \mathbf{y}_{k,2}^f, \dots, \mathbf{y}_{k,m}^f, \dots, \mathbf{y}_{k,N}^f]^T$ in Eq. 7 (please see Lines 175-177). In the analysis stage of EnKF, the real-time observation \mathbf{d}_k^{obs} perturbed by the specified noise was assimilated via Eq. 8 (please see Lines 228-235). Considering the difficulty of determining the observation noise under real-world circumstances, the noise level is artificially specified in this study. We have added the relevant explanations in the revised manuscript (please see Lines 315 and 370).

Line 375 - I don't have a specific comment here, but would like to highlight that the difference in data worth between a physical model and a machine learning model is very interesting and a key contribution of this study.

Answer:

Thank you for your valuable recognition. Considering that the data-worth analysis in physical models has been discussed in detail in our previous study (Wang et al., 2018), this study did not add the corresponding test cases, but directly compared the findings in (Wang et al., 2018) with the results of the proposed NP-DWA. As you mentioned, the comparison in data-worth between physical and machine learning models can help modelers better understand the impact of the ways data being utilized on its worth. We have accepted your suggestions and further highlight this difference in the revised manuscript (please see Lines 20-30, 420-428, 565-575, and 640-650).

References:

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.

Predictions (variables of interest) considered in this study are depth-averaged soil moisture. It would be of more interest to the broader hydrologic community to discuss the potential application of the presented framework for other types of predictions, e.g., ET, infiltration.

Answer:

Thank you for your constructive suggestions. This study used the GP model to reconstruct the nonlinear relationship between multiple variables (including time, depth, precipitation, and air temperature) and soil moisture. Therefore, the expected data-worth of future monitoring programs regarding the estimation of depth-averaged soil moisture can be evaluated. Our future study will further discuss the application of the NP-DWA framework for other types of predictions, e.g., ET, infiltration.

Line 365: “matrixes” should be “matrices”

Answer:

Thank you for your carefully reading. We have revised “matrixes” to “matrices”.