Reply to Reviewer #1

General comments

This paper presents an interesting and comprehensive study on the use of machine learning (ML) algorithms for predicting surface soil moisture content at a global scale. The authors use eight different machine learning algorithms trained with in situ data to predict soil moisture at the top 5 cm of the soil across the globe at a spatial resolution of 1 km and a temporal coverage of 2000-2018. The performance of various ML algorithms was evaluated. They also created 10 ensemble models from five optimized base models with the highest performance metrics and propose KNR_RFR_XB as the best ensemble for soil moisture prediction.

The paper provides a thorough description of the model building and validation process. However, prior to publication, the following questions should be addressed:

Thanks for your thorough review and detailed comments on our manuscript. Your comments are immensely valuable in enhancing the manuscript's quality.

Specific comments

Abstract

Please specify which soil layer, period (coverage), temporal and spatial resolution were considered in the generation of the machine learning models.

1. Line 15-20: Instead of using the term "best data product", it would be more appropriate to use an adjective such as reliable or well-validated, as the data product from machine learning may be produced empirically without explicit knowledge of the physical processes involved. This may introduce additional uncertainties into the final output. (O., S., Orth, R. Global soil moisture data derived through machine learning trained with in-situ measurements. Sci Data 8, 170 (2021). https://doi.org/10.1038/s41597-021- 00964-1)

Reply 1: Thanks for your advice. We agree with your comment. In line 18, we will change "the best" into "reliable".

2. Line 24-34: Please clarify what it means to have the best performance on "test_random" set, "test_temporal" and "test_independent-stations" and reword this sentence to focus on the meaning rather than the jargon terms. Also, please use full wording before notation (XB = extreme gradient boosting).

Reply 2: Thanks for your advice. As what we understood, here we need to clarify two things: (1) meaning of "best performance", (2) explain a bit of these 3 terms: "test_random" set, "test_temporal" and "test_independent-stations".

In line 23-24, we will modify the sentence "The result showed that K-neighbours Regressor (KNR) performs best on "test_random" set, while Random Forest Regressor (RFR) performs best on "test_temporal" and "test_independent-stations." to a more accurate description with quantitative indicators: "The result showed that K-neighbours Regressor (KNR) had the lowest Root Mean Square Error (0.0379 cm^3/cm^3) on "test_random" set, while Random Forest Regressor (RFR) had the lowest RMSE (0.0599 cm^3/cm^3) on "test_temporal" and AdaBoost (AB) had the lowest RMSE (0.0786 cm^3/cm^3) on "test_independent-stations".

We will explain the meaning of "test_random": for testing the performance of randomly split data during training, "test_temporal": for testing the performance on the period which were not used in training, "test_independent-stations": for testing the performance on the stations which were not used in training.

In line 33, we will replace “XB” with “Extreme gradient Boosting”.

Introduction

3. Please provide citations for all claims or statements that require them (e. g., Line 42:45, Line 47:49, Line 52:54)

Reply 3: Thank you very much for your valuable suggestion. We will ensure that all the claims and statements, which you've highlighted as requiring citations, will have proper citations in our revised manuscript. These citations will contribute to the accuracy and reliability of the paper.

4. Line 65-68: The statement “Another advantage of using ML techniques is that they can help to reduce the uncertainty…” may not be accurate, as machine learning may introduce additional uncertainties into the final output (see response to line 15:20). It would be helpful to discuss this issue in more detail and provide references to support your argument.

Reply 4: Thanks for your advice. In our study, when we said ML can help reduce the uncertainty, we initially meant that compared to traditional physical models, ML can alleviate uncertainties (e.g. on parameter optimization) to some extent. Nowadays people integrate ML and physical models to reduce the prediction uncertainty (Roy et al. 2023). However, we agree that machine learning itself has uncertainties. To provide a more accurate representation of our standpoint, we will remove this statement and will state the limitations of machine learning in the subsequent discussion section which is related to the 16th comment.

5. Line 75: Please specify which soil layer, period (coverage), temporal and spatial resolution were considered in the generation of the machine learning models.

Reply 5: Thanks for your attentive review. In line 78, we will add the information: for predicting SSM in 0-5 cm depth with daily and 1 km resolution from 2000 to 2018.

Data

6. Line 91:92: Did you filter out NAN values from the predictor data based on the in-situ soil moisture data? If not, what did you do? Please make the statement clear.

Reply 6: Thanks for your advice and sorry if this was unclear in the manuscript. We first organized the in-situ soil moisture data, and then extracted the predictor variables for each station in the period when they have in-situ soil moisture data. After this, we filtered out NAN values for each data pair (soil moisture and all predictor variables in one specific time and location). If there are NAN values in either soil moisture or all predictor variables, we removed it. (Because we organized the in-situ soil moisture data in a specific period (2000-2018), there is missing soil moisture values during this period as well).

To make this statement clear, in lines 92-93, we will change from “In this study, we extracted in-situ SSM from ISMN from 1 January 2000 to 31 December 2018, filtered the NaN (Not a Number) values of different predictor variables and kept 1722 stations for further analysis (Figure 1). Land surface information was also incorporated to prepare training and testing sets.” to “In this study, we extracted in-situ SSM from ISMN from 1 January 2000 to 31 December 2018, and then the predictor variables were incorporated to prepare training and testing sets. We filtered the NaN (Not a Number) values of SSM and different predictor variables, and after this 1722 stations were kept for further analysis (Figure 1).”

7. Line 104: Why is "Year/ DOY " a predictor?

Reply 7: Thank you pointing out this question. There are two reasons why we include Year and DOY as predictors.

(1) Firstly, they can help capture the seasonal variations in soil moisture. Soil moisture can exhibit significant changes across different seasons, such as summer droughts or winter wet periods. By incorporating year and DOY as input variables, the model can learn and capture these seasonal trends, leading to more accurate predictions of soil moisture changes at different time points.

(2) Secondly, Year and DOY help analyze the historical trends of soil moisture. Certain years or specific dates might exhibit consistent trends in soil moisture, such as prolonged dry spells or continuous rainfall. The model can leverage this historical trend information to make more accurate predictions of soil moisture changes.

In addition, we will implement the feature importance experiment (which is related to the 10th comment) to provide more details.

8. Line 140: To improve readability, could you please ensure that all temporal coverages are consistent (2000-2018). I noticed that line 117 mentions the years 2000-2019, but I believe only the years 2000-2018 were used for analysis.

Reply 8: Thanks for your careful check. The correct temporal coverate should be 2000-2018. In line 117, “2019” will be changed to “2018”.

9. Line 232: " The full data contained a total of 735,475 registries (each sample contained nineteen predictor variables) and was differentiated in the training set and three different test sets, based on the following strategy (also described in Table 2)" I am finding this a bit confusing and would appreciate some clarification. Is it correct that the combined data of all predictors and in-situ data is 735,475, and that this data was split into train and test samples as shown in Table 2?

Reply 9 : Yes, that is correct. We aimed to explain how we split the full data into 4 parts. Our full data has 735475 rows, and each row includes in-situ soil moisture and 19 predictor variables.

10. Additionally, I noticed that Table 1 shows fewer than 19 predictors. Could you please explain how you arrived at the number 19 predictors? Is it necessary to use all 19 predictors, or could feature or predictor importance be performed to determine the various influences of predictors and select important features if necessary?

Reply 10: Thanks for your careful check. Because in Table1 we wrote “Soil Texture”, actually it includes sand, silt, and clay content. In the fifth row and the first column of Table1, we will modify “Soil Texture” to “Clay Content/Sand Content/Silt Content”. In the fifth row and the second column, we will modify “Soil texture (proportion of clay, sand, silt)” to “Proportion of clay/proportion of sand/proportion of silt”.

For the feature importance, thank you very much for your valuable feedback. In response to your question regarding the necessity of using such a multitude of input variables, we will conduct a comprehensive feature importance analysis across different regression models to address this inquiry. The KNR+RFR+XB ensemble model performed the best, therefore we plan to compute the feature importance based on these three MLs, and then we calculate the mean feature importance of all features for these three MLs with permutation importance method (Li et al. 2021).

We believe by doing this experiment, we will observe the feature importance across different ML models and the overall feature importance. Then we can conclude that which input variables contribute to predicting the target variable. The result of this experiment will be added in the supplementary information.

Methodology

11. Line 279: Introduce the full wording e.g., coefficient of determination before using the associated notation r 2.

Reply 11: Thanks for your careful check. In line 279, we will add the full wording of r^2, from "the r^2 score" to "the Coefficient of Determination R-square (r^2) score".

12. Line 350: Could you please explain the reason for choosing both r and r 2 as evaluation metrics? What is the significance of using both in this context?

Reply 12: Thank you for your advice. The r score measures the correlation between two or more variables (A and B, for example). The r^2 score measures how much variation of B values can be explained by A values). They are two very different metrics and each measures a different thing. The r^2 score is more important in ML training. Based on your advice, we decided to keep r^2 score as evaluation metric to provide a comprehensive view of the model's performance, accounting for its ability to explain variance.

However, in the evaluation of individual independent sites (section 4.4 and 4.5), we observed instances where r^2 values could be negative due to poorer model performance on some sites because they were not included in the training and different climate conditions from training sites. These situations might potentially lead to misinterpretations of model performance. To ensure an accurate assessment, we will keep using r score values in this evaluation part, which directly quantify the linear relationship between predicted and observed values.

Based on above, section 4.4 and 4.5 which are related to individual independent stations evaluation we will keep same as before. Other parts we will remove the r score values.

13. Line 384: Arranging Table 3, as well as any other relevant tables, in ascending or descending order based on performance can improve readability.

Reply 13: That is a nice idea, sorting will be done based on increasing RMSE. Table 3 will be changed into sorting on increasing RMSE. Table 4 and 5 will be changed into sorting on increasing RMSE on "Test_random".

14. Line 430: 470: Could you please explain why the Kruskal-Wallis test is necessary when you can already compare models using various performance metrics such as RMSE and r 2 ?

Reply 14: We use the Kruskal-Wallis test to analyze if there is significant difference between the performance of the different MLs and ensemble models based on the squared errors of the predictions, therefore we find it important to show the results of this test. We see the metrics but those are single values and statistical tests are the way of distribution comparison.

15. Line 480: Please move Figure 5 into supplementary information.

Reply 15: Thanks for your advice. We agree with it. Figure 5 will be moved into supplementary information.

16. In addition to addressing these questions, it would also be helpful to compare the performance metrics and uncertainties of your proposed ensemble model with other existing soil moisture products or methods. This would provide more context for your results and help readers evaluate their significance. It would also be interesting to discuss the limitations and assumptions of your machine learning algorithms and how they affect the reliability and applicability of your soil moisture predictions. Finally, it would be useful to suggest future work or research directions to improve or extend your study.

Reply 16: Thanks for your valuable advice. We will modify like following:

(1) In line 596, we will add the comparison with other methods and soil moisture products. We drafted an initial paragraph could be added: "Furthermore, a comparison of ensemble models with other existing

soil moisture products and methods highlights their superior performance. For instance, evaluating the root mean square error (RMSE) on randomly selected test samples, our KNR_RFR_XB ensemble model achieves an RMSE of 0.0355 cm³/cm³. In contrast, the RFR model used for generating a global daily 1 km soil moisture product presents an ubRMSE (unbiased root-mean-square error) of 0.045 cm³/cm³ (Zheng et al. 2023) and the RFR model used for generating a global daily 0.25 degree soil moisture product presents an RMSE of 0.05 cm³/cm³ (Zhang et al. 2021). The RFR model used for reconstruction of a daily SMAP surface soil moisture dataset shows an ubRMSE of 0.04 cm³/cm³ (Yang and Wang 2023). The XB model used for generating a global daily 1 km soil moisture product presents an RMSE of 0.038 cm³/cm³ (Zhang et al. 2023). These comparisons demonstrate that our ensemble model has the potential to improve predictive accuracy compared to individual methods. This makes our model a candidate for further exploration as an effective tool for accurate prediction of soil moisture."

(2) For the uncertainty, limitations and future work. We will rewrite the content in lines 618-627, and move the content in lines 597-604 to lines 628-. We will add the future work in the end of this part. The final version of the uncertainty, limitations and future work will be like: "Although the proposed ensemble model has been demonstrated to be an effective solution to predict soil moisture, there are still limitations. Firstly, our ensemble model can be limited in the areas outside the training conditions such as climate zones BSh, BWh and BWk (Sungmin and Orth 2021). Secondly, hyper-parameters tuning is a computationally expensive operation that proved to have an important effect on the performance of each machine learning model. However, it involves the human factor with expertise in choosing the right ranges for each hyperparameter in order to achieve the best possible training. We recommend to carry out the training in at least two iterations, first selecting wider parameter intervals, and then narrowing it down to ranges in proximity to the best value detected in the initial experiments. Thirdly, the training of algorithmic implementations within ensembling environments requires more computational power. However, the increased and the more stable prediction behaviour is more desirable when tackling tasks where high performance metrics are expected. Lastly, depending on the base models, the performance of ensemble models can sometimes worsen when compared to well-performing optimised algorithms. For this reason, it is advised to optimise the base algorithms as much as possible for the chosen task. It is also observed that regression algorithms with a higher complexity generally displayed a higher generalisation capacity. The above four points highlight the limitations and challenges of our ensemble model in practical applications. Future research directions may include enhancing the generalization ability of the model to obtain more accurate predictions in areas outside the training conditions, such as increasing the training data or using transfer learning techniques. In addition, more efficient and automated hyperparameter tuning methods can be explored to improve the performance of the model. Addressing the computational demands required to train the ensemble models is also a key direction, possibly involving the use of parallel computing, distributed frameworks, or hardware acceleration approaches, all of which aim to further enhance the performance and applicability of our models."

The content we will add above is also related to the 4th comment.

References

Li, W., Migliavacca, M., Forkel, M., Walther, S., Reichstein, M., & Orth, R. Revisiting global vegetation controls using multi‐layer soil moisture. *Geophysical Research Letters, 48*, e2021GL092856, 2021

Roy, A., Kasiviswanathan, K., Patidar, S., Adeloye, A.J., Soundharajan, B.S., & Ojha, C.S.P. A physics‐aware machine learning‐based framework for minimizing the prediction uncertainty of hydrological models. *Water Resour. Res.*, e2023WR034630, 2023

Sungmin, O., & Orth, R. Global soil moisture data derived through machine learning trained with in-situ measurements. *Sci. Data, 8*, 1-14, doi:10.1038/s41597-021-00964-1*,* 2021

Yang, H., & Wang, Q. Reconstruction of a spatially seamless, daily SMAP (SSD_SMAP) surface soil moisture dataset from 2015 to 2021. *J. Hydrol., 621*, 129579, 2023

Zhang, L., Zeng, Y., Zhuang, R., Szabó, B., Manfreda, S., Han, Q., & Su, Z. In Situ Observation-Constrained Global Surface Soil Moisture Using Random Forest Model. *Remote Sens., 13*, 4893, doi:10.3390/rs13234893*,* 2021

Zhang, Y., Liang, S., Ma, H., He, T., Wang, Q., Li, B., Xu, J., Zhang, G., Liu, X., & Xiong, C. Generation of global 1&thinsp;km daily soil moisture product from 2000 to 2020 using ensemble learning. *Earth Syst. Sci. Data, 15*, 2055-2079, 10.5194/essd-15-2055-2023*,* 2023

Zheng, C., Jia, L., & Zhao, T. A 21-year dataset (2000–2020) of gap-free global daily surface soil moisture at 1-km grid resolution. *Sci. Data, 10*, 139, 2023