Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC1
https://doi.org/10.5194/gmd-2022-211-RC1, 2022

Comment on gmd-2022-211

Anonymous Referee #1

Referee comment on "Evaluating a global soil moisture dataset from a multitask model (GSM3 v1.0) with potential applications for crop threats" by Jiangtao Liu et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2022-211-RC1, 2022

The manuscript "Evaluating a Global Soil Moisture dataset from a Multitask Model (GSM3

v1.0) for current and emerging threats to crops" presents a LSTM-based machine learning model for global soil moisture estimation. The authors used a multi-loss framework to train the LSTM model, and employed multiple reference datasets for evaluation. Specifically, the work investigated spatial generalization ability by cross-validations, both randomly and continent-based sampling. The overall quality of the work is excellent and fits in the scope of GMD. With that, I do have the following comments.

The title is somewhat confusing to me. It indicates the authors also care about crops in addition to soil moisture, as soil moisture will affect the crops. So I was expecting to read something about agricultural applications (e.g., use the DL-based soil moisture to drive some crop models). However in the manuscript, it is only mentioned before the conclusion section that this model will be put into production, and the major focus of the paper is model benchmarks on soil moisture. I suggest the authors have more discussion on crop applications, or revise the title to make it clearer. In addition, if the main focus is to evaluate the dataset instead of the LSTM model, I suggest making the dataset publicly available and adding a link to access this dataset.

To clarify, what's the difference between the SoMo.ml model and the authors' LSTM model? Do they share the same dynamic and static input variables with the only difference as the loss function? In Figure 1 and 2, the authors compared the metrics derived from the training period (Multitask_train and SoMo.ml_train). The R value or RMSE from the whole training period is not important, as one can always overfit the model. I understand

that a barplot for Multitask_train shows the model is not overfitting, but a comparison between Multitask_temporal and SoMo.ml_temporal would make more sense to me.

What is ubRMSE in Table1? Does Corr represent Pearsonr correlation coefficient? In Figure 3 and 4, the meanings of colormaps are not the same. In the correlation map, greener represents better performance, but it is worse performance in the RMSE plot. I would suggest a uniform colorbar with light (dark) colors for high performance and dark (light) colors for low performance.

The LSTM model is optimized towards both the in-situ estimation and the satellite products. In Figure1 when comparing the model performance, the authors selected the SMAP and GLDAS products, which are gridded datasets. When evaluated against the ISMN dataset, it is not a fair comparison because of the coarse resolution of SMAP. How do the authors correlate in-situ measurements with gridded datasets? Will the result change when switching to 25-km resolution (i.e., with coarser resolution, the dataset will lose more representations)? A further question is that, what is the meaning of the LSTM outputs? Is it the best estimation over the grid points (e.g., an average over the grid spacing), or the best estimation for the in-situ observations?

The authors split the global dataset into 7 continents. Would it be a more straightforward comparison to have a similar 7-fold cross validation to match the number of continents in Section 3.2? I believe 5-fold is a common choice but 7-fold may be a more intuitive and fair comparison.

When analyzing the factor controls, the authors selected a random forest model. What is the performance/skill of this random forest model? Is Figure 5 using spatial R as target or temporal R, or is it a multi-output model? At line 329 it shows the target is temporal R, but line 330 shows R is from either temporal or spatial tests.

Regarding the random forest model, is the conclusion independent of the model choice (or what's the reason to choose a random forest model)? Will we get different results if we

switch to extreme boosted trees or other tree-based models? Also please check the name of the python package. It is used as “sklearn” in python but the paper referenced (Pedregosa et al., 2011) shows “scikit-learn” in the title.

The author mentioned the driest sites are hard to predict, and suggested the reason as scarce but sudden rainfall events. Do authors believe it may be related to the loss function used in the LSTM model (i.e., some loss functions would not emphasize the extreme high/low values)? I also don’t follow the logic behind Line 345. From the error type analysis, the authors mentioned the comparison between temporal and spatial tests, especially for error type B (nonstationary). But the boxplots of R from the temporal test and spatial test over driest sites are similar to me, with a median R of approximately 0.7.

Line 146 mentioned the use of different resolutions for the static terrain attributes. What is the aspect resolution used in the random forest model?

The code files in the Zenodo repository are not enough to replicate the experiments. I’d suggest the authors update their repository, either during or after the peer-review process.