Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2022-137-RC2, 2022


# Review of "An 8-day composited 36 km SMAP soil moisture dataset from 1979 to 2015 produced using a random forest and historical CCI data" by Haoxuan Yang et al. submitted to ESSD

Anonymous Referee #2

Referee comment on "An 8-day composited 36 km SMAP soil moisture dataset from 1979 to 2015 produced using a random forest and historical CCI data" by Haoxuan Yang et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-137-RC2, 2022

Yang et al. present a study recalculating global soil moisture data of the Soil Moisture Active Passive mission from data of the Climate Change Initiative programme and the International Soil Moisture Network by means of a Random Forest. The resulting dataset shall last back to 1979 at a resolution of 36 km and 8 days composites.

In general such a dataset appears to be a valuable and worthwhile aim of a study. However, the applied methods omit addressing substantial questions about the validity of the data. Most fundamentally, the approach assumes that the barely five years of data overlap between 2015 and 2019 can describe the global relationship for 1979 to 2015, although this exactly is the period in which global change starts to become traceable in data and although global change is known to happen non-uniformly across the globe. As such, validation of the derived data requires more detailed analyses than the rather rough screening presented here. I would not expect overall RMSE statistics to be applicable for the desired outcome.

Given the large amount of data in the ISMN database, the seasonality at most locations already providing a very simple first order dynamics and the very broad generalisation of a value for soil moisture at a grid of 36 km, I would be very interested about the ability of the model to reproduce deviations from the overall patterns. Since the data are spatially and temporally at least to some degree dependent/correlated, maybe LSTMs or other sorts of machine learning are more appropriate (cf. Fang et al. 2017,

https://doi.org/10.1002/2017GL075619; Abbes et al. 2019 https://doi.org/10.1109/IGARSS.2019.8898418; Breen et al. https://doi.org/10.3390/make2030016, Zang et al. 2022 https://doi.org/10.1080/10106049.2022.2105406)?

Moreover, there are already a number of global soil moisture products derived by machine learning. Among others, there are Sungmin and Orth 2021 (https://doi.org/10.1038/s41597-021-00964-1) in 0.25 degree resolution form 2000 to 2019 and Martens et al. 2017 (https://doi.org/10.1080/10106049.2022.2105406) ranging back to 1980 (GLEAM v3.6a). Given that these examples use very different and likely more sophisticated approaches, the authors should clarify clearly, what advances their dataset does provide. Hence in addition to the evaluation stated above, this opens a second area of analyses, which have not been addressed yet.

With substantial deficits in these three domains (extrapolation from 5 years to non-stationary system, evaluation of performance beyond mean characteristics of climate zones, evaluation against other soil moisture products) the manuscript and the data deserve fundamental revisions before publication.