Earth Syst. Sci. Data Discuss., author comment AC5
https://doi.org/10.5194/essd-2022-177-AC5, 2022

Reply on RC2

Qingliang Li et al.

Author comment on "A 1□km daily soil moisture dataset over China using in situ measurement and machine learning" by Qingliang Li et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2022-177-AC5, 2022

We are very grateful to Reviewer for reviewing the paper so carefully. These comments are very helpful to improve the quality of the manuscript. Please find my itemized responses in below and my revisions will be in the re-submitted files.

Major comments:

Comment#1: The paper writing is extremely poor. Language errors and statement repeats or inconsistency are found through the whole manuscript. For example, "dataset of China" in the title should be "dataset over China" or "dataset for China"; in Lines 14-15, the authors have stated "high quality gridded soil moisture products" are "usually available from remote sensing… with coarse resolution" but then they raise that "high quality" is characterized by "high-resolution…", which is obviously contradictory; there are many other cases like the usage of "… is acted as".

Responds□We corrected the errors mentioned by the reviewer. We have also carefully checked the whole manuscript and revised the inaccurate description. We will also invite a native English speaker to polish the manuscript. The sentence in lines 14-15 is corrected as:

High quality gridded soil moisture products are essential for many Earth system science applications, while the recent reanalysis and remote sensing SM data are often at coarse resolution and remote sensing SM data are only for the surface soil.

Comment#2: More importantly, the Results part contains many statements that are actually discussion while the Discussion parts contains too many results.

Responds□Thanks for your kind comments and helpful suggestions. We have moved Section 3.4 in the old manuscript to the discussion as Section 4.1, and put some short discussions in the results of the old manuscript into the "Discussion" part. The new section in the discussion is as follows:

4.3 Factors affecting the quality of SMCI1.0

Figure 2 and 2s shows that the result at 70 cm and 90 cm were significant worse than

those at other depths. The reason may be that RF model is difficult to estimate accurate SM for only a few in-situ SM stations. From Fig. S1 (b), we can see that the total numbers of data at 70 cm and 90 cm soil depths are quite small. In other words, more abundant of data were expected to help RF model 'learn' complete relationship between covariates and in-situ SM and further improve the quality of high resolution SM in China. Meanwhile, compared with the previous study of Sungmin et al. (2020), our SMCI1.0 showed the superior quality (Figure 4-6), because the larger numbers of in-situ SM data in China were applied for RF modelling. From Figure 5, during the rainfall near 91th day across the Tropical Monsoon Climate zone (Am) and near 1st day across the Snow climate with dry winter zone (Dw), the in-situ SM did not increase with high precipitation, but the SMCI1.0 product could capture the increase in SM (denoted in the light blue rectangle). The reason may be that the applied covariates had bias with in-situ measurement and further affected estimation by RF model. Meanwhile, we also found the RF model could overcome much bias in dry conditions, except for that from 196th to 305th days in the snow climate, fully humid zone (shown in the light red rectangle). In the case of 30 cm soil depth (Fig. S5), we could see an agreement between several peak events, it could be attributed to the soil texture homogeneity at the 10 and 30 cm soil depths. Almost all climatic regions had lower dynamic ranges at 30 cm soil depth than that at 10 cm, this may be attributed to the persistent behaviour of SM at 30 cm soil depth. In the case of 30 cm soil depth in Fig. S6, the SMCI1.0 product had higher accuracy than that at 10 cm soil depth (Figure 6), especially in terms of ubRMSE and MAE metrics. The reason may be the background aridity led to low variability of SM in the deeper layers (Karthikeyan and Mishra 2021) and the RF model can capture the variation in SM easier. Interestingly, it was inconsistent for the results of R, ubRMSE, and MAE in Fig. 2 and Fig. 4, which had the same phenomenon with the previous study (Sungmin and Orth 2020) (represented in their Fig. 4 and Fig. 5). For example, SMCI1.0 product had the ubRMSE, MAE and R being 0.046, 0.035 and 0.889 at 10 cm soil depth in Fig. 2. However, in Fig. 4, the box-plot represented the lowest ubRMSE, MAE and highest R of SMCI1.0 product were nearly 0.03, 0.02, and 0.7, respectively. The reason may be that the same metrics were calculated in different ways, the one in Fig. 2 was to count the results of all stations and temporal period, and the one in Fig. 4 was to count the results of only temporal period at one station. It was necessary to note that we also compared the RF model with other ML models, including CatBoost (Dorogush et al. 2018), XgBoost (Chen et al. 2016), and Neural Network (Rosenblatt et al. 1958) based models. We found that the performance of these models is very similar to RF models with a R2 around 0.79. In addition, RF has been widely applied and recognized in SM prediction and many other fields (Carranza et al. 2021, Lin et al. 2022, Ly et al. 2021) and it does not take too much computing time to make the predictions for the whole China. Hence, we only took RF model to produce the high-resolution SM data.

We also removed the Section 4.1 in the old manuscript and put the related text into the "Conclusions" part. The new expression is as follows:

In this study, the gridded soil moisture was estimated through RF method in China based on the ERA5-Land reanalysis, USGS land cover type and DEM, reprocessed LAI and soil properties from CSDL, which included soil depths from 10cm to 100cm and had 1km spatial and daily temporal resolution over the period from 1 January 2010 to 31 December 2020.

Finally, we set "Sensitivity to precipitation, air temperature and radiation" as Section 4.2, as it is close to the new Section 4.1. We set Section 4.3 as "Factors affecting the quality of SMCI1.0". We combined the original Section 4.3 and 4.4 as the new Section 4.4 "Requirement of further validations and improvements". In addition, we have added the Section 4.5 providing some thoughts on our product about implications for the soil moisture modeling and attribution, meanwhile, in this section, we have also added the discussion about comparison between our product and previous products. The new expression is as follows:

In this section, we mainly discussed the comparison between SMCI1.0 and previous products, and

the implications for the soil moisture modeling and attribution. From the previous results in Section 3, we can see that SMCI1.0 generally outperforms the existing SM products (ERA5-Land, SoMo.ml and SMAP-L4) at most cases. The most important uniqueness of SMCI1.0 is taking the in-situ SM data as the training target with abundant sample size. Even though we used the ERA5-land to correct their means and standard deviation at each site, the temporal variation still came from the observations. We have also tested to train the RF model with the original SM observations and found that the performance of the model decreased dramatically with a R2 of 0.67 compared to the model with correction (a R2 of 0.79). And more importantly, the resulting SM maps demonstrated unreasonable noisy spatial distribution. These indicates that the in-situ SM in China have essential data inconsistency and the correction according to ERA5-Land is necessary which has physical consistency. Furthermore, SMCI1.0 is provided with relatively high spatial and temporal resolution (1-km and daily) for ten soil depths, which makes it possible for wider applications at finer scales and deep soils for the whole China, while reanalysis and remote sensing SM data are often at coarser resolution and remote sensing SM data are only for the surface soil. However, SMCI1.0 estimated by machine learning model cannot always reflect the variation of SM well, especially for some extreme events or so called "tipping points" (Bury et al. 2021). From Fig,5, we can see that SMCI1.0 deviated from the in situ SM in some cases, though this also happened to the other three SM products. For example, from 35th day to 61th day across the Snow climate, fully humid (Df), SMCI1.0 and SoMo.ml overestimated, while SMAP_L4 underestimated. "Tipping points" denoted that slowly changing SM sparks a sudden shift to a new (Bury et al. 2021). This is a huge challenge for estimating in-situ SM by ML models, because "tipping points" make the dynamics of complex system simplify down to the limited number of possible "normal forms" (Bury et al. 2021). ML models cannot accurately capture such extreme events. Hence, for these extreme events, we hope ML models trained on a sufficiently diverse database of possible SM variation, so that complex relationship between SM and predictors will be captured better and "tipping points" will be approached. In the future work, a possible solution is to apply a Land surface model, such as Common Land Model (Dai et al. 2003), to simulate large numbers of SM data and select the local bifurcations in SM variation as supplementary samples.

Comment#3: Additionally, much discussion in the Results and Discussion sections actually lacks sufficient evidence support (e.g., Lines 361-362).

Responds：We have carefully checked the whole manuscript and deleted the inappropriate discussions including line 361-362.

Comment#4: The soil moisture product has a spatial resolution of 1 km while the input data, ERA5-Land product, has a resolution of 9 km. So how did the `authors pre-reprocess the ERA5-Land data?

Responds：Thanks for your kind comments and helpful suggestions. We have described the pre-reprocessing of the ERA5-Land data as follows:

All covariates were processed to the same 1km by 1km grid system. For ERA5-Land with 9 km resolution, we resampled it into 1 km by the nearest neighbor method. For MODIS LAI with 500 m resolution, we aggregated it into 1 km by averaging.

Comment#5: They mentioned that in-situ observations were adjusted to ERA5-Land soil moisture but did not introduce the specific methodology.

Responds：For adjusting the in-situ observations to ERA5-Land soil moisture, we have added the specific methodology as follows:

In this method, we first obtained a weight by dividing the standard deviations of the in-situ SM at each station by that of ERA5-Land SM at the corresponding grid, and then multiplied the original in-situ SM by this weight. After that, we computed the difference between the average value of the in-situ SM at each station and the ERA5-Land SM at the corresponding grid, and subtract the in-situ SM by the computed difference.

Minor comments:

Comment#6: The soil moisture product ranges from 2010-2020 but this time coverage is still too short for analysis in related fields, for example, the occurrence of droughts. I am wondering why the authors chose such a target period.

Responds：Thanks for your kind comments and helpful suggestions. The in-situ measurements before 2010 may be available from China Meteorological Administration (not open to us) and the number of stations is less than 800. If we produce the SM data set without any in-situ data (or only a few hundred stations), the quality of the data may be poorer as it will be extrapolation in time. However, we agree that it is proper (assuming the relationship between SM and covariates remains the same in the last two decades) to extend the present time period to 2000-2020. We did not extent it before 2000 taking a conservative attitude. But it is possible to extend it as long as in-situ SM is available in the future. The extended data is still available at http://dx.doi.org/10.11888/Terre.tpdc.272415. We have added the following contents in section 2.4:

In addition to the period of 2010-2020 when in situ SM data are available, we also produced the gridded SM for the period of 2000-2009 when in situ SM data are unavailable, assuming that the relationship between SM and covariates remains the same in the last two decades. It is proper to deem that the data quality during 2000-2009 is poorer than that of 2010-2020.

We also list a future work in the conclusion as follows:

It is also possible to update and extent the time coverage of this data set before 2010 as long as in situ SM data becomes available.

Comment#7: In the text, the authors mentioned terms such as “Liaoning province”, “Sichuan province” and “the plateau”, which are not friendly to readers that have no the background knowledge.

Responds：Thanks for your kind comments and helpful suggestions, we have added the detailed longitude and latitude to the mentioned regions. The new expression is as follows:

Meanwhile, SMCI1.0 product often underestimated in north China and overestimated in Sichuan province (97°21'E-108°12'E, 26°03’N-34°19’N)

Additionally, air temperature had significant positive partial correlations with SM in the northwestern China, and negative partial correlations in north China and Liaoning province (118°53'E-125°46'E, 38°43’N-43°26’N) for SMCI1.0.

Qinghai province (89°35'E-103°04'E, 31°09’N-39°19’N) belongs to the tundra climate zone, where some soils are wet and other soils are dry.

In some of the plateau areas (73°19'E-104°47'E, 26°00’N-39°47’N)