Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-59-RC1, 2020
© Author(s) 2020. This work is distributed under
the Creative Commons Attribution 4.0 License.

Earth System Science Data Discussions

Interactive comment on "A new dataset of satellite observation-based global surface soil moisture covering 2003–2018" *by* Yongzhe Chen et al.

Anonymous Referee #1

Received and published: 9 June 2020

Review comments for essd-2020-59 "A new dataset of satellite observation-based global surface soil moisture covering 2003–2018 " by Chen, Y., Feng, X., and Fu, B.

The authors propose a global dataset of top (0-5cm) soil moisture with 10 day temporal and 0.1° spatial resolution, covering the period 2003-2018. The dataset was produced gradually, backward in time, through machine learning methods (neural networks) for 5 periods that correspond with the availability of 11 different passive and active satellite remote sensing soil moisture products. Besides satellite observations (starting with SMAP in step one), 9 environmental properties were fed to the neural network, and from step two on, previously modeled soil moisture was included to enable the expansion backward in time. The final product is evaluated with observations of the international soil moisture network and, in comparison to other merged products,

rated superior, however, the potential for further improvement is also emphasized. Altogether, the work seems sound and the developed method and dataset appear valuable for further scientific studies and applications. Nevertheless, some of the steps in the processing chain need further clarification and the data structure needs to be improved before the manuscript can be considered to suffice for publication.

Specific comments

Title

I suggest to remove "new" from the title, since all dataset proposed in this journal are somewhat new. You may consider to name it "combined" or "improved" or "complete" or "optimal". Have you thought about giving the product an acronym? That improves recognizability and makes it easier to reuse it in other studies and publications.

Abstract

L14: more than 10**6 not correctly displayed in the online abstract (here it reads 106)
L15: Please state also the temporal resolution (10 days)

Introduction

L32: resolution of ERA INTERIM is rather 0.75°

L34-37: I agree that these products have many shortcomings, but other than the dataset provided by the authors, the models provide also information about the deeper soil layers. This important point should not be omitted here.

L75-76: "data averaging" - what type of averaging is meant, spatial or temporal? "can hardly unify the temporal variations." Please specify what the "temporal variations" refer to. Is it the temporal variations of the different soil moisture data products?

Data an methods

Instead of Table 1 or in addition, it would be good to have a timeline figure from 2003-

2018 that shows a bar for every dataset used in the process of creating the final product, including the 11 soil moisture products, the time-varying quality impact factors and the intermediate modeling products (SIM-1T, SIM-2T,...).

L110: Specify why SMAP is mentioned as the "best product" here. Is it because of the spatial resolution, the algorithms or with respect to the in situ observations? Can you add a citation to corroborate this statement?

L143: change to "reference coordinate system"

L182: "based on the correlation between soil dielectric conductivity" - do you mean soil dielectric permittivity or or soil electric conductivity?

L186-188: "Because ..." this sentence is unclear.

L205: Figure 1 is never referenced in the manuscript. This should be done here or later at L225.

L219: Do the 140x360 zones include water (ocean) areas?

L220: A subnetwork has 100 pixels, but ("for a 0.1° pixel in a given 10-day period, if all the subnetwork inputs have valid..."), how can one pixel have more subnetworks? Please improve the formulation.

L222: What is an "individual neural network"? Is it the collective of all zonal neural networks for one simulation (SIM-T1, SIM-T2, ...)? Is the maximum possible number of subnetworks 50.400 or less because of ocean cells?

L223: For reproducibility, it is required to state exactly the MATLAB version and the toolbox version and method/function name that was used for training the neural network.

L256: "we classified all pixels" -> "we classified all 0.1° pixels", I suggest to add the resolution information that it is clear which of the different grids is addressed.

L259: Again, I thought that a pixel is the smallest unit in the process (i.e. subnetwork). So how can a pixel have a subnetwork? Not clear to me.

L261-262: "Hence, it is a ..." sentence seems incorrect. I think you should better write "neural network collocation" or "neural network constitution" to make it more clear that these are neural network realizations with identical configuration but different ingredients.

L272,815 and other occurrences: it is not clear how the 10 day periods are defined and how they relate to the ordinal numbering. A month has between 29 an 31 days, so how are the periods split and how does that affect the last 3rd where the number of days is variable? How does this variable length averaging affect the results and what are the implications for validation?

L270-292: also this section would greatly benefit from a timeline bar plot that shows all the soil moisture products and simulated models, so that the overlaps can be grasped immediately

L318: define how R**2 is computed (based on Spearman or Pearson).

L321: lower case r should be used for the correlation coefficient (based on Pearson?). Why are you mixing r and R**2 and do not use R**2 for all analyses?

L322: please provide formula for A.R computation

L326: "in all grids", grids or pixels (1 x 1 or 0.1 x 0.1°)?

L326: please provide formulas for spatial pattern validation (at least in the supplement)

Results

Figure 3: Use identical labels for the x-axis, add missing lower frame.

Figure 4: If the color key is put below the figure, the figure can be increased in the horizontal direction which leads to wider bars. You could even remove the x-axis labels

and names and leave only the lowermost. By this you can increase the size of the bars and hence the readability (reduce redundancy).

L381: How is the performance of SIM if the SMAP training period is omitted, i.e. from 2003 until 2015D01, as compared to ASCAT-SWI?

Figure 7,10: As for fig. 4 place color key below the plots and increase the bars horizontally

Discussion

Do you see any chance to improve the temporal resolution of the product in the future? If not, what are the constraints?

L499-500: Is SIM also superior to the other products if only the prior to SMAP period is considered (2003 until 2015D01)?

Are there plans to update the data-set on a regular basis?

Dataset

The dataset is organized as an archive of geotiff files. The problem with this structure is that the time identifier is only contained in the file name, but without practical formatting. If one wants to import a time series for a region or a single pixel, the data structure is quite unhandy. Also from the readme file and the metadata it is not quite clear what the 10 days ordinal numbering means exactly. Is it always the [1-9],[10-19],[20-29] or [1-10],[11-20],[21-30] periods? How are the months with variable length considered (28,29,30,31 days)? That's not clear also not from the manuscript.

Further, I would suggest to add a table (csv) that links the different file names to their specific period using ISO 8601 https://en.wikipedia.org/wiki/ISO_8601 notation:

e.g., a file named inventory.dat with a list like the following one:

Period, Filename 2003-01-01/2003-01-10, SMY2003DECA01.tif 2003-01-11/2003-01-

20, SMY2003DECA02.tif ...

Also the numbering should be formatted as %02d so that, e.g.,

SMY2003DECA1.tif becomes SMY2003DECA01.tif. This is important if one wants create a chronological file list for looping over time. With the current scheme, the order would become SMY2003DECA1.tif SMY2003DECA10.tif, SMY2003DECA11.tif, ...

This should be also applied to all tables in the manuscript (e.g., 2005D01 instead of 2005D1).

Supplement

Figure S1: The figure and description is not completely clear. I assume that every number (yellow and blue frames) is one pixel (0.1° x 0.1°)? I think it would become more clear if you superimposed a light gray mesh for the pixels over the 1x1° zones. But then, why are there 4 steps required to smooth the borders? It means that every boarder gets smoothed twice, and every corner point even four times.

Figures S5, S8, S11: put the color-key to the bottom of the figure (a single key would be sufficient for all sub-figures), you could even remove the x-axis labels and names and leave only the lowermost. By this you can increase the size of the bars and hence the readability (reduce redundancy).

Language and bibliography

There are often blanks missing between words. DOIs are completely missing in the reference list.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-59, 2020.