Earth Syst. Sci. Data Discuss.,
https://doi.org/10.5194/essd-2020-353-RC4, 2020

Open Access Earth System Science Data Discussions

Interactive comment on "SGD-SM: Generating Seamless Global Daily AMSR2 Soil Moisture Long-term Productions (2013–2019)" *by* Qiang Zhang et al.

Anonymous Referee #4

Received and published: 31 December 2020

The authors present a method and dataset to fill daily AMSR2 soil moisture product gaps with a CNN for the years 2013-2019.

The abstract jumps quickly into the topic, but is somehow ambiguous by not being clear to which current soil moisture products (". . . the acquired daily soil moisture productions") the introduction relates? It would be good to explicitly state that it relates to the AMSR2 products (not productions) . . . this lack of language clarity (e.g. "reliable cooperativity. . ." etc.) traces through the whole manuscript and needs to be strictly revised before considering acceptance. Otherwise it is really hard for the reader to understand, and thus, to estimate the usefulness of this dataset. However, independent

of the language a few other comments as follows:

Still the abstract creates more questions than answers. The evaluation measures are difficult to interpret. Why stating 2 evaluation measures, with one for original data? Also the choice of units (m3/m3) is not immediately clear as the dataset only produces percent values for soil moisture?

The introduction is not clearly introducing the AMSR2 original dataset. I think it would be of great value if at least basic technical cornerstone of the original dataset are described.

Stating that the codes are also published is misleading so far. "The related Python codes of this dataset are also available at https://github.com/qzhang95/SGD-SM." (authors) only holds an example of code of extracting data. I think it would be really helping the transparency of the data quality if the physical network implementation (TensorFlow, Keras, SKLearn, PyTorch?) would be also open in the spirit of open data and open source and reproducibility. Understandably, a trained neural network model is not 100% reproducible, but the model could be also archived on Zenodo? It only makes sense, because it would be very feasibile to update the dataset on a yearly basis with the developed model. In the end the idea of ESSD is "living data".

The year 2013 folder only contains 362 files, not 365. May 2013 only seems to have 28 files? Please check your upload on Zenodo.

(Zhang et al., 2020) This citation is not in the references list, there are only Zhang 2020a and 2020b. Please add, presumably it is your data citation: DOI: 10.5281/zenodo.396042

"More details of this work are released at https://qzhang95.github.io/Projects/Global-Daily-Seamless-AMSR2/." . . . your current paper should reflect the most important and up to date source of information until it is published. It would be ok though to refer to it as a "technical supplement" maybe? However, those URLs are not reliable. Thus, if

you have a technical supplement with more details, it could be added to your Zenodo archive (which will be reliable and has a DOI).

Section 3.1 starting on page 7 is the methodological main part of your neural network implementation. While I'd like to acknowledge that technical level of description, I have two contradicting issues with it:

1) p7 ll137-139: "This network includes 11 layers (3D partial CNN unit and ReLU (Rectified Linear Unit)) in Fig. 4. The size of 3D filters isall set as $3\times3\times3$. Number of feature maps before ten layers is fixed as 90, and the channel of feature map in the final layer is exported as 1".

That is very technical, yet, it is not clear why those dimensions where chosen. The discussion section does not discuss the CNN and the design choices at all and what effect they have. For example, how can this capture the comparatively big gap areas of the original AMSR2 dataset?

2) On the other hand, much of this could also go into a technical supplement and you could provide a much higher-level overview for the reader in the paper. The paper is the data description, and many readers and future users of the dataset will not have the technical understanding of judging or even reading through the technical low-level design of the CNN – nevertheless this still also needs to be documented.

Also, why not an LSTM type network?

Calling it spatio-temporal 3D might be misleading, as it is areal 2D and then a temporal dimension. Spatio-temporal indicates that already, the added 3D might lead to think of spatial 3D plus time.

What was the reason to chose the 0.25 dec degrees as spacing for the data files?

Last but not least, I'd like to advocate for a bit more metadata in the netcdf files, because netcdf provides great means for metadata. For example, you could adhere a bit more to the NetCDF-CF conventions, or at least add e.g. attributes such as title,

reference and a time stamp in the dataset, not relying on the filename for example. You could also join at least the yearly slices into a "cube" that follows conventions of the Earth Sciences community (e.g. longitude instead of lon as variable name). Also the Zenodo deposit could have more fields filled out for improved discovery, more keywords (e.g. "soil moisture"?) and terms from controlled vocabularies, such as GEMET (https://www.eea.europa.eu/help/glossary/gemet-environmental-thesaurus) or similar.

I think there is a lot of potential in this paper and dataset (even if "only" 6 years length). However, many technical aspects make it hard to grasp the content and estimate the quality in the first place.

Interactive comment on Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2020-353, 2020.