REVIEW

Manuscript ID: hess-2023-33

Title: Evaluation of model-derived root-zone soil moisture over the Huai river basin

Authors: Liu, E., Y. Zhu, J.-C. Calvet, et al.

The paper evaluates 7 global root-zone soil moisture products and one regional product over the Huai river basin in China for the period from April 2015 to March 2020. The products are evaluated against each other and against in situ measurements from 58 sensor profiles. The differences in the products and their performance against in situ measurements is related to the skill of the products' forcing inputs (precipitation, air temperature), model structure, and model parameters (soil texture).

The authors find that the GLDAS_CLSM product performs best overall and has the highest R and lowest ubRMSE values. Root-zone soil moisture is overestimated w.r.t. the in situ measurements in the 7 products based on land surface modeling, which is traced to an overestimation of the precipitation inputs, too much drizzle, underestimation of the air temperature inputs, and errors in the prescribed soil texture.

The paper addresses a topic of interest to HESS readers and presents some important results about the skill of several widely-used global root-zone soil moisture products and one regional product. Unfortunately, the paper falls short in several aspects, as described in the major comments below. The paper *may* be suitable for publication in HESS after major revisions. However, given the gaps and errors in the submitted manuscript, it is more realistic to REJECT the paper at this time and encourage resubmission later if the authors are willing and able to address the shortcomings. This gives the authors a better chance to produce a revised manuscript of sufficient quality without being constrained by the customary 2-month period for major revisions.

Major comments:

1) There are clear errors in the authors' description of the product characteristics that impact the interpretation of the results.
   Lines 26-27, also Lines 592-593: "[MERRA-2 and SMAP L4] are based on the same LSM and on the same surface meteorological forcing…". While both products use the Catchment model, and both use atmospheric forcing data generated with GEOS, the model versions and the forcing data are not the same. MERRA-2 uses an older version of the Catchment model than SMAP L4 (Reichle et al. 2017, 10.1175/JHM-D-17-0063.1; Reichle et al. 2019), including (but not limited to) the difference that soil parameters are based on FAO data in MERRA-2 and on HWSD in SMAP L4. Re. surface meteorological forcing, MERRA-2 uses an older GEOS version 5.12 at 0.5-deg resolution, whereas the forcing data in SMAP L4 is from the GEOS weather analysis ("Forward Processing", or FP) system with evolving versions 5.17 to 5.25 and at 0.25-degree resolution. In MERRA-2, the CPCU precipitation is used in its native climatology whereas in SMAP L4 the CPCU precipitation is rescaled to an independent reference climatology.

These differences are not discussed in the paper but are critical for the interpretation of the results, especially those re. errors in the products' soil parameters and the cause of the RZSM bias. For example, Figure 8 shows that SMAP L4 and MERRA-2 precipitation metrics vs in situ measurements are different, but the authors simply ignore this result and keep repeating that SMAP L4 and MERRA-2 have the same forcing data (e.g., Line 592). The fact that MERRA-2 also uses CMAP precipitation inputs is irrelevant (Line 493) because CMAP is used in MERRA-2 only over Africa and the ocean.

I outlined the above differences between MERRA-2 and SMAP L4 in detail because I am most familiar with these products. I did not exhaustively review the accuracy of the other products' characteristics. Descriptions of the other products may or may not contain additional errors.

2) Besides the errors in the product descriptions outlined above, there is insufficient information in the description of the products and analysis methods:
    a. Section 2.3 is very uneven across the subsections (individual product descriptions) in terms of the level of detail provided. E.g., in some cases the horizontal resolution of the product is not provided.
    b. For MERRA-2 the native 0.5-deg data is interpolated to 0.25 degree, but there is no comparable statement for SMAP L4. Are the SMAP L4 data used in their native (9-km) resolution? Aggregated to 0.25 degree?
    c. ERA5 includes a soil moisture and screen-level temperature and humidity analysis. This analysis clearly impacts the ERA5 soil moisture estimates but is not mentioned even once in the paper.
    d. Much of the text in the subsections of section 2.3 appears copied from the blurb of the descriptions on the products' websites. In some cases, the text includes to the motivation of products' development from the preceding version. E.g., MERRA-2 is described in reference to the long-obsolete original MERRA data (Lines 159-163), and NCEP CFS v2.2 is described in reference to NCEP R1 and R2 (~Line 176). The product descriptions in this paper should focus on each product's characteristics and how they differ from the other products examined here, not on how the products differ from older versions that are not examined here.
    e. The information on the specific air and soil temperature variables provided in the manuscript is not sufficient. What are the air temperatures that are used in the comparison? Are they at the 2-m screen level (T2m) or at the atmospheric model's lowest level? The SMAP L4 product only provides the latter, which is used to drive the Catchment model.
    f. As for the simulated soil temperature, what is the layer depth for each product? And what exactly is the in situ soil temperature shown in Fig 9? The strong short-term variability of the in situ soil temperature (after averaging across 58 stations!) is highly suspect. Could this be the surface (or skin) temperature? The comparison here is almost certainly one of apples vs. oranges.
    g. What is the spatial resolution at which the temperature is validated?
    h. The in situ soil moisture measurements are taken at 8am local time (Line 109), but what about the in situ soil temperature measurements? If the in situ soil temperature is taken at 8 am local time, are the soil temperatures from the products (Fig 9) also

sampled at 8am local time? Are daily average soil temperatures from the products used?

i. It is not clear at what time step the second-order skill metrics (R, ubRMSE) are computed. Are the metrics computed at daily time steps after aggregating all sub-daily products to daily time steps?

j. The in situ soil moisture data are QC'd (Lines 111-112) but there is no information on how many data points are actually used at each station. The footnote of Table 2 suggests that 1827 (daily?) values are used at each of the 58 stations, but then no data would have been excluded by the QC process. This is contradictory.

3) Many references are plainly incorrect or inappropriate, or are missing altogether:
   a. Many of the references used for the products are not the first-hand references. Rather, the references used are simply about applications of the data product. This is not acceptable. The authors must use first-hand references for the data products examined. In other cases, inappropriate references are cited, or references are simply wrong. Here are some of the problematic references (not meant to be a complete list):
      i. Bi et al. 2016 is used in several places as a reference for GLDAS product characteristics but the paper is not a first-hand description of the GLDAS products.
      ii. Line 51: The main ERA5 reference is Hersbach et al. QJRMS 2020.
      iii. Line 46: Rienecker et al 2008 is not an appropriate reference for SMAP L4.
      iv. Line 56: Reichle et al. 2017 is not an appropriate reference for NCEP CFSv2.
      v. Line 80: Koster et al. 2016 is really McCarty et al. 2016.
      vi. Lines 83-84: Need a reference for this statement about "layer stratification" (and also edit the statement for clarity); also in Lines 479-480.
   b. There is no reference for the HWSD soil data!
   c. References by the same lead author written in the same year are not distinguished properly. E.g., there are two papers by Wang et al. 2016 and three papers by Reichle et al. 2017 in the list of references. In the text, they are universally referred to as Wang et al. 2016 or Reichle et al. 2017, leaving the reader guessing which paper the authors refer to in each instance.
   d. Data products such as SMAP L4 that have a digital object identifier must be cited with proper references that include the DOI.

4) The version of some of the products examined here are unclear. E.g., what SMAP L4 version is used in the paper? This has important implications on the interpretation of the results. The precipitation forcing data in SMAP L4 versions 5 and 6 differs considerably, which is well documented in the products' validation reports (disseminated by NSDIC along with the data).

5) Like SMAP L4 and CLSM, the GLDAS_CLSM product also uses the Catchment model and precipitation forcing based on observations, but there are clear differences in the skill metrics between the three products. One distinguishing feature of GLDAS_CLSM is the assimilation of GRACE TWS observations, which inherently contain information about root-zone soil moisture. There is no discussion in the text on what the impact of the GRACE data assimilation may be on

the skill of GLDAS_CLSM vs. the other two Catchment-based products. (By the way, the Catchment model versions in GLDAS_CLSM and MERRA-2 are different but closer to each other than to the version used in SMAP L4.)

6) There is no reference for the in situ measurements other than that "data are available upon request". This should not be acceptable to HESS. For the study to be reproducible, the in situ measurements used here must be made readily accessible to interested researchers.

7) The discussion in section 5.4 is rambling and does not add useful information. All land models underpinning the products are "1-dimensional" models in the sense that grid cells (or computation elements) are not coupled horizontally other than through the forcing data. The authors make a mess of this simple fact by selectively pointing this out for some products (e.g., for CSLM in Line 506 and for HTESSEL in Line 511) but not others (e.g., Noah, Lines 509-510). The entire section is worthless and should be deleted.

Minor comments:

1) English style and grammar need some attention. Overall, the paper is readable but there are occasional English language errors.

2) The graphics generally lack clarity and quality.
    a. The axes labels (and other text within the graphics) are often too small to be readable (e.g., Figs 3 and 11)
    b. Color/symbols are sometimes difficult to distinguish (e.g., Figs 2 and 9)
    c. The resolution of the images embedded in the pdf is generally poor (e.g., Figs 2, 5, 7)

3) Inconsistent use of product names ("SMAP" vs "SMAP L4") (e.g., lines 322-323; Fig 7h label)

4) Line 337/Figure 3: Is the "nonoutlier minimum Q1 – 1.5 x (Q3 – Q1)" meant to reflect the "whiskers" of the box plot? If so, I don't see how this can be correct. Same for the "nonoutlier maximum" (Line 338)

5) Line 21: "The underestimated SMOS L3 SSM associated with low physical temperature triggers…" This is unclear. Do you mean "The underestimation of SMOS L3 SSM associated with…", or, perhaps more clearly, "The underestimation of the SMOS L3 SSM during cold conditions…" ???

6) Table 3 does not need a column for "Soil layer" because it is the same for each product.

7) Table 3 needs units for the bias. m3/m3 ??

8) Table 3: What is the "Bias (anomaly)" in the final column? Isn't this zero by construction?

9) Line 370: “histograms of normalized RZSM”: How was RZSM normalized? I could not find this information.

10) Lines 101-103: The average annual precipitation in the Huai basin is listed as 888 mm precip and the average annual evaporation is listed as ranging from 900 to 1500 mm. A relatively small fraction of the Huai basin is irrigated. How can the average evaporation exceed precipitation by that much?