Earth Syst. Sci. Data Discuss., referee comment RC2
https://doi.org/10.5194/essd-2021-456-RC2, 2022

Comment on essd-2021-456

Anonymous Referee #2

Referee comment on "CAMELE: Collocation-Analyzed Multisource Ensembled Land Evapotranspiration Data" by Changming Li et al., Earth Syst. Sci. Data Discuss., https://doi.org/10.5194/essd-2021-456-RC2, 2022

This paper presents a Land ET product that has been generated by merging multiple ET data sets using different collocation-based appraoches. While such a product would certainly be of great interest to the community, I have various major concerns about the methodology and the evaluation approach.

General comments:

- My biggest concern is the brute-force nature of the approach. Various collocation approaches are thrown blindly at various products with no regard given to the properties of either the products or the methods (see specific comment to L248). It seems that all possible combinations are applied and averaged, and then a selection is made (Supplement 5) without further justification or demonstration of relative performance (see below). Why selecting exactly these combinations of products and methods in these periods? What were the criteria to deem these best-performing?

- Much related to this comment: All the employed collocation approaches are very sensitive to error cross-correlations. While some variants tolerate/estimate cross-correlation, they typically require the assumption that at least some product errors are uncorrleated. Notwithstanding, the authors seem to just apply QC to all combinations for all possible cross-correlation scenarious and then just average the results, which most likely fails terribly. This is because in all cases, cross-correlation estimates will be biased, because the aforementioned assumption will be violated in either case.

A proper application of such methods would require careful consideration of the product properties. For example: If four products are considered, only two of them are allowed to exhibit non-zero error cross-correlation. QC can be applied accordingly to estimate error variances of each of the four products as well as this one error cross-covariance, but exactly which errors are correlated has to be chosen a priori. Unfortunately, if more than

two products exhibit correlated errors, or if the wrong data pair is assumed to have correlated errors, the whole thing breaks down and both error variance and error covariance estimates will be strongly biased. Consequently, the merging weights will also be strongly biased.

I think there's very good reason to expect strong error correlations between many products. For example, FLUXCOM is using ERA temperature for conversion, and PMLV2 uses GLDAS as an input. What about forcing data of ERA5 and GLDAS? I know that at least soil moisture simulations from ERA and GLDAS have highly correlated errors in many regions, so I don't think it will be any different for ET. Testing this could possibly be done by selecting triplets with supposedly uncorrelated errors, estimating erorror variances, then replacing one product, and assessing whether the error variance estimates remain unchanged.

- The description of the merging methodology in Sec. 3.2 is very unclear. In L281, omega is called optimal weight, even though there is never just "omega", only omega_ij, which appears to be the weight when using two data sets only. Later, in L286, omega_i is introduced as "arithmetic mean for each product", yet the equations claculate arithmetic means between weights (of data pairs), not between products. So, if I understand correctly, the authors calculate a weighted average between products, where the weights are calculated as unweighted averages between weights of data pairs that do account for error cross-correlations.

I don't know where this comes from (not from the afore-cited Kim et al. (2020), and I cannot access Bates and Granger (1969)), but I'm fairly certain that this is not a valid least-squares solution. A least squares solution for an arbitrary number of products is provided, for example, in Eq. (2) of Gruber et al. (2019). This requires to take into account the cross correlation between all products in a properly constructed error covariance matrix.

Is the described approach, perhaps, to account for the different possible implementations of the various method, e.g., the 30 possible options to implement quadruple collocation? For the reason stated above, I don't believe that this would be valid, and most likely does more harm than good.

- The issue of bias is left entirely undiscussed. The method of least squares minimizes the random error variance, but doing so requires the data to be free of bias. Gruber et al. (2019) attain this by rescaling (which is only one possibility). However, as evident from e.g., Figure 11, bias is certainly present and will have a large impact on the merging. This is a problem because relative weights are calculated from random error variances and disregard biases altogether. However, when applied in the merging, they are also used to weight the biases by the same amount. Therefore, the fact that CAMELE follows FLUXNET so closely in Figure 11 appears, in my opinion, mostly serendipitious, possibly because it just so happens that - during this period - weights are fairly evenly distributed across products. In other periods, things would look very different because in the other merging periods, much more weight is put on ERA5 (see Supplement 5). I believe this is also the

reason why results appear best in the KGE, because the KGE puts a much higher weight on the contribution of bias than do the other performance metrics.

- Related to the previous comment: The validation is insufficient and does not justify the selection of products and collocation strategies as shown in Table 2. No performance metrics are shown other than KGE statistics. How do the individual available input products perform in the different periods where data are available? How do merged products using the different collocation methods perform relative to one another, and to the performance of the individual input products? Most importantly: How would a simple unweighted average perform? For the above-described reasons, I suspect that the proposed approach cannot estimate relative weights accurately enough to outperform an unweighted average. All these aspects should be evaluated and shown separately for bias and for correlation characteristics. Least squares merging aims at improving the latter, while the largest impact appears to be in the former (which is, in fact, often found for model ensemble averages, because their bias seems to scatter rather randomly around the truth, hence averaging tends to improve that, especially in an unweighted case). Lumping the effect of bias and correlation together in the KGE actually hampers a proper assessment of the impact of the merging algorithm.

- Supplements are not referenced properly. S3 is quite unclear.

Specific comments:

L31: What about superiority / inferiority w.r.t. all the others? Why only mention one (second-best), and then only KGE?

L33: should this be "inconsistent"?

L43: Rephrase "As the intermediate variable of soil moisture affecting air temperature"

L61: I would strongly disagree with this statement. SA is arguably the best bet if weights cannot be estimated accurately. In other words, unweighted averages often outperform badly weighted averages, and this is observed across disciplines. The authors actually point this out in L65.

L79: should be: "Su et al. (2014) proposed..." Check citation style throughout the document (The same error happens again several times in the lines that follow as well as lateron).

L83: Gruber et al. (2016) doesn't propose "quadruple collocation" in particular, they propose collocation with an arbitrary number of n>3 data sets, referred to as extended collocation, and only demonstrate it for the case of n=4 as an example.

L125: Change to "more elaborate descriptions"

Sec 2: I'd be good to be very clear about the input of all the employed models, especially to understand potential error cross-correlations. Which RS data are used for FLUXCOM?

L238-240: The log-transformed multiplicative error model has been preferred for precipitation products, because they are assumed to exhibit a multiplicative error structure. This is not the case for other variables such as soil moisture, where the additive structure is indeed more common (and arguably more appropriate). Is there any good rationale for which to assume for the ET products used in this study?

L248--: This is a mere repetition of the introduction that doesn't provide any understanding of the respective methods other than how many data sets are needed. I think the readers could benefit greatly from a more thorough explanation / illustration of the differences between these approaches. What are their strengths, limitations, and assumptions? How do these relate to the properties of the products used in this study? Which would you expect to perform how? (The supplement provides mere mathematical derivations, but no insight into the properties / differences between methods.)

Table 2: Does this selection of products/methods during different merging periods emerge from the validation? If so, I think it'd be better to make this part of the results section alongside the validation of the different approaches. This is necessary to actually justify this selection.

L314: How's a standard deviation a validation metric? Is there any reason to believe that a low SD equates "better"? Also, no SDs are ever shown.

L324: Bootstrapping cannot improve uncertainty, it can only provide confidence intervals, which is not done here.

L325: How (and why, see above) was a multiplicative error model used? L331 shows additive errors.

L328: Do you mean "Poisson distribtion"? Does ET generally follow such a distribution?

(I'm not an ET expert, so I don't know). The referenced Kim et al. (2020) used a uniform distribution, but I believe that doesn't tell much anyway other than a sanity check.

Figure 2: I have the feeling there's something fishy about the synthetic experiments. For example: Why would delta_rho increase with sample size? Isn't a lower number better, i.e., closer to the truth? Also, why should there be discontinuities in the bottom two panes?

L428: Do the authors mean "less influenced by antecedent conditions"? This would, in fact, be a problem, because lagged TC approaches REQUIRE the variable itself to be highly auto-correlated while ERRORS should be temporally uncorrelated.

Table 4: I don't understand what is shown. The description says "Correlations against in situ", but why the columns for the different input products? And which products are actually being merged? All of them in all possible combinations?

L473: I'd recommend scaling the axis, not the values themselves.

Figure 5: I don't understand what is shown. What does it mean to compare an additive and multiplicative errors structure? This hasn't been properly described in the Methods section. Also, the figure is very busy and hard to read. Also, what's meant with RMSE_TCA? Is TCA again used to evaluate results, after first using it for merging?

L505: Not sure why the results section starts here... A lot of results are already shown before that.

Figure 12 is the same as Figure 11 (caption seems to be correct, but the figure seems to be wrong).

Figures 13-14: Hard to compare visually... Would it make sense to show difference maps?

Figures 15-16: How confident are you that trends aren't introduced by the merging algorithm? I understand from Table 2 that different products/methods are used in different periods. This could introduce trends just by having jumps in the data, especially if no bias correction is applied (see general comments).

Data repository:

- The description should be a description of the data to make it easier for people to understand/use them, not a mere copy of the abstract of the manuscript.

- I couldn't open the data in panoply because "Axis includes NaN value(s)". This seems to be the case for all 3 dimensions. Please fix the data files so that dimensions include only valid data.

References:

Gruber, A., Scanlon, T., van der Schalie, R., Wagner, W., and Dorigo, W.: Evolution of the ESA CCI Soil Moisture climate data records and their underlying merging methodology, Earth Syst. Sci. Data, 11, 717–739, https://doi.org/10.5194/essd-11-717-2019, 2019.