Hydrol. Earth Syst. Sci.
hess-2019-103

Dear Reviewer #1,
Thank you for your review and the detailed comments. Following please find our point by point response to your suggestions and questions. The Reviewer's comments are in regular font and our response is in bold.

Response to Referee #1
The manuscript presents a global-scale assessment of the effects (i.e., bias) of large-scale averaging of atmospheric forcing, namely P and PET, on evapotranspiration rates. The assessment relies on the application of a previously published approach based on the Budyko framework and using existing datasets of P and PET. Additional analyses are presented at the continental- (i.e., CONUS) and regional-scale (i.e., Switzerland) in order to explain the role of climate and the scaling of evapotranspiration bias with the grid resolution, respectively.

The work addresses an important subject of research (i.e., how to quantify averaging effects in large-scale models), which is certainly of relevance for HESS readership and for a broad scientific community. However, in its current form the work appears too much a mere application without offering new insights on such a relevant subject. In my opinion, this makes the contribution (after substantial revisions) more suitable for a technical note. I provide below a list of points that support my evaluation.

We thank the reviewer for his/her interest in this work. We leave it up to the Editor to determine whether it would be more appropriate for this work to appear as a technical note or a research article.

1. Authors motivate the need of their work (i.e., quantification of averaging effects on evapotranspiration) and discuss their results in light of well-known ESMs simplifications that do not take into account the fine-scale spatial heterogeneity in the atmospheric forcing and land surface characteristics. However, the averaging effects are assessed under steady-state conditions and neglecting non-linear land surface processes, two assumptions that do not reflect the actual way ESMs (and large-scale integrated models) are implemented. While authors clearly disclose these limitations (see lines 225-226; lines 229-230; lines 271-274 lines 361-364, etc: : :), my impression is that at the end of the article the reader is left to wonder about the maturity of the work. For instance it is not clear how current results can be exploited to calculate bias correction factors for ESMs simulations. Sentences as those reported at lines 363-364 (": : : we are now working to quantify aggregation bias in ET: : :) support somehow my impression that reported results are in a sort of "intermediate" stage.

Our research in this area is indeed continuing, in ways that build on the present paper. We nonetheless think it is useful because it demonstrates, at global scale, an overall framework for estimating how averaging over heterogeneity in atmospheric forcing at the land surface affects evapotranspiration estimates. We use Budyko framework even though it is not actually used in ESM's, because it is a simple "see through" test case for quantifying these biases. These results are clearly not directly transferrable as estimates of the averaging biases in ESM's – which, indeed, we suspect are substantially larger, for reasons that we explain in the paper.

We state in the paper that the current results can not be directly exploited by ESMs to correct for averaging bias, although the proposed methodology sheds light on the potential

ways one can account for this bias (depending on what ET calculation algorithm ESMs or land surface models use and the scales they average over sub-grid heterogeneities) (line 374-382). We will try to make the revised paper even more explicit on this point.

We agree that it would be more informative to estimate the averaging bias for physically based, time-varying ET calculations as used in ESM's. Our paper highlights a general methodology that can be used to estimate the systematic bias due to averaging, but not the precise magnitude of this bias because it will change significantly depending on ET formulation used.

2. In a similar vein to the previous point, one of the main conclusions is that from 'an atmospheric perspective' averaging is for sure an approximation in those regions characterized by strong topographic gradients. This is probably not new and authors could have made an effort to explain better the different degree of sensitivity between P and PET.

Equations 7 and 8 of Rouholahnejad-Freund and Kirchner (2017) show that averaging biases in Budyko ET estimates are equally sensitive to the same <u>percentage</u> variability in P and PET. Thus we do not try to explain the different degree of sensitivity, per se, between P and PET (because, at least in percentage terms, these sensitivities are not different in the Budyko approach). If P is more variable than PET (in percentage terms) across landscapes, then the variability in P will make a larger contribution to the averaging bias. (At least in the Budyko approach; whether this is true for more physically based ET estimates remains to be seen.)

3. Another point of the work (summarized in Figure 4) is that using different datasets we end up with different averaging effects. Again, was this not expected? Could authors provide an explanation of the different degree of sensitivity between P and PET datasets? If you do not provide any insight on this how can you claim (in the abstract) that your work discusses the underlying mechanisms of such differences? I have the same concern for Figure 5. The ordering (see lines 275-276) is really not informative.

Briefly, in figure 4 the choice of P datasets is more consequential than the choice of PET datasets, because the two P datasets differ more from one another (specifically, the percentage variability in P is substantially bigger in PRISM than in WorldClim), whereas the two PET datasets are about equally variable. If desired, we can include a discussion of these points in the revised manuscript (although this will probably require also revisiting the mathematical development of the earlier paper, and thus would add somewhat to the length).

4. According to Figure 5 it seems that when estimating ET the spatial heterogeneity matters for a minor portion of CONUS domain and when using certain datasets. Is this realistic? These results seem in conflict with efforts (of several groups) that have emphasized the need of including lateral moisture distribution in ESM simulations. Including this lateral effects may completely change the "picture", am I wrong? This concern brings me to the introduction (lines 145-146) when authors highlight the need of a general framework for systematically quantifying biases in ET estimates due to spatial averaging. Is it the case at the current stage of the work?

We treated lateral transfers in some detail in our 2017 paper and we currently have nothing to add on this subject. Lateral transfers may of course be important, but unless and until we have reliable quantitative estimates of how big lateral transfer fluxes actually are (and where they are), it will be difficult to estimate their impact on ET heterogeneity biases.

5. The discussion around Figure 6 is not clear. If I compare at single grid points the color scale at 1/32, 1/16, 1/8 and so on with the ones at 1 and 2, I see that the magnitude of the bias is not increasing, isn't? Further, the bias extends over larger areas because you're increasing the resolution, am I wrong? In any case, here it is important to implement some statistics that accurately quantify the scaling of the bias with the grid resolution.

These inferences are correct, and are stated in the text and the figure caption (although in different language). We will try to make the paper even more explicit on these points. We can add a panel to this figure that plots the mean error as a function of the grid resolution.

Specific points: - The title is a bit misleading because you're accounting just for land surface heterogeneity related to P and PET.

Potentially can change to: Global assessment of how averaging over heterogeneity in precipitation and potential evapotranspiration affects modeled evapotranspiration rates

- Key points cannot be defined as long and multiple sentences

Key points will be shortened in the revised manuscript.

- In the abstract you cannot make a long discussion of a previous work findings. Please revise.

We will streamline the abstract in the revised manuscript.

- Line 70: What do you mean with "grid-averaged land surface parameterizations"?

To our understanding, the authors used parameter values that correspond to the average of the given property over each grid cell.

- Paragraph between lines 99-108 appears disconnected from the main flow of the introduction.

The paragraph is on the application of remotely sensed data in estimating ET at several scales. We will try to make the connection clearer.

- Lines 126-128: The issue of spatial averaging is also due to the fact that atmospheric and land surface components of ESMs "work" at different resolutions. In order words, it is not " just" a problem of data volume and computational limitations.

We can bring this point into the introduction of the revised manuscript.

- Line 122: what do you mean with "likely" magnitude?

This will be revised to "potential".

- In Eq. 1 "n" is not defined. How is it estimated?

"n" is a dimensionless parameter that modifies the partitioning of P between E and Q and was assumed to have a value of n=2 in our analysis (literature value). This will be added to the manuscript.

- How PET is calculated in the two datasets? This point has to be discussed in order to provide additional information about possible discrepancy between the many existing approaches to calculate PET

The Worldclim PET dataset (Hijmans et al., 2005) is based on the Hargreaves method (Hargreaves and Allen 2003). The MODIS PET product (Mu et al, 2007) is based on Penman–Monteith equation. The text will be revised to include this information.

- Figure 3e and Figure 4: How did you calculate the percentage averaging error?

Figure 3e is the approximated averaging error calculated using Eq. 3. This is written in the figure caption and will be added to the main text too. The estimates of biases obtained by direct calculation of average ET from finer resolution data were functionally equivalent to those obtained by Eq. 3 (R2=0.97), as we state on lines 214-217

- Lines 269-270: Saying that "perhaps due to the sharp gradient: : :" is a weak statement that conveys the impression that you're not really sure about the interpretation of the results.

We will revise this.