Review of "Implications of Model Selection: A Comparison of Publicly Available, CONUS-Extent Hydrologic Component Estimates" by Saxe et al.

This study compares the main components of the terrestrial water balance over the CONUS using a range of data products and model simulations. Not only do the authors provide a review of publicly available datasets, they also compare and cross-evaluate them. They use a range of approaches and indices (including innovative metrics like a measure of unalikeability) to explore the variability across these datasets and the resulting uncertainties in the water balance components. They show that significant gaps in the water balance can exist depending on which datasets are relied upon and they stress the importance of adopting a multi-model approach.

Overall, the analysis is thorough, pertinent and timely. The text is well written and the figures are carefully crafted. It is a challenge to summarise the results stemming from this variety of data sets, states/fluxes, time scales and hydroclimates, so some compromises have been made (e.g. seasonal variations are not explored and some ecoregions are quite large), but overall I support the decisions made by the authors. This study provides a rather comprehensive overview and comparison of publicly available and hydrologically relevant datasets, and hence constitutes a useful resource for future studies in the US and beyond. It fits within the scope of HESS. I congratulate the authors for this impressive data collection and analysis effort.

My comments below are mostly clarification requests and suggestions.

Major comments

CV I: The authors use the coefficient of variation (CV) to measure the variability among the different data products/models for each component of the water balance (Eq. 1). But while CV is provided for two 15-year periods, I understand that it is computed for each water year (L263), so it can potentially be used to quantify the inter-annual variability of each component (as suggested by the caption of Figure 2). Can the authors clarify if their intention with CV is to compute the inter-model spread and/or the inter-annual variability? If it is the former only, they might want to refer to the variability measured by CV as "uncertainty" to avoid confusion with (temporal) variability. Can they also clarify if/how CV values for individual water years were aggregated over 15 years?

CV II: One might argue that using CVs inflates the importance of (comparatively small) absolute differences between data products (e.g. in R, SWE and RZSM). For instance, the contribution of SWE [in mm] to the total water balance is small over the US (Fig 3d), and so is the SD among the SWE products (Fig 2a), yet the CV is quite large (Fig 2b). So, when it comes to reducing uncertainties and closing the water balance, should we first tackle components with large uncertainties (but potentially modest contribution to the water balance) or components with potentially smaller uncertainties but a larger contribution to the water balance? This is important when making recommendations in the abstract and conclusions. I see the value of the analysis as it currently stands, but to get a better sense of the uncertainty in each component with respect to the total water balance, it might be worth completing the analysis by computing a modified version of CV, using total input P as denominator

(similarly to Eq. 7) instead of x bar (Eq. 1). Showing these new results using a figure similar to Figure 5 could put the differences between datasets in the wider context of the water balance and possibly provide insights into the reasons behind the residuals shown in Figure 8. To some extent, this is already achieved by Figure A2.1, but the different units and ranges covered by the x-axes make component comparison difficult.

Sample size: What is the influence of the sample size on the CV estimates? There are fewer RZSM estimates than P, AET, R and SWE estimates, can this influence the conclusions? Can this be estimated by boostraping, e.g. by resampling with replacement using the same sample size for all the water balance components? This uncertainty could be shown in Figure 2.

Soil moisture: I agree that differences in RZSM can be related to the soil depth used by different models (L337), and in this respect, differences between models are less surprising than for the other state variables and fluxes. Still an important result, as hydrologists using a single model are likely to find it useful to know how their model compares to others. But why “rootzone” soil moisture (defined on L17 and used throughout the study) and not soil moisture? Is it because SM was computed only over the depth accessible by roots, and not over the whole soil thickness? Is the root depth/soil depth distinction made by all the models considered? Also, please comment on the soil depth actually covered by remote sensed products. Is it correct to say that the whole “rootzone” is covered, or is it less, which could bias soil moisture estimates and partially explain low correlations with hydrological model estimates (Figs. 7f and 7g)?

Water balance residuals: changes in groundwater storage are not quantified in this study, and instead, they are lumped into the water balance residual. The authors mention Gravity Recovery and Climate Experiment (GRACE) in the discussion but did not use GRACE data although it would have provided a first estimate of changes in terrestrial water storage and enabled them to go beyond soil moisture. Why was this key part of the water balance not characterised?

Eco-regions: please clarify why the Environmental Protection Agency Ecoregions were used instead of a classification more focussed on hydrology (e.g. Sawicz et al., 2011; Berghuijs et al, 2014; Knoben et al., 2018)?

Figure 7 provides an elegant overview of the correspondence between the remote-sensed data and the other datasets. Any ideas why MOD16-A2 and SSEBop correlate particularly poorly with the other datasets in the N. Amer. Desert and Medit. CA, respectively?

Uncertainties in all components of the terrestrial water balance exist and their existence is well documented (here and by others before). And many datasets are now publicly available (as highlighted here). Yet, studies that account for these uncertainties (e.g. using a range of data products) are still rare. What would you recommend to make such an approach more systematic? Can you suggest a way to select a subset of data sets for users who can’t download and process all the datasets used here? Please also see my note at the end of this document.

Minor comments

L193: Currently there seems to be no section 2.

L608-610: What does the range provided for each variable correspond to? I assume there are retrieved from Figure 5. Please also clarify this in the abstract (L23-25)

Figure 2a: Is the SD for RZSMV really close to 0 (perfect agreement between models)?

Figure 7: maybe clarify whether each circle/square corresponds to one dataset and whether these datasets are sorted based on rho (and hence their order changes from one panel to the next).

Figure 8: “Histograms of eight ecoregion water budget relative residuals”, maybe clarify that it is relative to the total precipitation (Eq 7).

Figure A2.5 to A2.9: I suggest

- removing caption mentions to colours not used in the plot (e.g. green in A2.5)
- clarifying whether the boxplots are made using significant tau values only
- clarifying whether the dots represents significant tau values
- re-evaluating whether box plots are well-adapted to show 1 or 2 values

References

Berghuijs, W. R., Sivapalan, M., Woods, R. A. and Savenije, H. H. G.: Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales, Water Resour. Res., 50, 5638–5661, doi:10.1002/2014WR015692, 2014.

Knoben, W. J. M., Woods, R. A. and Freer, J. E.: A Quantitative Hydrological Climate Classification Evaluated with Independent Streamflow Data, Water Resour. Res., 54, doi:10.1029/2018WR022913, 2018.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. a. and Carrillo, G.: Catchment classification: Empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15(9), 2895–2911, doi:10.5194/hess-15-2895-2011, 2011.

Note to the authors

I co-created the CAMELS dataset to facilitate access to hydrometeorological time series and landscape attributes of 671 CONUS catchments (see https://hess.copernicus.org/articles/19/209/2015/ and https://hess.copernicus.org/articles/21/5293/2017/). CAMELS is a community resource and we are looking for ways to expand it. Reviewing your manuscript, I thought that it would be great if time series for the CAMELS catchments could be extracted from the data products you downloaded and processed, and made publicly available. It would remove many barriers currently preventing water scientists and

practitioners from characterising uncertainties in the water balance. Feel free to contact me to discuss this further. Of course, your decision will not influence my review of the revised version of this study.

Best regards,

Nans Addor - University of Exeter, UK