How well are we able to close the water budget at the global scale?, F. Lehmann, B. D. Vishwakarma, J. Bamber

Reply to referee 1

We thank Christof Lorenz for his detailed comments on our manuscript. Our reply is attached below.

Major comments

Using the GURN-dataset as the only source of information for runoff is a bit disappointing. […] , if possible, it would make the study even more comprehensive and convincing if the authors also include a similar evaluation with observed runoff at least for a subset of study basins.

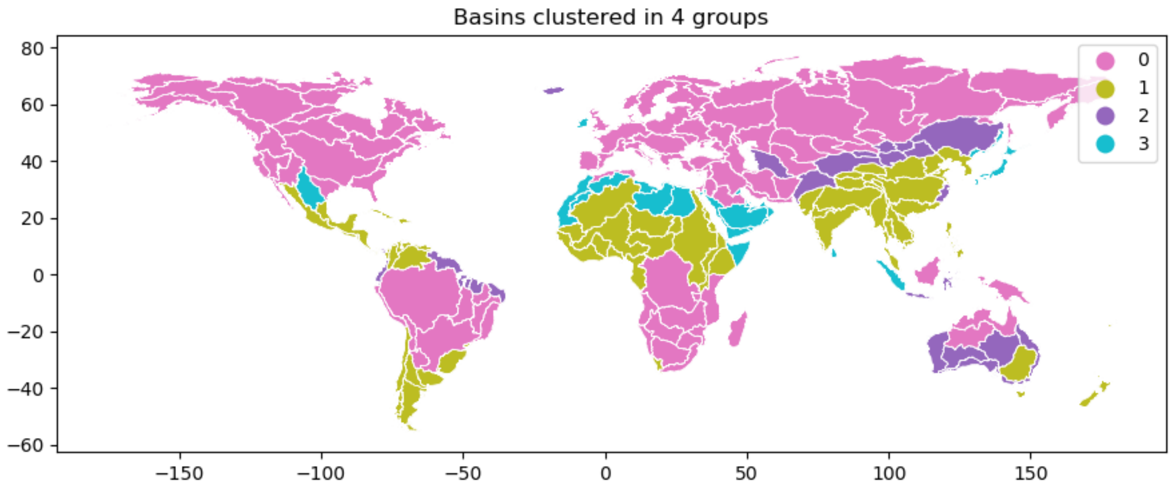
The reason for using GRUN was to have temporally consistent time-series for all basins. However, we agree that it does not represent real variations. We selected basins with a discharge area similar to the boundaries of the basins we studied and computed the runoff time-series from GRDC measurements. We combined these time-series with P and ET data to obtain 154 combinations. The table below shows that including runoff measurements improved the water budget closure in this selection of basins. The maximum NSE and cyclostationary NSE were computed over the 154 combinations containing GRDC measurements. The values were compared with the maximum NSE and cyclostationary NSE over the 1694 combinations used in the manuscript (with R from models).

	max NSE	max NSE GRDC	max NSEc	max NSEc GRDC	months with available runoff data (%)
AMAZON	0.907397	0.980957	-1.490078	0.487929	100.0
AMUR	0.586040	0.898816	0.459075	0.867783	31.9
CONGO	0.860302	0.867065	0.178806	0.218559	66.0
DANUBE	0.903778	NaN	0.628426	NaN	0.0
LENA	0.825520	NaN	-0.229911	NaN	0.0
MACKENZIE	0.889386	0.911745	-0.010903	0.193439	100.0
MISSISSIPPI	0.932459	0.940384	0.469833	0.532042	84.4
OB	0.915694	0.957197	0.227638	0.607863	66.0
ORANGE	0.341260	0.205192	0.196090	0.030036	100.0
PARANA	0.902606	0.899179	0.688144	0.677171	95.0
VOLGA	0.918098	0.937325	0.445893	0.575976	66.0
YANGTZE	0.747095	NaN	0.214342	NaN	7.1
YELLOW RIVER	0.740573	NaN	0.501178	NaN	7.1
YENISEY	0.920923	0.935921	0.269030	0.407666	74.5
YUKON	0.904029	0.931050	0.106887	0.358348	100.0

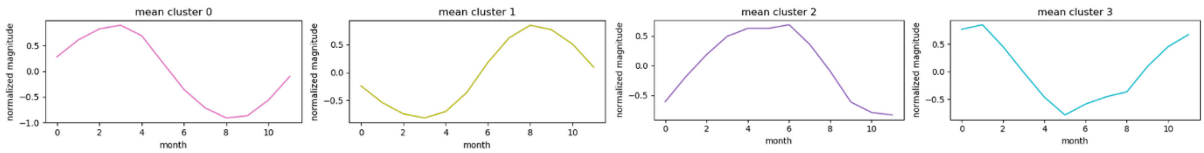
I find the separation into Köppen-Geiger-classes a bit problematic. […] If possible, it would be worth to check if some of these metrics give a clearer picture of the regional performance.

With the Köppen-Geiger classification, we tried to understand which factors could explain similarities between basins. It is not perfect but one of its advantages is that its assumptions are well-known. We have already explored some variables exposed in the HydroSHEDS database (annual discharge, lake volume, reservoir volume, degree of regulation) and we did not find any significant relationship between those variables and the imbalance error.

We also tried a clustering method based on the shape of the TWS seasonal signal that did not help further. Below is an example of the four clusters detected by the algorithm (K-means clustering) with the mean seasonal signal representative of each cluster. However, the clusters were highly dependent on the initialization of the algorithm and we were not able to understand why a basin belonged to the selected cluster. Therefore, this idea was not explored further.



Basins clustered in 4 groups using a K-means algorithm based on the shape on the mean annual cycle of TWS



Mean annual cycle of TWS representative of each cluster

Concerning the fact that a basin may be comprised of several climate zones, we specifically studied the Amazon sub-basins that are of smaller sizes and less diverse climates. The best combinations in the sub-basins were similar to those found for the entire Amazon basin, leading us to think that aggregating basins with some heterogeneous climates was not a major drawback.

[…] Future studies and regional applications must therefore use the findings from this study (e.g. from figures A12-A15) as a starting point to further explore strengths and weaknesses of individual datasets across different regions. Only then are we able to see improvements in our global hydrometeorological data sources (as the authors also state in the abstract).

We entirely agree that our study is only a starting point. It highlights a need for independent validation data that we hope to develop in the future.

Minor comments

The paper is very long and contains a huge amount of quite detailed information! I would hence really urge the authors to reduce the number of pages!

As suggested, we have reduced the length of the article and clarified which material belongs to the supplementary information.

The authors mention that they bring every dataset to a resolution of 0.5°. This, however, could lead to issues for coarser datasets (GLDAS, GPCP, etc.), particularly for smaller catchments. Are there any relationships between the resolution of the input datasets and the performance metrics?

In this study we have explored more than 1600 combinations. Having the same resolution for all datasets was necessary to ensure the consistency of the computations that included numerour loops. Since we are only interested in basins above the GRACE spatial resolution (>65000 sq. km), we believe that changing resolution should not have a significant impact on basin-mean time-series on a monthly time scale; and we did not notice any result that could suggest it might be the case. A more detailed reply is given below comparing GPCP and ERA 5 Land.

Page 10, line 223: I guess err_cst are simply anomalies, right? Then, err_cst^2 is simply the temporal variability of total water storage changes. If this is the case, I would not call these errors as this sounds misleading.

Page 10, equation 6: Similarly, err_cyc^2 is just the deviation from the annual cycle and, again, using the term error for such anomalies is quite misleading.

It is right that the terms err_cst and err_cyc represent deviations and therefore may be misleading. We have modified those names.

Page 11, line 260: The two consecutive enumerations (i.e. lines 257 - 259 and 260 onwards) look a bit weird... Please add at least one sentence for separating these two parts.

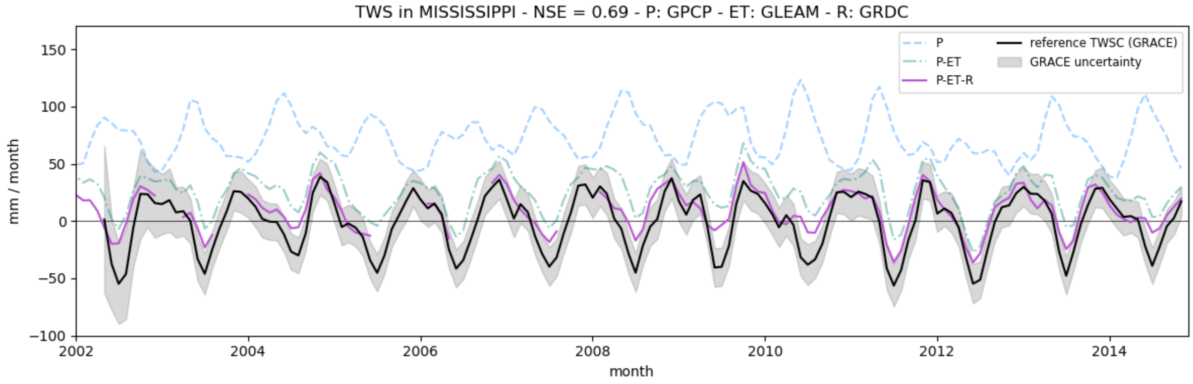
page 12, line 263: ...is within the confidence interval from GRACE TWSCs.

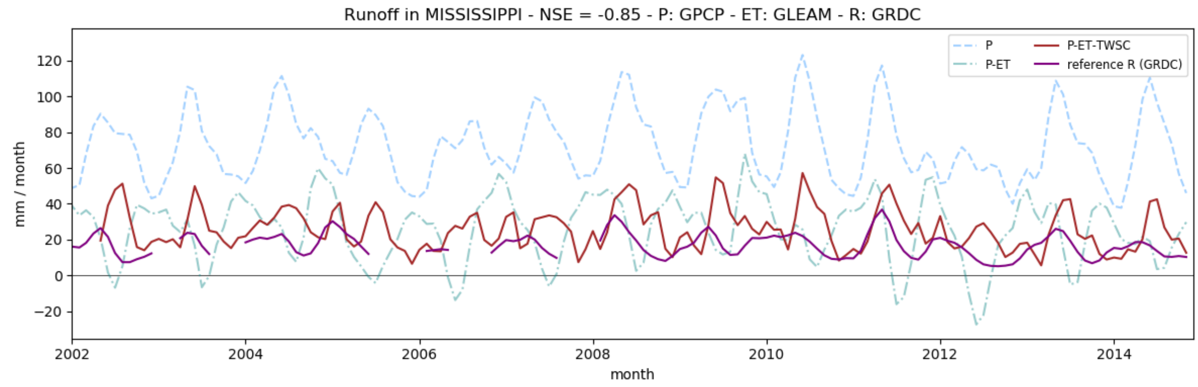
The suggested corrections have been made.

Page 13, lines 294 - 295: Do you have an explanation why the performance has improved? Is it due to improvements of the consistency or the performance of the hydrometeorological datasets?

The main improvements come from the reference variable used to compute the NSE. Since R is generally small compared to P and ET, the NSE is higher with TWS as the reference rather than with R. The two figures below illustrate this with a combination found in [Lorenz et al., 2014].

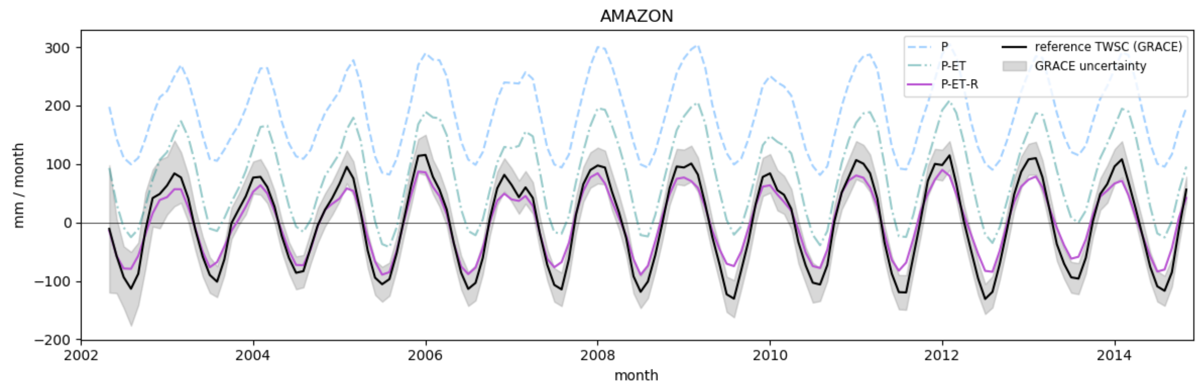
The top figure shows TWS as the reference and leads to a high NSE (0.69) whereas the bottom one has R as a reference and shows a negative NSE, similarly to [Lorenz et al., 2014].

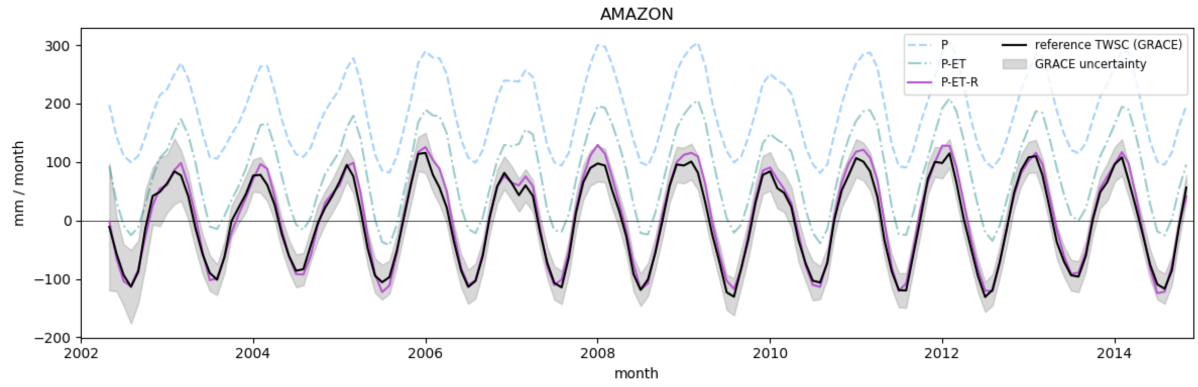




Page 14, lines 315 - 325: Do you have any explanation why TWSC is too low in the wet and too high during the dry season, respectively? According to Fig. A3, this under- and overestimation seems to be quite systematic.

In the Amazon basin, the under- and over- estimation was coming from inappropriate values of R computed by the GRUN dataset. Top figure: with GRUN, bottom: with GRDC gauge measurements. However, we also note that the basin used by the GRDC is 20% smaller than ours.





Page 15, line 350: Important for what?
Page 18: line 398: For this analysis, we focus on a subset of 132 basins out of the 189, where an excellent budget closure could be achieved.
The sentences have been rephrased as suggested.

Page 19, line 405: This is a dangerous conclusion as it indicates that two very "bad" datasets can still lead to good water budget closure, if there occurs a cancellation of biases (i.e. right for the wrong reason), right? This would mean that e.g. the datasets in Figure 12, that satisfy a cost lower

than 0.1, must not necessarily be realistic datasets but, by combining them with other suitable datasets, only achieve a reasonable water budget closure.

When examining individual datasets, the risk of bias cancellation seems unavoidable since we cannot *a priori* exclude some datasets. This is why we implement the clustering approach (see the reply to the next question).

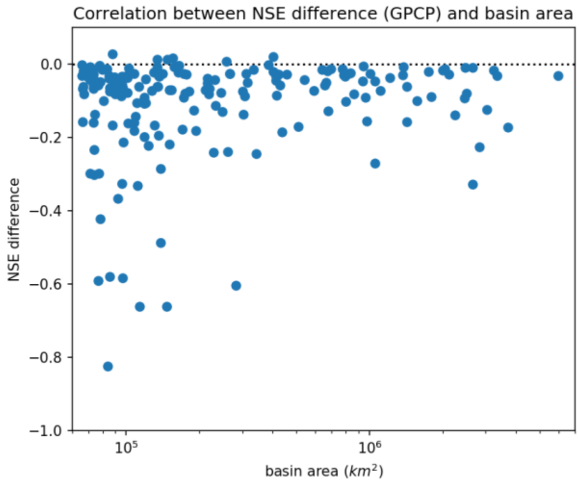
Page 19, lines 12 - 16 and Figure A7: I did not really understand the clustering approach. What exactly are the authors trying to do here? Do they want to define 13 representative catchments and then identify smaller basins that achieve a similar performance? If this is the case, why did they choose 13 "artificial" clusters instead of using e.g. similar climatic or topographic conditions?

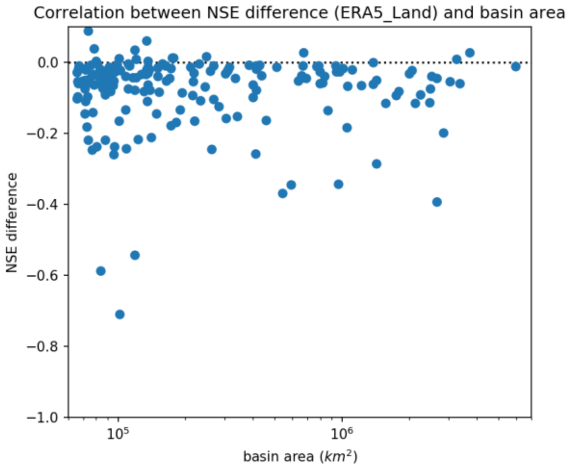
When examining datasets leading to a small imbalance error in a given basin, it appeared that many dataset combinations performed equally well. However, they would frequently lead to a low performance in a neighbouring basin with seemingly similar conditions (size, climate zone, or shape of the mean seasonal TWS). This prevented us from drawing any conclusion on the ability of this combination to effectively close the water budget.

Therefore, we tried to increase the constraints on the combinations from closing the water budget in a single basin, to closing the water budget in several basins. Since no combination was able to meet this condition in every basin of a climate zone, we relaxed the constraint to closing the water budget in several basins sharing some similar patterns that we do not necessarily understand. We then used the clustering approach to find groups of basins following the same patterns, i.e. satisfying a low imbalance error for similar combinations.

Page 19, line 430: This is an important statement as GPCP has by far the lowest spatial resolution of 2.5° (around 250km). Claiming that GPCP (approx. 250km) performs similar than ERA5 Land (9km) indicates that resolution does not play a big role, even this is generally assumed by the community (especially over complex terrain). Could you find any relationship between the performance of GPCP and the size of the catchments?

Assessing the performances of the datasets with the NSE difference, we found no correlation between the size of the catchments and the performances of GPCP (R^2=0.09). There is no significant difference with ERA5 Land (R^2=0.06).





Page 21, lines 450 - 455: The authors suggest that the discrepancies between the TWSC from water budgets and GRACE are somehow related to overfitting of the CLSM. But there is also a huge temporal shift between the two time-series. Are there any explanations for this?

We realized that our interpretation was maybe too superficial. Therefore, we deleted our explanation and instead suggested that this remains an open question for the community.

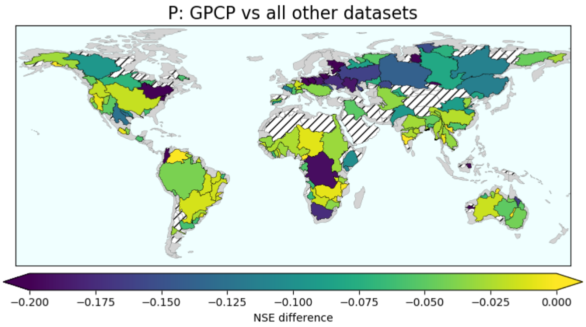
Page 21, Figure 12 (and A12-A15).: Is there any meaning of the length of an individual section (or dataset)? And at least on long-term averages, we often assume that P should be equal to ET + R but this is not the case for several clusters. Can you explain why this happens here? Or did I misinterpret the figures?

The length of a section represents the proportion of combinations in which each dataset was found. When one dataset has a much longer section than the others, it means that it was able to close the water budget when being combined with several datasets. Hence, this reduces the risk of error cancellation with this dataset. We will clarify this legend.

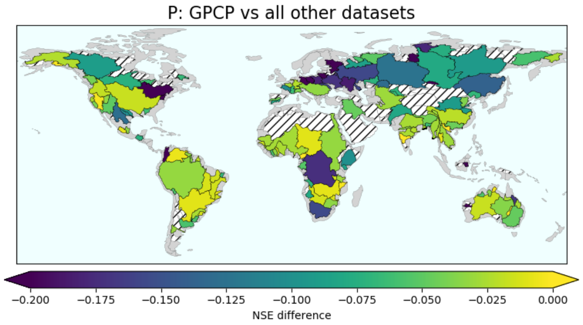
The total length of the P, ET, and R sections were computed using the annual mean over all datasets. Therefore, we cannot expect that different datasets lead to P=ET+R. Since this is only for representative purposes (to express the x-axis as a percentage and make basins comparable), it should not be misleading.

Page 22, Figure 13: As the distribution of NSE-values might be highly skewed, wouldn't it make more sense to show the median of the 10 best-performing combinations?

As shown below, there is very little difference between using the mean or the median since it is computed over the 10 best-performing combinations over 1600 combinations. The possible skewness would appear with a larger number of combinations.



Using the median of the 10th highest combinations



Using the mean (as done in the article)

Pages 21 and 39, Captions for Figures 12 and A11: Why are MSWEP, PGF and GRUN outside the boxes?

Datasets were gathered by “type”. Since MSWEP and PGF are built using a combination of several methods, they do not really belong to any category among “rain-gauge”, “satellite”, or “reanalysis”. Similarly, GRUN is the only machine learning dataset for runoff and therefore was separated from land surface models.