Petch and co-authors study the water and energy balance of 20 large river basins. Given my own expertise, I will mainly comment on the water balance part. Using GRACE-derived terrestrial water storage changes and globally available observation-based product of precipitation, evaporation and runoff they first study the imbalance between these products.

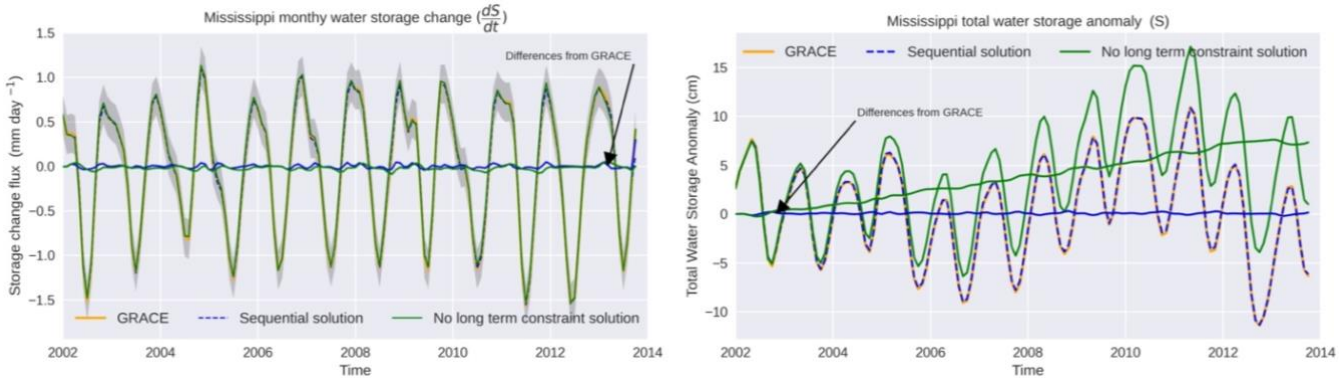
Thanks again for the review of our manuscript. The responses given for the general comments are as in our earlier response during discussion phase. We now add responses to the specific comments below.

Then, I failed to understand what the usefulness of the "Our Optimised Storage" and associated figures 4 and 5 are. It seems quite circular to me that if you force it to match GRACE it matches GRACE better than other products. From a hydrologist perspective I would rather see the optimized P, Q and E compared to other products. Perhaps even an independent better regional precipitation or river discharge dataset for specific basins, to see whether the optimized fluxes match that better and which would then clearly demonstrate the strength of their method. I am not sure if this is going to require major changes to the paper, or just a clearer explanation of the objectives and results.

'Our optimised storage' is the total water storage produced from integrating our optimised P, Q and E fluxes. The usefulness of this quantity is that it is more directly comparable to GRACE products, and is able to capture information that cannot easily be seen when looking at individual monthly flux components. Small imbalances in monthly fluxes which would be hard to validate with any single month of data, can still imply unrealistic long-term changes in a basins water storage e.g. associated with rising/lowering water tables. GRACE data does track these longer timescales. For this reason, a set of fluxes consistent with GRACE is more useful for evaluating modelling products for example, using which we may seek to attribute both short- and long-term water budget variations. Previous studies which have sought to develop observational flux products consistent with GRACE have failed to match low frequency variations which are important to understand through hydrological modelling.

Previous optimisation studies have only use GRACE on a monthly timescale. Using GRACE total water storage anomalies, rather than only monthly storage changes, enables us to use more information to provide stronger constraints than previous studies.

I have plotted the figure here to help explain further. This compares two optimisation outputs for the Mississippi basin, both of which use GRACE as input. In blue, we show our sequential optimisation approach described in the study, along with an optimisation that enforces monthly water balance but with no long-term constraint in green. The left figure shows monthly flux imbalances (P – E – Q) compared to the GRACE dS/dt. Both solutions are very close to GRACE. However, in the "optimised



storage" plots (right), only our sequential method fits low frequency GRACE storage changes. Figure 5

in the manuscript brings out this low frequency information, including from other products, that would not be seen if we only compare monthly P, Q and E estimates.

The main objective is a methods advancement on how to use both short and long-term GRACE data to adjust/optimise the monthly fluxes to achieve a GRACE consistent long-term budget.
In the manuscript we will emphasise that a timeseries of monthly fluxes is not adequate to identify long term consistency hence the accumulated storage figures are shown instead. Will that satisfy your concerns?

A second major, but not difficult to solve issue, is that I find the authors somewhat sloppy regarding equations and symbology. Unfortunately, this set of guidelines has disappeared recently from the HESS manuscript preparations guidelines online: https://iahs.info/Publications-News/Otherpublications/Guidelines-for-the-use-of-units-symbols-and-equations-in-hydrology.do, but I personally still appreciate it if we all try to follow this as much as possible. Of serious concern are Eq. (1) and Eq. (5), which should obviously read dS/dt instead of dS as the fluxes are per unit of time. It would also be helpful if the equations would contain the dimensions, thus, e.g. length^3 per time [L3 T -1] for Eq. (1) so this becomes obvious. Even expressed per unit area [L T-1] would also be fine of course as long it clearly remains a flux and not a stock. Moreover, please use single italicized symbols, so something like Sfi instead of FIS, which makes it directly clear we are talking about a storage. I know many other papers invent funny acronyms as well, maybe it is even the rule rather than the exception, but in my opinion, it is simply not pleasant for any reader.

All equations and symbology will be corrected to follow appropriate guidelines. We will also rename FIS to S_fi, we appreciate this suggestion to help make our manuscript easier to follow.

A third major question is why the authors chose the data they chose and whether it matters for the main point they are trying to make. For precipitation and evaporation, many more observation based products exist, so did they select the 'best' according to some previous studies or did they just select 'good' data and does it not matter a lot whether it is really the 'best'. I hope the authors can explain. Moreover, the runoff data is even dependent on precipitation and evaporation from GSWP3, which is a bit of a vague product in terms of how it was constructed and I think it may even rely partly on GPCP and FLUXCOM, making the estimates of P, E and Q not completely independent. Moreover, I fail to see why spatially varying runoff is necessary at all, as on the basin scale, the actual river discharge measurements at the river mouth would suffice for which, for example, the GSIM archive (Do et al., 2018) could have also been used.

As our main aim was to present a methodological advancement so we chose a "good data set" and we noted that many previous studies have used ensemble products because there is no "best data". The choices were not critical to the main points. Although we only used single data products, uncertainties are applied based on previous multi-product studies. We used EO based datasets where possible partly due to the nature of our funding, which led to the decision to use the GPCP product, however we also sought global gridded products as this ensures uniformities of uncertainties across all basins and it is a future ambition to move towards a gridded version of our own product.
For Runoff the GRUN product is the only available global gridded runoff dataset we are aware of; it uses GSIM observations as input and has been validated against many river flow datasets using monthly river discharge from the GRDC, see Figure 3 from Ghiggi et al., (2019). It is true that the runoff data is not completely independent from the precipitation dataset, and this is explicitly discussed in lines 500 - 509 in the manuscript. This has also been noted in other papers, however we go further in explaining the likely qualitative impact on results, which is still quite small for the cases shown.
Does this address your concerns if we include more discussion along these lines in the manuscript?

Specific comments

L1-2: "improving climate and earth system models" I would say 'validating' or 'assessing the capability of' which is to be done first before anything can be improved.

We agree with comment and this statement will be changed in manuscript to 'validating climate and earth system models.'
As well as validation, optimisation can also be useful prior to data assimilation of products into such models.

L6: "the corresponding turbulent heat fluxes ranges between ± 10 W m−2 " I suppose something should range between value x and value y, thus this sentence misses something.
This statement in abstract will be edited to give absolute maximum and minimum imbalances.

L8: "This exposes mismatches in seasonal water storage" Mismatches between what and what exactly?
We are referring to the mismatches between seasonal water storage seen by GRACE, and the seasonal water storage implied from raw P, Q and E observations. We will make this clear in the revised manuscript.

L12: "The optimization also reduces formal uncertainties on individual flux components" Sounds great, but I failed to clearly identify this result in the paper itself.
This is shown in Table 3 and discussed in section 5.2 'Uncertainty estimate' in the manuscript. In the revised manuscript we will more clearly emphasise this result.

L14: "The FIS metrics" What are 'the FIS metrics'?
Here, the FIS metric was referring to results gained when calculating the accumulated storage implied from optimised fluxes from other studies. This line in abstract will be rewritten to avoid using this phrase as it is not fully explained at this point.

L23: "Water is a conservative quantity" Technically speaking this statement is incorrect. Water is used by plants for photosynthesis and released by decomposition or fire. Probably it is an order of magnitude lower than the errors made in the products of P, E and Q, but not entirely negligible.
What we meant by this is that the mass of water will be conserved. We will change text to be clear on this.

Table 1 "present" and general period statements It is rather irrelevant whether e.g. GRUN is available until 'present' or that it starts in 1902, what matters is which years you used for the analysis.
The period column will be removed from Table 1, and we will instead state the period we downloaded each dataset for. "All datasets have been downloaded for the months between October 2001 and December 2013"

L91 "evaporation" I strongly support the use of evaporation over the ambiguous term evapotranspiration, see Miralles et al. (2020) for the arguments why that is, so perhaps you could simply use evaporation also elsewhere in the manuscript.
Thank you for pointing us to Miralles et al. (2020), we agree with this point and will use the term 'evaporation' instead of 'evapotranspiration' throughout the manuscript.

Equation 5 The integral is between what and what? What does the to the power 0 between brackets mean? Is this equation supposed to present a time series? Then it would be clearer if Sfi,w(t) was explicitly written.

This equation will be rewritten as following.

$$S_{fi,w}[t] = \int_0^t \left(\frac{dS}{dt}\right) dt + S[0] = \int_0^t (\mathrm{P} - \mathrm{E} - \mathrm{Q}) dt + S[0].$$

We integrate between time 0 and an arbitrary time t. Where 0 represents the first month in our period. We have also added S[0] which is consistent with the step of initialising with GRACE storage from January 2002.

The superscript 0 was to indicate a first order correction hae been applied, this was only associated with the detrended flux inferred storage shown in figure 4 and is not relevant for the optimisation, so it has been removed from equation 5. Text has been changed to make this clear.

For the detrended flux inferred water storage, the first order correction was to match the mean of P-E-Q to the mean of dS/dt from GRACE. This quantity was calculated to help identify imbalances in the seasonal storage cycles between GRACE and the other flux observations. For the detrended flux inferred energy storage, the first order correction forced the long-term NET to be zero.

Technical corrections

L93: "Earths" Earth's

This technical correction will be made.

L101: "Land" land

This technical correction will be made.

References

Ghiggi, G., Humphrey, V., Seneviratne, S. I., and Gudmundsson, L.: GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth Syst. Sci. Data, 11, 1655–1674, https://doi.org/10.5194/essd-11-1655-2019, 2019.