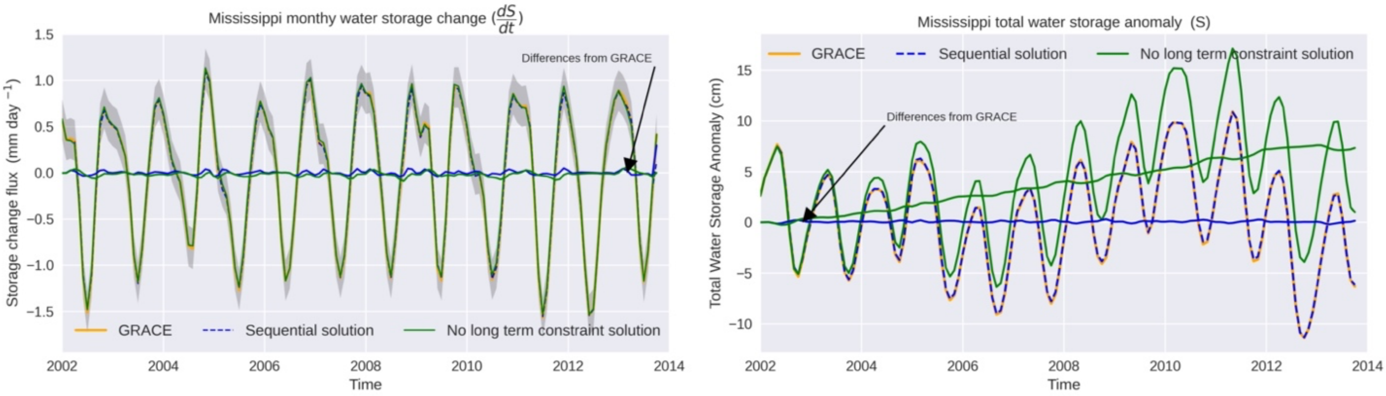
Thanks for giving a detailed review of our manuscript, we really appreciate all your comments. We would also like to thank you for encouraging a quick discussion and will use the opportunity to respond to your general comments. We will address the other comments in our response to all of the reviews later on.

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 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 with which we may wish 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.

And so, ‘our optimised storage’ is used to demonstrate how we are able to achieve long-term balance with only small adjustments to the monthly fluxes. Previous optimisation studies have only use GRACE on a monthly timescale. By using GRACE storage (S) rather than the monthly storage change (dS/dt) derived from GRACE, it enables us to use more information to provide stronger constraints than these other studies.

I have plotted the figure below to help explain this further. The figure 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 methodological advancement about 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 can try to emphasise more clearly that a timeseries of monthly fluxes is not adequate to identify long term consistency hence the storage figures that are perhaps less commonly shown. Would 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. We chose what should be good data and we do note that many previous studies of this type produce ensemble products because there is no such thing as the "best data". The choices of datasets were not critical to the point we were trying to make. Although we have only used single data products, uncertainties we applied are based on previous studies that have used multiple products to estimate uncertainties. We chose to use EO based datasets where possible due to the nature of our funding, which led to the decision to use the GPCP product. We also chose to use global gridded products as this ensures uniformities of uncertainties across all basins and it was 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 correct 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 however quite small for the cases shown.
Would it address your concerns if we included a discussion along these lines in the manuscript?

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.