Authors' reply to Anonymous Referee #2

We thank the referee for his/her constructive and detailed review of our paper. We have adapted the manuscript in accordance with the referee's suggestions, as detailed below. Referee comments appear in *italics*, our reply in normal font, and changes to the manuscript in blue.

Review of "Intercomparison of freshwater fluxes over ocean and investigations into water budget closure" by M. Gutenstein et al.
This is a nice intercomparison study of various satellite-based precipitation and evaporation products and ERA5. The authors show that there is a large spread among the different products, and most of them fail to satisfy global budget constraints, especially when combining different products for estimation of P, E, and the transports. ERA5 performs best with a remarkably good agreement of forecast-based fluxes (P and E) and analysis-based transports (moisture flux divergence). The paper is well-structured and -written and therefore easy to follow.
Thank you!

My only major comment is on the obvious error in the computation of ERA5 moisture transports as detailed below.
Moreover, the reader would probably like to see stronger conclusions. I know it is hard to make a ranking, but e.g. SEAFLUX with its clearly unphysical P-E over ocean could be ruled out as clearly unrealistic. Also HOAPS appears to be a bit of an outlier, especially in terms of variability. The budget constraints are an objective measure to rule out poor data, and this helps to better constrain the best estimate of the water budget, without unnecessary inflation of the error bars. Otherwise I only have a number of minor comments.
We would certainly have liked to conclude with more clear recommendations, but believe that, apart from the finding that the global total E from SEAFLUX2 is unrealistic, our results do not permit any kind of ranking. First, there are not enough truly independent data with which to assess the quality of each data set. For example, although HOAPS precipitation is an outlier compared to other data sets, it is not certain that its variability (e.g. dependence on ENSO) is erroneously large, as there are reasons to believe that ERA5 and GPCP underestimate variability. This, however, is the topic of a future study. Second, each data set has its particular strengths and weaknesses, and HOAPS comes closer to water budget closure than OAFLUX or IFREMER. For these reasons, our conclusions cannot go beyond the statement that observational data sets (and associated uncertainty estimates) need to improve.
We added this information to Section 5: "Although it is tempting to make a ranking from the results of our inter-comparison, there are good reasons to resist. First, there are not enough truly independent data with which to objectively assess the quality of each data set. And second, each has its particular strengths and weaknesses: for example, HOAPS comes closer to water budget closure than OAFLUX or IFREMER (panel c of Fig. 3)"

And we modified our final statement as follows: "In general, the quality of observations of the water cycle needs to improve before attempts at assessing effects of climate change from those data can be undertaken. The importance of accompanying high-quality uncertainty information cannot be overstated."

Major comment: The authors use VIMD from ERA5 to compute ocean-to-land moisture transports. For this it should not matter whether on integrates VIMD over all land points or over all ocean points. Another constraint is that the global average of VIMD must be zero. This is a mathematical constraint independent of data quality. However, the authors obtain inconsistent results for land and ocean integrals of VIMD (table 4). So either the archived VIMD fields are flawed (which I doubt as I am using the ERA5 data myself and could not find a similar problem) or there is some error in the author's

processing chain that leads to these erroneous results. If this problem is really due to interpolation errors, as suspected by the authors, these interpolation errors are clearly unacceptably large. In short, this error must be corrected.

We could not agree more with the referee: summed over the globe, VIMD should equal 0. However, two of us calculated the area-weighted sum over the globe in three different ways and every time we find the rather constant value of -0.04 mm/day. These calculations were performed with monthly mean VIMD, regridded at ECMWF to a regular, 1x1 degree grid. This puzzled us and we contacted the ERA5 team about it at the beginning of 2020. Paul Berrisford gave us the explanation that we eventually wrote into the manuscript, i.e., the error occurs during transformation from the model grid to a regular lat/lon grid. Following the referee's comment, we contacted the ERA5 team again. Paul Berrisford confirmed the finding that global mean ERA5 VIMD equals -0.04 mm/day and commented: "Anton [Beljaars] also points out: 0.04 mm/day is not very big. I do not think accuracy at that level can be claimed on the fields of the water cycle. Also observationally, I doubt whether global precipitation is known to within one tenth of a mm/day."

Whereas this is certainly true for a single grid cell, we, like the referee, feel that at the global scale 0.04 mm/day (which sums up to about $10 \cdot 10^3$ km^3/year, as seen in our Table 4) is large – perhaps even unacceptably large. However, a detailed discussion of the issue is not within the scope of our manuscript and should take place elsewhere, preferably with direct involvement of the ERA5 experts.

To address this issue more clearly, we changed lines 413-415 to: "However, we find global total ERA5 VIMD to be -0.04 mm d^{-1}: a small value within the standard deviation of the ensemble of single grid boxes, but significant and on the order of the amplitude of the seasonal cycle of net E-P on the global scale. The deviation from zero is due to the fact that VIMD is calculated in grid point space (and not in the model's spectral space), where the mathematical constraint of net zero divergence is not enforced (P. Berrisford, personal communication, Oct. 2020)."

And line 497-499 to: "As observed above, the fact that ERA5 VIMD is calculated in grid point space causes $\nabla(vq)$ to be about $10 \cdot 10^3$ km^3 yr^{-1}, and not zero. In addition, due to the tighter observational control over land, analysis increments may be larger over ocean than over land and may cause $\nabla(vq)$ to be very close to net *E-P* over land, but less so over ocean (P. Berrisford, pers. comm. Oct. 2020). "

Minor comments:

*Equation 1 and everywhere else: as it stands, the VIMD terms looks like the moisture gradient. I suggest to replace with the more appropriate nabla * (vQ).*

The referee is right. We replaced the inaccurate term with $\nabla(vq)$ in the equations and text.

L17: I presume you use monthly values. Please say it clearly, as the correlation strongly depends on the considered timescales.

Correct. We changed the sentence to reflect this:

"On a monthly time scale, linear regression of $E_{ocean} - \nabla(vq)_{ocean}$ with P_{ocean} yields $R^2 = 0.86$…"

L37: The term "model reanalysis" seems an uncommon term to me. I suggest to drop "model". If you want to give an attribute, it may be better to say "climate reanalysis" or "dynamical reanalysis".

Done.

L46: Isn't there an author on the GPCP document?

This is not a document, but the DOI reference to the data set itself; no single author is indicated. In fact, GPCP-1DD v.1.3 was not correctly cited in the manuscript, and we updated it accordingly:

"Mesoscale Atmospheric Processes Branch/Laboratory for Atmospheres/Earth Sciences Division/Science and Exploration Directorate/Goddard Space Flight Center/NASA, and Earth System Science Interdisciplinary Center/University of Maryland: GPCP Version 1.3 One-Degree Daily Precipitation Data Set, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, doi: 10.5065/PV8B-HV76, 2018. Accessed June 2019."

L58: "moisture divergence" is sloppy terminology. Moisture itself cannot diverge. It should be "moisture flux divergence".
Corrected.

L58: another nitpicky comment: VIMD is technically not identical with advection, although it is an excellent approximation of it. I suggest a slightly more cautious wording.
Corrected.

L85-86: "model runs" sounds it does not use any observational info. Simply say it consists of ten ensemble members.
Corrected.

L108: Le is usually called "latent heat of evaporation".
Corrected.

L133: Do you mean SST averaged over the top 0.5m? Please clarify.
We wrote: "(…) a bulk SST at 0.5 m", which refers to the minimum depth at which the sondes measure to which satellite data are calibrated. This was clarified in the manuscript by changing the text to: "(…) the SST at a depth of 0.5 m".

L183-184: The statement about forecast skill is hard to understand for a non-expert.
Agreed, but it is not of importance for the rest of the study, so we will not explain it further here.

L184: "model runs" -> see comment above.
Corrected.

L184: It should be mentioned that the ERA5 ensemble members have a lower resolution than the stated 30km. How would the results change when using the high-resolution ERA5 data?
The referee might have misunderstood that we did, indeed, perform the study with high-resolution data (although those were interpolated to a regular 1 degree grid by ECMWF prior to our investigations). Only uncertainty estimates were made using (lower-resolution) ensemble data that was similarly interpolated to a 1 degree grid. The text was modified to make this more clear: "ERA5 encompasses data from ten reanalysis runs at a reduced spatial resolution of 62 km, allowing estimation of the uncertainty range from ensemble statistics. The analysis presented here is performed with the ECMWF ensemble mean, whereas uncertainty is determined from the ensemble. Both data sets were interpolated to 1° resolution at ECMWF."

L189: I suggest to delete "on single levels"
Done.

L192: It should be noted that monthly P and E from ERA5 is averaged from short-term forecasts (12 or 24 hours? Needs to be clarified as well!)
We changed the text on lines 191-195 to: "Monthly averages are calculated from daily means starting at 00 UTC and ending at 00 UTC the following day (ECMWF, 2020). Evaporation rates are derived from the gradients of specific humidity between the surface and the lowest model level (10 m for ERA5) as described above (ECMWF, 2016). The main differences between the satellite-based retrievals described here and ERA5 determination of E are the consistency of atmospheric variables involved (u, q_a, q_s) and the high temporal sampling rate: monthly means are determined from (daily means of) hourly data from forecasts initialized daily at 6:00 and 18:00 UTC."

L198: If the TCWV tendency is computed from monthly means (rather than instantaneous values at beginning and end of the month), one should use centered differences.

The referee is right. We re-computed TCWV tendencies as suggested and updated Fig. 7. The figure changed only slightly:

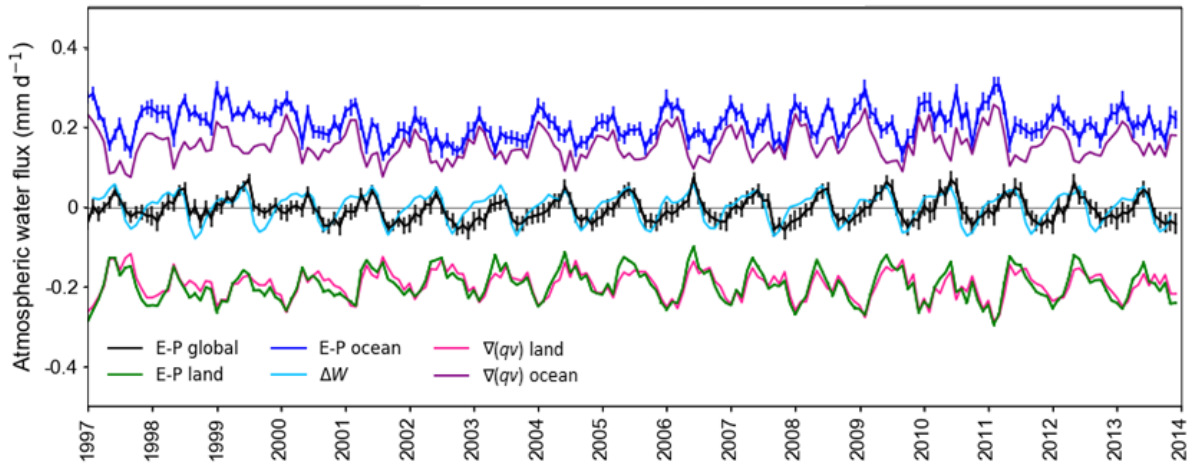

Figure 7. ERA5 monthly mean E-P over the whole globe (black), land only (green), and ocean (blue); global mean ΔW (light blue), mean ∇(vq) over land (pink) and ocean (purple). The mean values over the globe and land were scaled by their surface area relative to the ocean surface area (i.e., they were multiplied by 510/350 and 160/350, respectively) to obtain consistency with the over-ocean means shown in Fig. 3. Error bars represent the standard deviation within the 10-member ensemble, which is smaller than the graph's line width for E-P over land, ΔW, and ∇(vq).

Line 198 now reads: "We calculated the TCWV tendency in month *x* from monthly mean ERA5 data by subtracting TCWV of month *x*+1 from TCWV of month *x*-1, then dividing by 30 days/month (…)"

L243: When deriving E from monthly Q fields, what is the error from neglect of sub-monthly covariance between E and SST?
A test with HOAPS-4.0 data shows that the error due to the derivation of E from monthly LHF and SST is systematic, but of negligible magnitude. This is shown in the figure below, which depicts the differences between E derived from monthly means and monthly mean E from instantaneous values. Monthly mean difference maps are very similar to the left panel below, with the mean global difference 0.005 mm d^{-1} and grid point differences barely exceeding 0.01 mm d^{-1}.

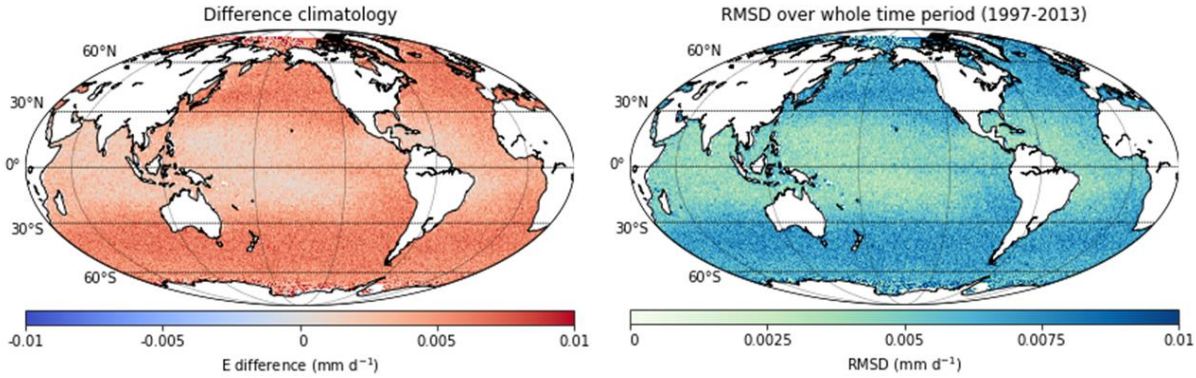

Figure A. Difference between E derived from monthly mean HOAPS LHF and SST and monthly mean HOAPS E for the time period 1997-2013. Left, climatological difference; right, root mean square differences.

We included this information into the manuscript by inserting the following statement on line 248: "Applying the same method of calculating E from HOAPS monthly mean LHF and SST data causes

negligible differences with monthly mean E determined from instantaneous LHF and SST data (root mean square differences of ≤0.01 mm d^{-1} for individual grid boxes during 1997-2013)."

L258: I suggest to replace "relative" with "area-specific" and "total" with "area-integrated".
Done.

L260: Please clarify how sea ice is treated in general. I presume it is masked out? Is this a seasonally varying mask?
As mentioned in Section 3, we neglected the seasonally changing number of observations screened out by the sea ice mask in our study. We clarified this by changing the statement on line 260 to:
"Seasonally varying numbers of observations screened out due to sea ice are neglected."

L273: "at the ITCZ" maybe better "in the ITCZ"?
Corrected.

L280: Figure 2: One could make the simple statement that HOAPS differences are generally larger (RMS values of the field would be useful), but areas of disagreement are smaller because of the larger uncertainties.
That is an excellent suggestion. We added the total RMSD to the text and inserted a statement similar to the one suggested by the referee to line 280:
"(…) with collocated ERA5 data. Although HOAPS differences with ERA5 appear larger to the eye, the root mean squared (RMS) differences are 0.6 mm d^{-1} for each of the three comparisons: 0.60 mm d^{-1} for HOAPS, 0.58 mm d^{-1} for SEAFLUX, and 0.57 mm d^{-1} for OAFlux. As already seen in Fig.1, differences are not homogeneously distributed over the globe."

L285: similarly -> similar
We changed the word to the more appropriate also.

L317: In terms of flow of reading, it may be better to move the sentence about ENSO correlation further down to around L340.
Done.

Figure 3 and in general: I suggest to change the panel labelling to small letters, as capital letters have potential for confusion, especially "E". Best would be E -> (e)
Panel labelling was changed to small letters.

Figure 3: It would be interesting to see the ENSO correlation for every curve. This could be given in the legend, ideally with the lag at which the maximum correlation occurs.
Good suggestion, we updated the figure with correlation coefficients and corresponding time lag. We also improved the readability of the left panels of Fig.3 by changing the depiction of the uncertainty information.

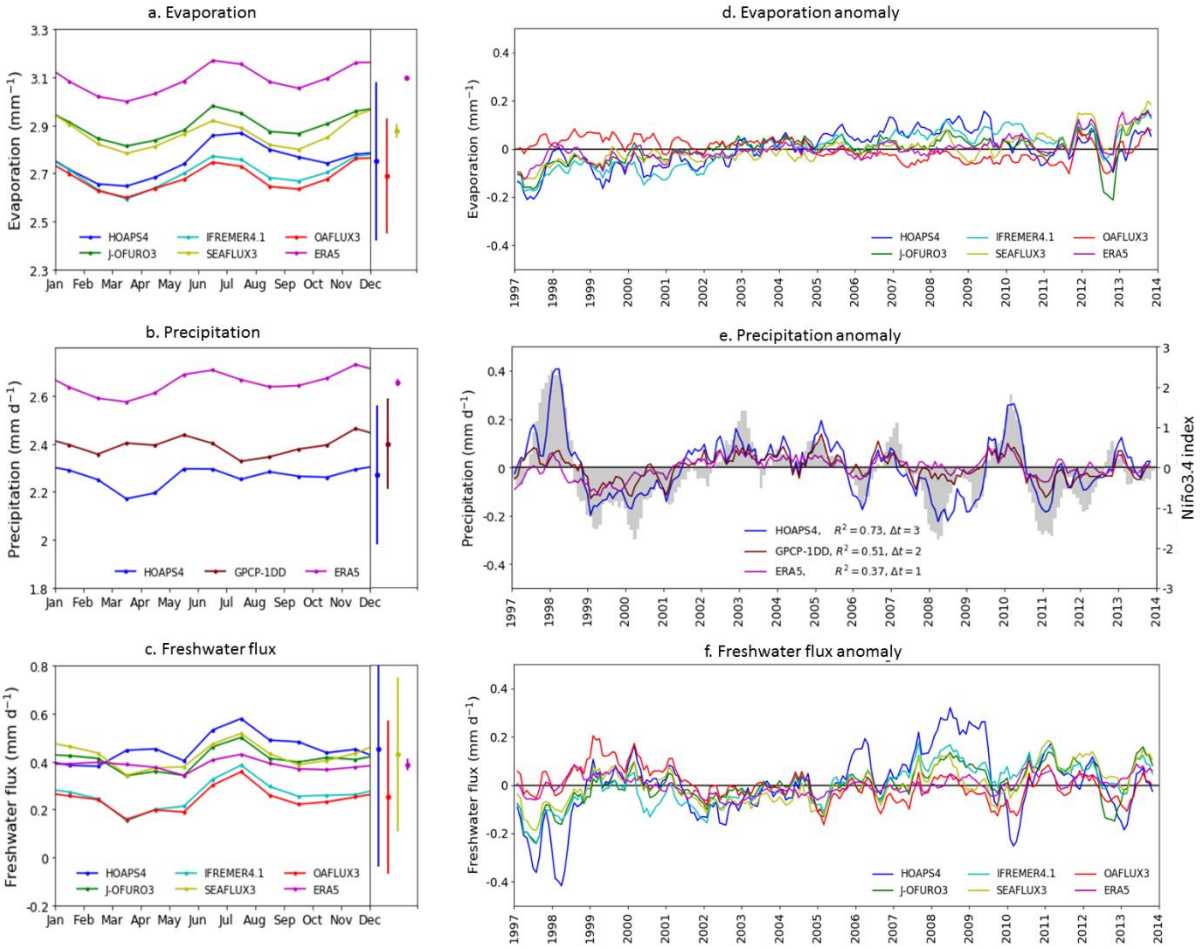

Figure 3. Climatological (1997--2013) seasonal cycle of global ocean mean evaporation rate (a), precipitation rate (b), and freshwater flux (c). HOAPS, ERA5, OAFlux, SEAFLUX, and GPCP yearly mean values and associated 1σ uncertainty ranges are shown in the boxes to the right of the panels. Monthly mean anomaly (w.r.t. the climatological seasonal cycle depicted at left) over the global oceans (80°S--80°N) of evaporation rate (d), precipitation rate (e), and freshwater flux (f). The anomaly data are smoothed using a three-month running mean. Panel e additionally displays the Niño3.4 index shifted by +3 months (right y-axis). The legend additionally displays the correlation coefficient of the Niño3.4 index with P anomalies and the time lag of highest correlation (Δt in months). Ticks on the time axis mark January of the indicated year.

L319: Is "bias" the right term? We see differences, but still one of the datasets could be unbiased.
The referee is right. We replaced the word "bias" in the text with "difference" or "deviation".

L337: "biased low". SEAFLUX seems to be low in general (according to the mean annual cycle figure). So better to change to something like "particularly low".
The statement was changed completely after updating the SEAFLUX data to version 3.

L373: Is there a reference for the statement on detection of snow in HOAPS?
The remark was actually meant generally for all global satellite-based precipitation data sets (or at least the two used for the present study). In an inter-comparison study, Tapiador et al. (2017) show that compared to satellite-borne radar observations, GPCP detects too little precipitation in the higher latitudes. HOAPS performs slightly better, but still underestimates precipitation near the poles. We added this information to the manuscript by changing line 373 to:
"(…) and in part to difficulties pertaining to the detection of snow by passive microwave instruments (Tapiador et al., 2017; Kidd and Huffman, 2012)."

L389: remove "are"
Done.

L413: This statement would be correct if VIMD was the 3D-divergence, i.e. including fluxes at top of the atmosphere, where there theoretically could be an exchange with space. However, your VIMD is 2D and its global average is 0 according to the sentence of Gauss.
Please see our answer to the first major comment above.

L415: See my major comment.
Please see our answer to the first major comment above.

L503: I think it should be "right-most"
Absolutely. We corrected the error.

L575-576: Please provide a reference for this statement.
We corrected the estimates to 2%-3% K^{-1} and added a reference to a paper recommended by referee #1: Allan, R. P. et al. Advances in understanding large-scale responses of the water cycle to climate change. Annals of the New York Academy of Sciences (2020).

L586: This statement is on "observation-based attempts". Please clarify.
The referee is right: as it is, the statement is inaccurate. As mentioned above, we modified it to: "In general, the quality of observations of the water cycle needs to improve before attempts at assessing effects of climate change from those data can be undertaken. The importance of accompanying high-quality uncertainty information cannot be overstated."

Figure 4: middle and right columns: Would it be possible to use color schemes that are really white in the middle?
It is possible to use a different color scale, but we would like to keep gray as the center color to distinguish from missing values, which we prefer to show in white.

Table 4: How is runoff from ERA5 obtained? Is this the area-integral of all grid point values?
Yes, it is. We clarified this in the manuscript in Sect. 3: "Global total runoff from ERA5 and other data sets was determined by calculating the area integral of all points."