Author's response to Referee #1

This study quantified the contribution of each pixel to the global TWS IAV of GRACE observations and two selected predominantly data-driven models, SINDBAD and H2M, as well as its modeling errors. The results show that the global TWS IAV is mainly driven by humid tropical and semi-arid region. The hotspots of modeling errors of the global TWS IAV are mainly located in tropical regions that span across climatic regions. The study provides an improved understanding of the global TWS IAV and its modeling error. Generally, the topic is important, and the study is well written and easy to follow. My comments are as follows.

AC: We would like to thank the reviewer for positive feedback on the study, as well as for the suggestions to improve it further. We address the comments here, and will include the changes in the revised submission.

1. In the high latitudes of the northern hemisphere, glacier changes contribute to TWS, whether the SINDBAD model and the H2M model have a glacier module.

AC: Though glacier changes significantly contribute to TWS, especially to its trend (Rodell et al., 2018; Scanlon et al., 2018), in the high latitudes of the northern hemisphere, the two models, SINDBAD and H2M, do not consider the contribution or process. To account for this limitation, grid cells with > 10% of permanent snow and ice cover were excluded from the study area (Kraft et al., 2022; Trautmann et al., 2022). We will clarify it further in the revision.

2. It needs to be further pointed out that the model is inconsistent with GRACE in typical irrigation areas, such as the western United States, northern India, etc.

AC: We agree about the potential inconsistency between GRACE and models in typical irrigation areas. That is why the analysis excludes regions with the largest anthropogenic influence on TWS trends according to Rodell et al. (2018). We will address this aspect in the manuscript as:

[...] Lastly, the two models covered different land pixels due to independent data filtering (Kraft et al., 2022; Trautmann et al., 2022). The data filtering excludes land pixels with 1) significant fraction of ice, snow, water body, bare land surface or artificial land cover, or 2) a large human influence on trend in GRACE TWS mainly by groundwater extraction. Therefore, only the common land pixels between two model simulations were used in this analysis, and the same land mask was applied to all forcing and constraints.

We will clarify the potential inconsistency also in the discussion as:

Lastly, the additional sources of errors include anthropogenic influence and uncertainties in the GRACE data. Though the analysis excludes pixels with a large anthropogenic influence

on TWS such as Northern India following Rodell et al. (2018), the human impact still remains out of the excluded pixels. In Africa, a decrease in TWS IAV in 2003–2006 (Fig. 6c) was due to the expansion of Nalubaale Dam as well as La Niña (Stager et al., 22 https://doi.org/10.5194/hess-2022-284 Preprint. Discussion started: 5 August 2022 c Author(s) 2022. CC BY 4.0 License. 2007; Awange et al., 2013, 2019). In the Indian subcontinent (Fig. 6d), the TWS changes have been reported due to human 485 impacts such as reducing groundwater abstraction and surging reservoirs as well as increased precipitation (Meghwal et al., 2019; Munagapati et al., 2021). In addition, typical irrigation areas such as the Corn Belt in the USA and northern India (Fig. 4) [...]

3. Figure 2(a) shows that the two models are in good agreement, and they both have some differences from GRACE. Does the input of precipitation significantly affect the simulation results of the model? If other precipitation products are used as input, will the results be different?

AC: Thank you for raising an important point. In the original manuscript, we evaluated the association of TWS variability with different precipitation products and found little influence on the results (Fig. B8 and B9 in the manuscript).

Nevertheless, we have tested the robustness of the results of SINDBAD and H2M by forcing the models using an independent precipitation estimate from MSWEP v.2.8 (Table 1 in the manuscript) instead of GPCP1dd, and repeated the whole analysis. We could again verify that using another precipitation forcing does not significantly change the results and findings of our study. We will highlight this finding in the manuscript and include the relevant analyses (Fig. 1~5, below) in the appendix.

Specifically, two models are still in good agreement when they are forced by MSWEP. The performance of two models has been slightly improved, shown by R^2, the distribution of errors, and the slope of regression equations (Fig. 1). However, the main findings and conclusions are not affected by that and stay the same. For example, spatial contribution to TWS IAV (Fig. 2) and its modeling error (Fig. 3) remain similar in general, so do the systematic larger $wRiver_{max}$ (Fig. 4) and more contribution of remotely-recharged groundwater to transpiration (Fig. 5) in error hotspots.

This is in line with Kraft et al. (2022) who showed that H2M gave almost the same performance when it is trained on different precipitation products.

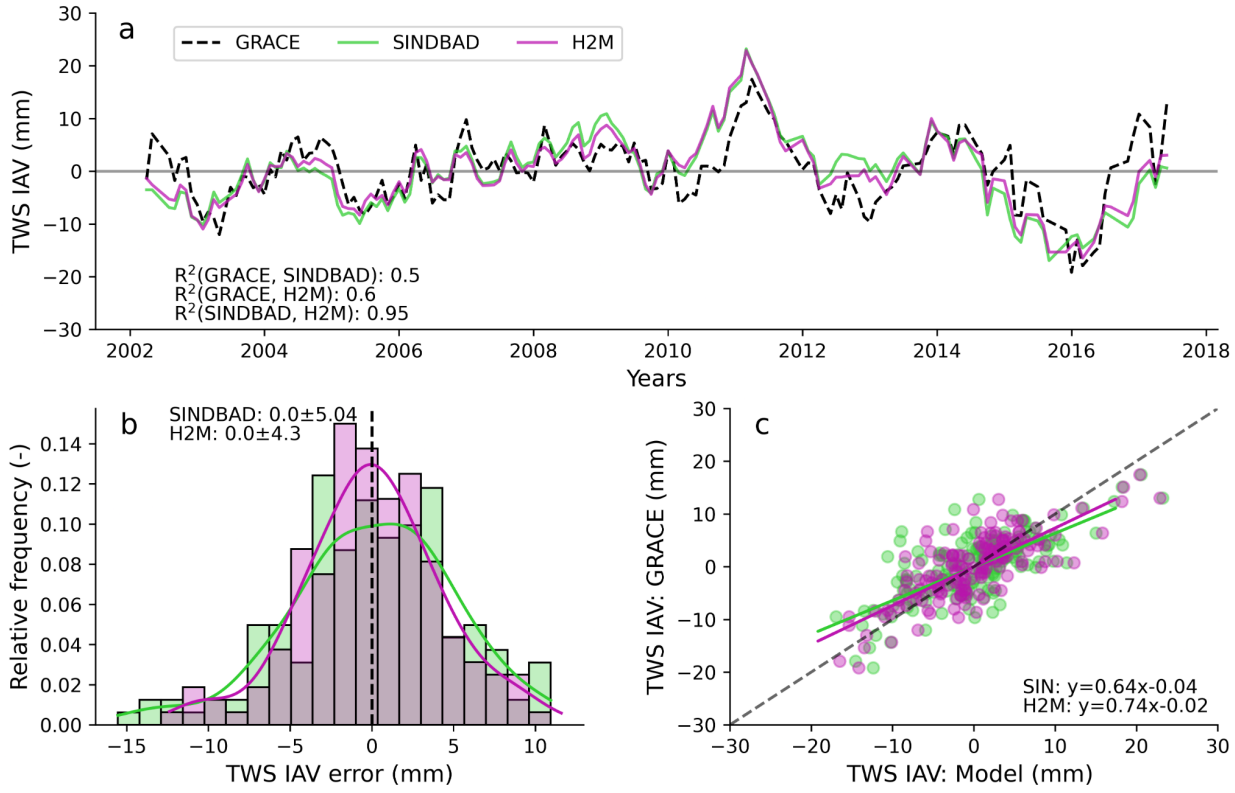

Figure 1. **The same as Fig. 2 in the manuscript, but using MSWEP for the precipitation forcing.** Comparison of monthly global terrestrial water storage (TWS) interannual variability (IAV) from GRACE observations and two data-driven hydrological models (SINDBAD and H2M). (a) Time series comparison of monthly global TWS IAV. R^2 statistics in the bottom-left is calculated as the square of the Pearson correlation coefficient. (b) Histogram of errors of the global TWS IAV (Eq. 3) with smoothed kernel density curves estimated using the Gaussian kernel and the Scott's rule of thumb to determine the bandwidth of the kernel. The sum of all bar heights (different models in different colors) equals unity. Shown text in the upper-left is the mean±standard deviation of the distribution of each model. (c) Scatter plot of monthly TWS IAV by GRACE and models. Equations in the bottom-right are from a robust linear regression using Huber's T estimation for downweighting outliers.

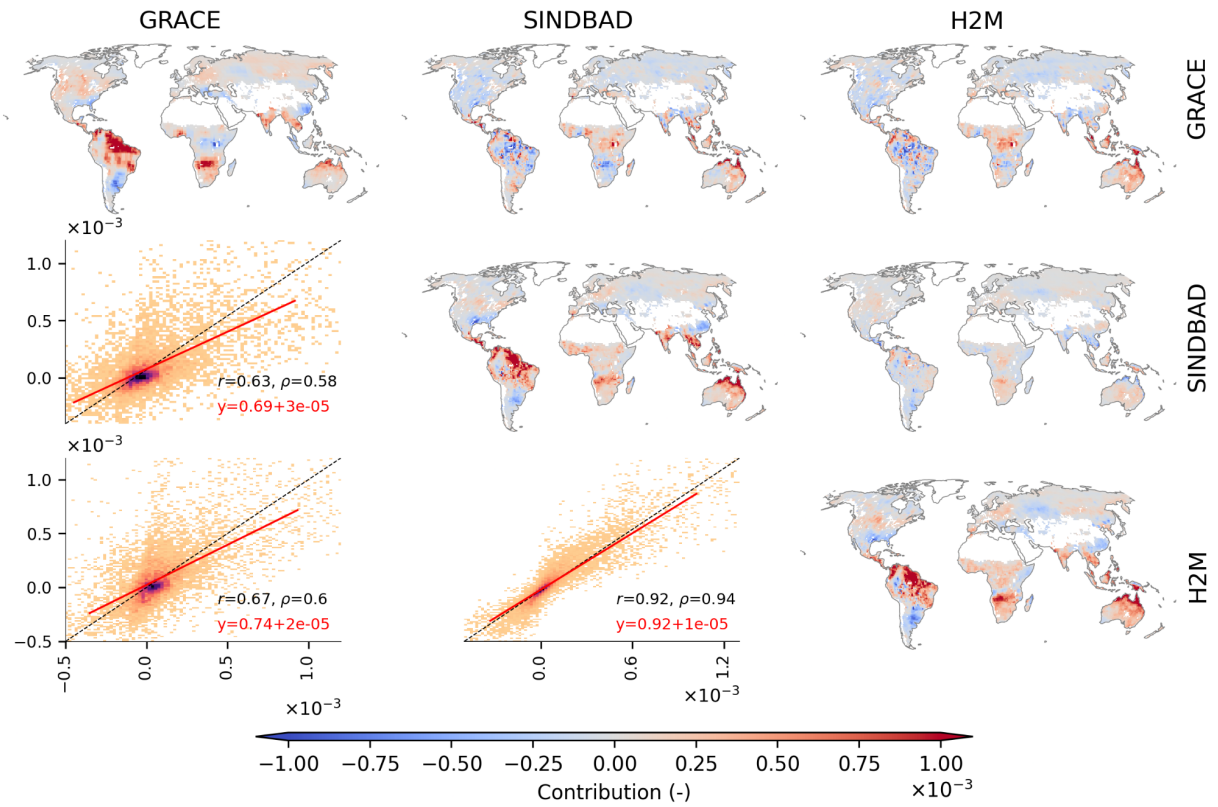

Figure 2. **The same as Fig. 3 in the manuscript, but using MSWEP for the precipitation forcing.** Global distribution of pixel-wise contributions to the variance of the global terrestrial water storage (TWS) IAV. Along the diagonal, maps of the pixel-wise contribution in GRACE, SINDBAD, and H2M are shown. Above the diagonal, maps of the difference (i.e., column - row) are shown. Below the diagonal, scatter plots comparing the corresponding column (x-axis) versus row (y-axis) are shown. In the scatter plots, colors indicate the density of points, r is the Pearson correlation coefficient, and ρ is the Spearman correlation coefficient. Red lines are linear regression fit and red texts are corresponding equations.

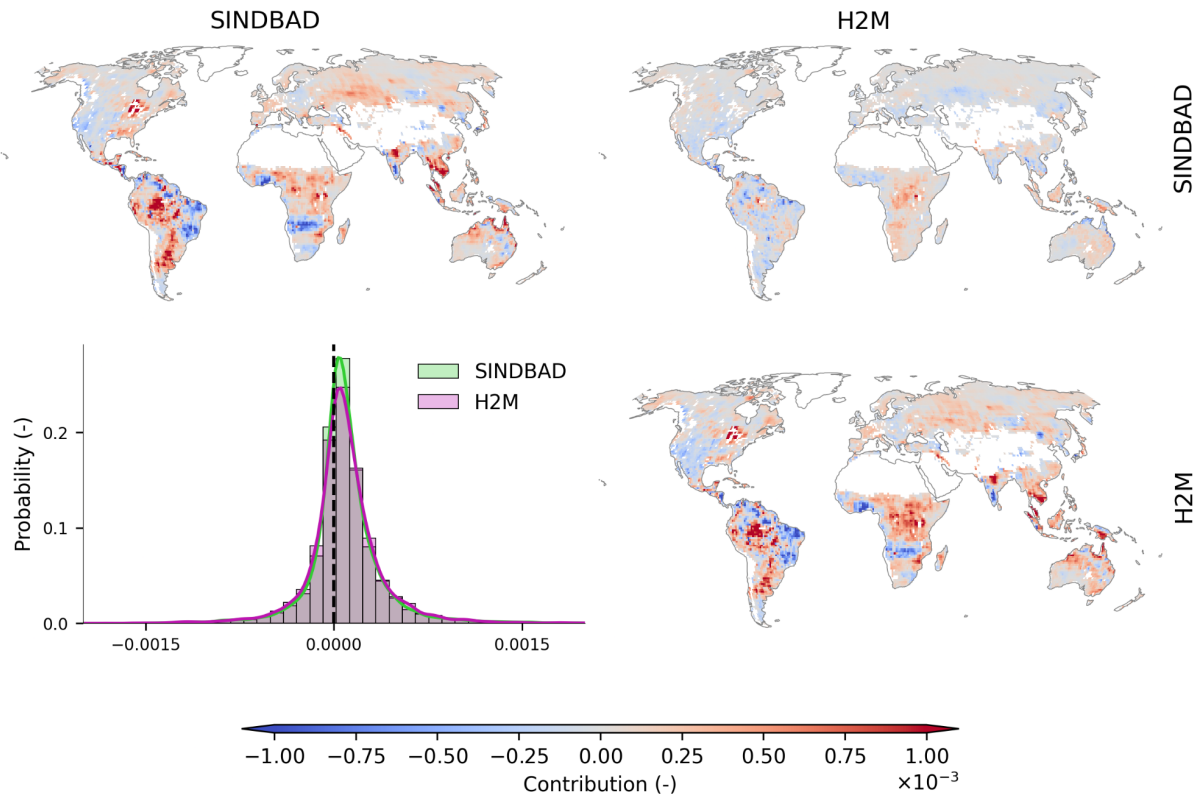

Figure 3. **The same as Fig. 4 in the manuscript, but using MSWEP for the precipitation forcing.** Global distribution of pixel-wise contributions to the variance of the modeling error of global terrestrial water storage interannual variability. Along the diagonal, maps of the pixel-wise contribution to the global TWS IAV modeling errors in SINDBAD and H2M are shown. Above the diagonal, a map of the difference (i.e., column - row) is shown. Below the diagonal, a histogram comparing the corresponding column (x-axis) versus row (y-axis) is shown. The probability density curves were estimated using the Gaussian kernel and the Scott's rule of thumb to determine the bandwidth of the kernel.

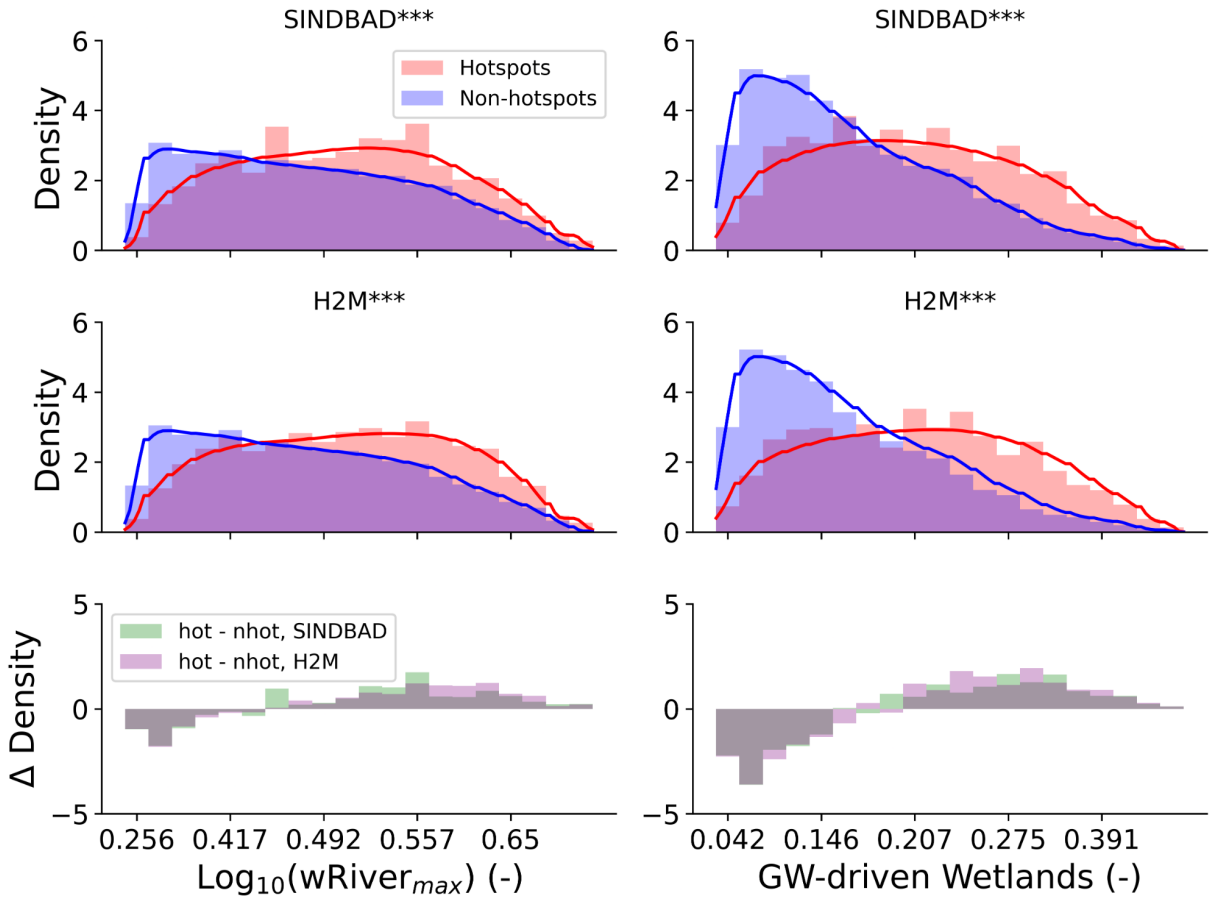

Figure 4. **The same as Fig. 7 in the manuscript, but using MSWEP for the precipitation forcing.** Comparison of probability density distributions of the log-transformed maximum river water storage (left), and wetlands fraction (right) between the error hotspot pixels and non-hotspot pixels. Top and middle lows are distributions of each model; the bottom row is the difference in bar heights between hotspot and non-hotspot (positive means occurrences are larger in error hotspots). The probability density curves were estimated using the Gaussian kernel and the Scott's rule of thumb to determine the bandwidth of the kernel. Asterisks (*) beside model names show the significance of the difference in distributions between error hotspots and non-hotspots using the Kolmogorov-Smirnov two-sample test; all results show significant difference in distributions (***, p-value < 0.001). Note that the x-axis is normalized using the maximum and minimum of variables so that the range becomes zero to one, and comparisons can be made across variables. The river water storage (wRiver) was calculated using the Total Runoff Integrating Pathways (TRIP) river routing model (Oki and Sud, 1999) with the input of runoff from SINDBAD. The maximum wRiver of a pixel ($wRiver_{max}$) during the entire period (April 2002–June 2017) was used with log transformation to use the skewed distribution of $wRiver_{max}$ for the comparison. The fraction of groundwater-driven (GW-driven) wetlands was provided by Tootchi et al. (2019).

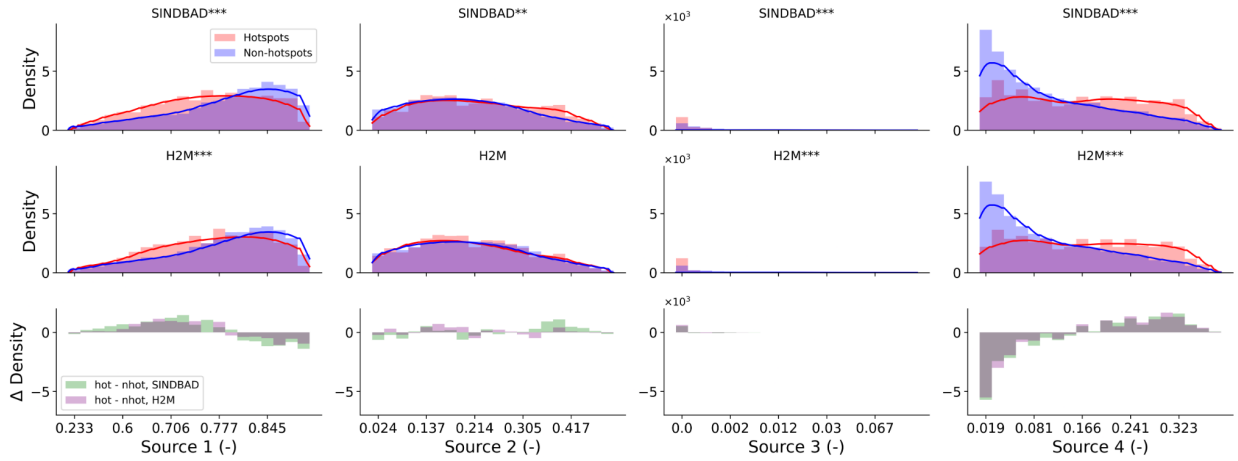

Figure 5. **The same as Fig. 8 in the manuscript, but using MSWEP for the precipitation forcing.** Same as Fig. 4, but for the contribution of four water sources to the water usage by vegetation. Data by Miguez-Macho and Fan (2021) was used for the four sources. Source 1 is soil water from recent (< 1 month) precipitation; source 2 is soil water from past precipitation; source 3 is locally-recharged groundwater via capillary flow; source 4 is remotely recharged groundwater from uplands to lowlands.

4. The abscissa and ordinate of the scatter plot in Figure 3 have no text description

AC: The text labels for the x and y axes were omitted for clarity and stated only in the figure caption. We further improved the related part of figure caption which now reads:

[...] Below the diagonal, scatter plots comparing the pixel-wise contributions of the corresponding column (x-axis) versus row (y-axis) are shown. [...]

5. How much different precipitation inputs affect the modeling error of global terrestrial water storage interannual variability? Does the precipitation input or the different model structure affect the simulation error more?

AC: As in the response to the comment 3, we agree that model input and structure are definitely important aspects that may potentially affect the results.

First, regarding the precipitation input, we show that the use of different precipitation products does not significantly alter the main results and findings of the study. This was shown in Kraft et al. (2022) as well.

Second, the potential effect of model structure is large. We assume that the two different models used in this study cover that source of uncertainty to a certain extent. Interestingly, we found that both models showed a large consistency in the findings despite having vastly different model structures with SINDBAD rooted on traditional hydrological concepts, and H2M formulated on modern machine learning methods.

Despite showing and presenting the effects of each of these factors separately, we cannot say with a large confidence if the uncertainty due to input is larger than that due to model structure or vice versa, especially based on what is presented in the current manuscript. For such an analysis, one would envisage a comprehensive factorial analysis of different modeling structures as well as the use of different input data but within a consistent seamless framework rather than comparison of two or more different models (as presented here, and in many model intercomparison projects to date). We will clarify this in the discussion with the following:

Lastly, the additional sources of errors include 1) anthropogenic influence, 2) uncertainties in the GRACE and forcing data, and 3) model structure. In Africa, a decrease in TWS IAV in 2003–2006 (Fig. 6c) was due to the expansion of Nalubaale Dam as well as La Niña (Stager et al., 2007; Awange et al., 2013, 2019). [...]

With respect to the uncertainty from forcing and model structure, we show that the use of different precipitation products does not significantly alter the main results and findings of the study (Fig. B14-B18). This was shown in Kraft et al. (2022) as well. In addition, the potential effect of model structure is large. We assume that the two different models used in this study cover that source of uncertainty to a certain extent. Interestingly, we found that both models, which have vastly different model structures, showed a large consistency in the findings. Despite showing and presenting the effects of each of these factors separately, we cannot conclude with a large confidence if the uncertainty due to input is larger than that due to model structure or vice versa, especially based on what is presented in the current manuscript. For such an analysis, one would envisage a comprehensive factorial analysis of different modeling structures as well as the use of different input data but within a consistent seamless framework.

, as well as in the introduction:

Both modeling frameworks are heavily rooted on using observations, and include GRACE observations in the model parameter estimation and evaluation. As such, under ideal conditions, the models have a potential to simulate aspects of hydrological cycle that agree the most with relevant observations, and the model errors, if any, could be attributed to either model structure (e.g., missing model processes or the way to formulate and connect processes) or observational uncertainties.