Author's response to Referee #2

The authors presented a study to diagnose the modeling errors by comparing GRACE and model TWSA based on IAV. The motivation of this study is nice, since Scanlon's PNAS study revealed an interesting question on the discrepancy between GRACE and models. The focus on interannual is a good complementary to the focus on trend by Scanlon et al.. Generally, this study is interesting. However, I have some critical questions related to the methods used for analysis, which may largely affect the reliability of the findings.

AC: We would like to thank the reviewer for a positive outlook of the study. Please find our responses below, which will also be included in the revised manuscript.

- **Generally, WGHM, PCR-GLOBWB, and maybe some other LSMs that include GW module, are more popularly used than the two models used in this study. I am not going to say the two models used here is not good enough, but I guess many researchers would be more interested on what will it like if we use WGHM, or PCR-GLOBWB, or CLSM. Besides, it is not clear how the including of GRACE in model parameter estimation and evaluation (Line 77) will impact the comparison between GRACE and the two models.**

  AC: While we agree that it will be interesting to look at the results of sophisticated land surface models or hydrological models, we focused on the two models because they are data-driven and represent the best-case scenario in terms of model performance against state-of-the-art observations. Note that Kraft et al. (2022) and Trautmann et al. (2022) conducted a direct comparison of SINDBAD and H2M, respectively, with global hydrological models from the eartH2Oserve ensemble (Schellekens et al., 2017), which reveal that the two models are at least in par or better than the GHMs in the Earth2O ensemble. So, the analysis presented in our study can be expected to be representative of other GHMs as well.

  Instead, the aim in our study was to go beyond the good performances of the two models used, and rather understand if there are underlying model assumptions and shortcomings that result in error of interannual variability of the global TWS. For this, we also needed the models to be forced by identical data, and be given a fair opportunity to learn from the observation. In fact, use of GRACE data in parameter estimation, theoretically, allows for the modeling framework to produce TWS simulations with no error. Note that this would not be possible with other GHMs/LSMs with uncalibrated parameters. We included this in the introduction and methods sections of the original manuscript and will improve the introduction in the revision as:

  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 provide the simulations that agree the most with observations, and the model errors, if any, could be attributed to either missing model processes or observational uncertainties. As we employ a covariance matrix analysis (see Sect. 2.1.2), we not only evaluate the global IAV, but also identify the regions that are most relevant to the IAV of global TWS in GRACE observations and the two models. SINDBAD and H2M are appropriate for this purpose as it requires models to be forced by identical data, and be given a fair opportunity to learn from the observation. In addition, as two models cover different aspects of modeling approach (i.e., process-based vs. physics-guided machine learning), using SINDBAD and H2M can cover the uncertainty of model structure to a certain extent.

Lastly, we, in fact, think that the framework presented in our study can be used for every model in the earth2Observe ensemble to identify the regions of largest TWS interannual variability and its error, which can then be utilized for model improvements. But, this is out of scope of the current study. Instead, we analyzed the error hotspots of TWS IAV modeling by four GHMs in the earth2Observe ensemble (W3RA, LISFLOOD, SURFEX-TRIP, and PCR-GLOBWB) as shown in Fig. 1 below. Four models show strong positive contributions in the humid regions of northern South America as SINDBAD and H2M do, but the four models disagree with SINDBAD and H2M for some aspects like hotspots in Central Africa. However, a direct comparison of the results of SINDBAD and H2M with those of the four GHMs are not reasonable because different products of forcing were used, and observations were either not used at all, or only used for model validation. We include Figure 1 as appendix, and we will address the issue in the discussion as:

[...] In SINDBAD, vegetation indirectly accesses secondary 435 water storage with capillary rise, which contributes to a larger evapotranspiration over some regions, including the regions in Africa (Fig. 9 in Trautmann et al., 2022).

As shown above, the model structure is an influential uncertainty of the location of hotspots. Though using SINDBAD and H2M cover the uncertainty to a certain extent, more sophisticated hydrological models may have different hotspots and sources of errors. Following Kraft et al. (2022), among 10 global hydrological models in the eartH2Observe ensemble, we selected four GHMs with groundwater storage in the structure to identify the hotspots of TWS IAV modeling error: W3RA (Van Dijk and Warren, 2010), LISFLOOD (Van Der Knijff et al., 2010), SURFEX-TRIP (Decharme et al., 2010, 2013), and PCR-GLOBWB (Van Beek et al., 2011; Wada et al., 2014). In Fig. B19, the four GHMs show strong positive contributions in the humid regions of northern South America as SINDBAD and H2M do, but the four models disagree with SINDBAD and H2M for some regions like hotspots in Central Africa. However, as different sets of forcing and constraints with different spatiotemporal domains were used for the simulation of the four GHMs, and given the complexity of their structure, further research will be required to investigate the hotspots and sources of TWS IAV modeling errors of each GHM.

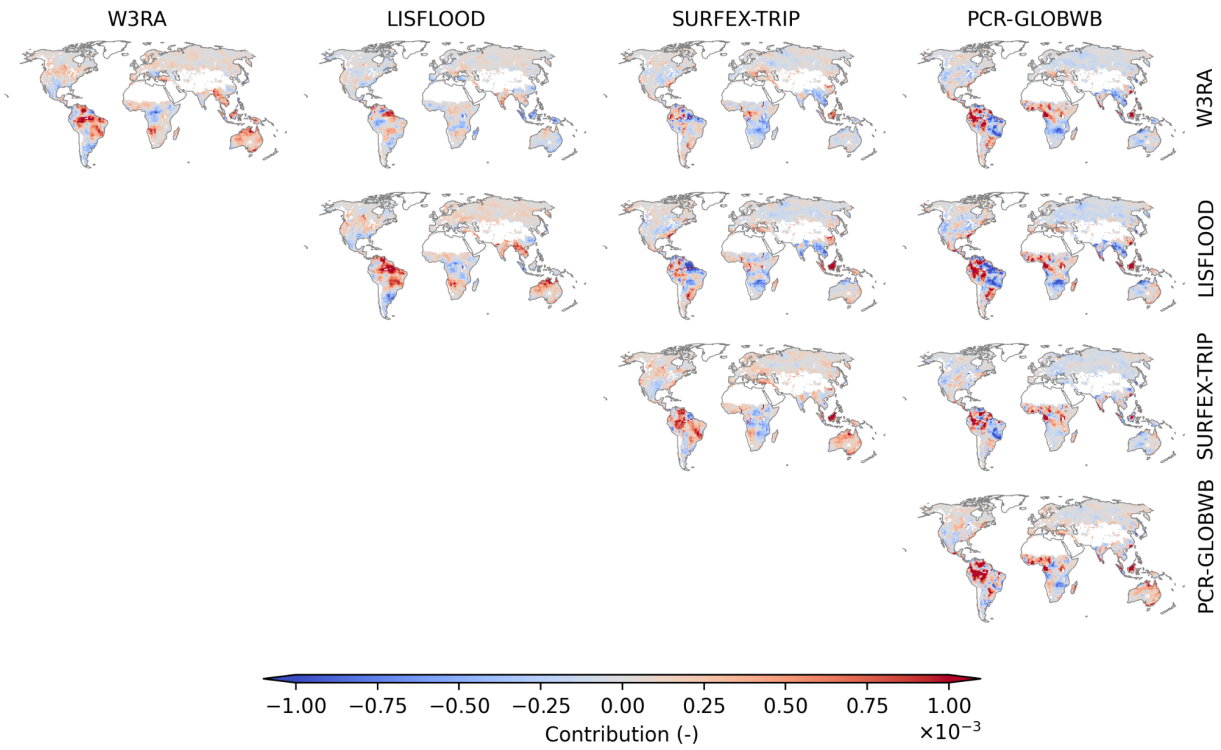

Figure 1. **The same as Fig. 4 in the manuscript, but for four global hydrological models in the eartH2Observe ensemble.** Global distribution of pixel-wise contributions to the variance of the modeling error of global terrestrial water storage interannual variability. Along the diagonal, maps of the pixel-wise contribution to the global TWS IAV modeling errors in W3RA, LISFLOOD, SURFEX-TRIP, and PCR-GLOBWB are shown. Above the diagonal, a map of the difference (i.e., column - row) is shown.

- **It is not clear why using Equation (1) to derive the IAV for analysis. I cannot understand the physical meaning of subtracting long-term trend (fit ()) from monthly values. So, the question comes that what is interannual variability, and how to define it? Can we just subtracting long-term average from monthly values? I am not sure my understanding is correct or not. Please verify it.**

  AC: Thank you for pointing this out. It is important to define a term clearly and evaluate the model at a proper aspect. First, we think that subtracting the long-term average from monthly values is not suitable because it cannot remove the trend. We want to remove the trend because SINDBAD and H2M do not properly account for the trend as it is significantly driven by human activities (Rodell et al., 2018; Scanlon et al., 2018) and long-term processes such as vegetation (Pokhrel et al., 2021) and glacier melt (Rodell et al., 2018; Scanlon et al., 2018). Instead, by interannual variability, we want to quantify how much each value deviates from the seasonal

mean condition including the trend as Fig. 2 illustrates below. This definition of IAV will also make this study more suitable as a complementary to the study by Scanlon et al. (2018) as well as other relevant studies (e.g., Jung et al., 2017; Humphrey et al., 2018). To this goal in mind, for each month of a year, we calculated the linear regression fit that represents the seasonal mean value including trend. We then got the IAV by subtracting the fitted value from each monthly TWS.

For clarification, we will add Fig. 2 as appendix and the definition of IAV above in the manuscript as follows:

In this study, IAV quantifies how much a value (e.g., TWS) deviates from the seasonal mean including the trend. Accordingly, we calculated the globally integrated GRACE TWS IAV as follows: [...]

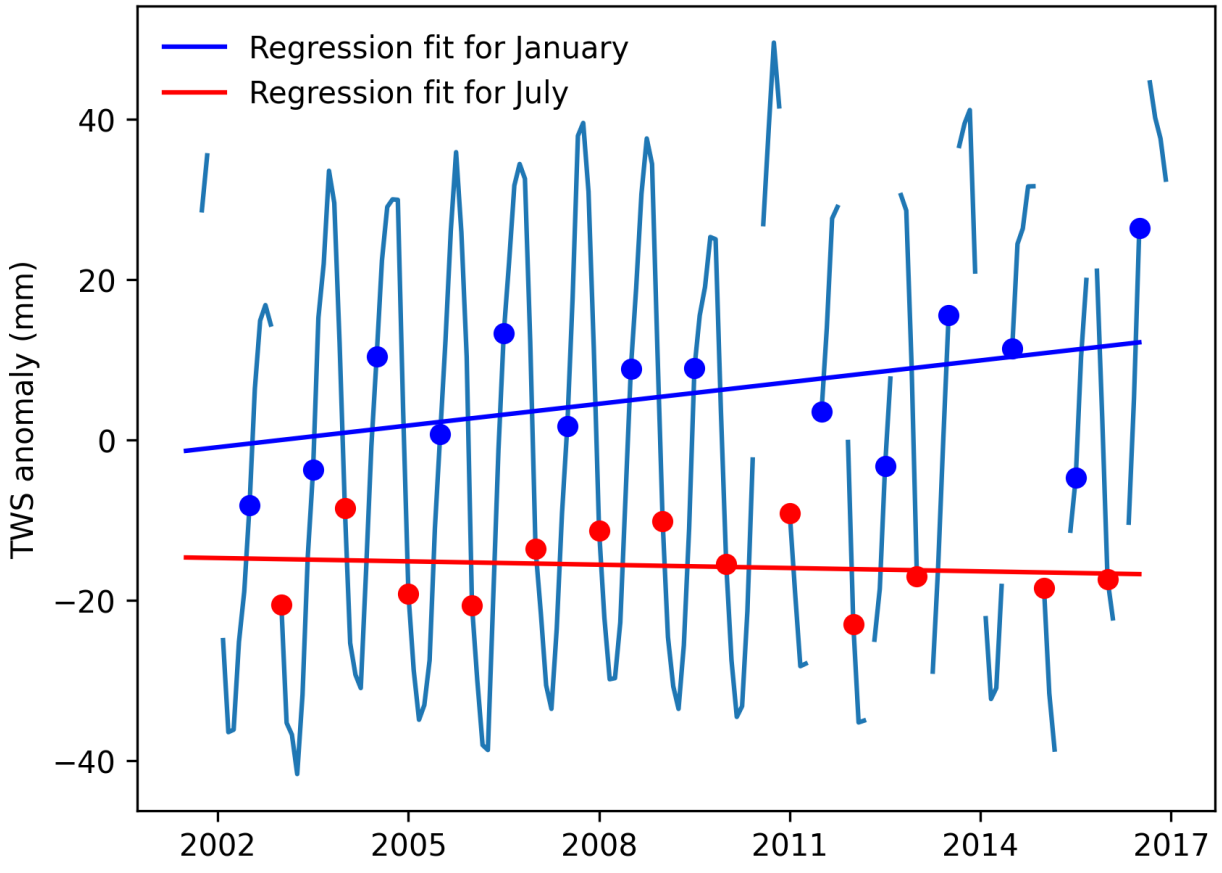

Figure 2. Illustration of the calculation of interannual variability for the global terrestrial water storage (TWS) anomalies.

- **Since GRACE Level-3 data has been already processed by subtracting the mean of a period (2004-2009?) from monthly TWS to get TWSA. If the authors again do subtracting (2002-2017) for GRACE and models, it may lead to mismatch between GRACE and model, because different subtracting were done for GRACE (subtracting 2004-2009, and then subtracting 2002-2017) and models (subtracting 2002-2017).**

  AC: The reviewer is correct that the time period used to calculate the TWS anomaly is crucial and different time periods would affect the comparison between GRACE data and models. Exactly due to this difference, a suggested necessary step in the GRACE data usage is to align both GRACE and modeled TWSA to the same time mean anomaly (https://grace.jpl.nasa.gov/about/faq/), which, in the study, is 2002-2017. By removing the time period 2002-2017 from each series, both time series of TWS anomalies are consistent and reflect the deviations from the same baseline condition.

- **Line 128: I am not sure it is the best way to evaluate model performance by comparing the IAV derived from GRACE and models. How about compare TWSA?**

  AC: We agree that TWSA can be evaluated as well, but as clearly mentioned in the introduction, the main aim of the study is to diagnose the error in IAV of TWS, which is still reproduced relatively poorer in the models compared to TWSA. The evaluation of TWS IAV presented here complements previous studies evaluating the anomalies (e.g., Scanlon et al., 2018) with climatic processes, and trends with anthropogenic influences. We also note that the original TWSA includes the trend that should be removed before the model evaluation as SINDBAD and H2M do not account for important processes that affect the trend, e.g., human influences.

- **Before Figure 2, people may be interested on seeing spatial distribution map of TWSA from GRACE and models, as well as the distribution map of IAV, which both can help we better understand the difference and consistence between GRACE and models.**

  AC: Thank you for the suggestion. We will add figures for spatial distribution of TWSA (Fig. 3, below) and TWS IAV (Fig. 4, below) to appendix, and will mention them in the manuscript as follows:

  SINDBAD and H2M reasonably reproduce the observed time series of global TWS IAV by GRACE (R^2 of 0.49 and 0.51 for SINDBAD and H2M, respectively) (Fig. 2), as well as the spatial pattern of TWS anomaly and TWS IAV (Fig. B20 and B21). [...]

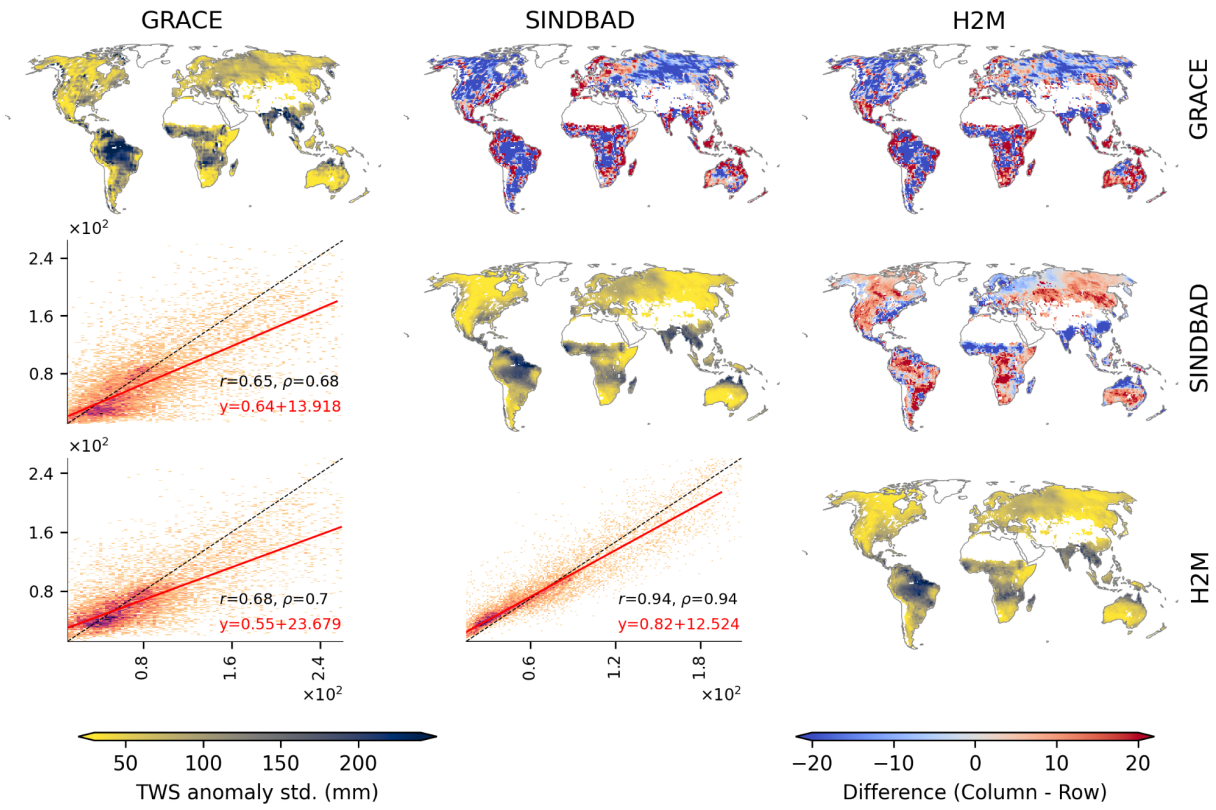

Figure 3. Global distribution of the standard deviation (std) of the global terrestrial water storage (TWS) anomalies. Along the diagonal, maps of the pixel-wise std of TWS anomalies in GRACE, SINDBAD, and H2M are shown (indicated by the label of row or column). Above the diagonal, maps of the difference (i.e., column - row) are shown. For example, the map of the first row and the second column is for SINDBAD (column) minus GRACE (row). 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. White pixels within land boundaries in maps are invalid as they are out of the study area.

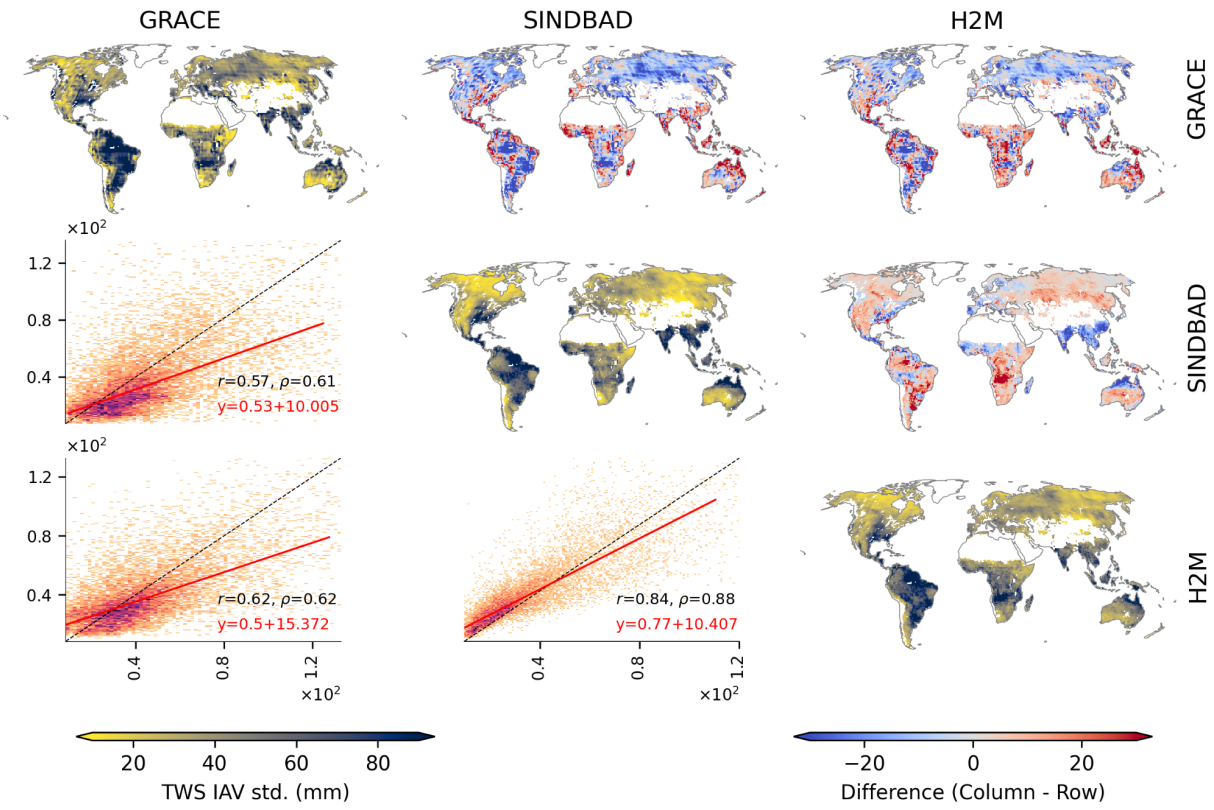

Figure 4. Global distribution of the standard deviation (std) of the global terrestrial water storage (TWS) interannual variability (IAV). Along the diagonal, maps of the pixel-wise contribution to the global TWS IAV modeling errors in SINDBAD and H2M are shown (indicated by the label of row or column). Above the diagonal, maps of the difference (i.e., column - row) are shown. For example, the map of the first row and the second column is for SINDBAD (column) minus GRACE (row). 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. White pixels within land boundaries in maps are invalid as they are out of the study area.

- **Figure 3: Sorry, but I do feel difficult to understand what the exact meanings of the spatial maps are. Maybe more information can be added to the figure showing who minus who, something like that. Besides, I guess the white blank areas here are the grid cells with positive covariances, is it true?**

AC: Thank you for pointing out the confusion. The detail (i.e., who minus who) was omitted for clarity, but was explained in the caption. We will improve the caption of Fig. 3 and 4 in the manuscript as follows:

Figure 3. 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 (indicated by the label of row or column). Above the diagonal, maps of the difference (i.e., column - row) are shown. For example, the map of the first row and the second column is for SINDBAD (column) minus GRACE (row). [...] Red lines are linear regression fit and red texts are corresponding equations. White pixels within land boundaries are invalid as they are out of the study area.

Figure 4. Global distribution of pixel-wise contributions to the variance of the modeling error of global terrestrial water storage interannual variability. Along the map of the first row and the second column is for SINDBAD (column) minus GRACE (row). [...] White pixels within land boundaries are invalid as they are out of the study area.

We will also add the same sentence about white pixels to captions of relevant figures such as Fig. B3 and B12 in the manuscript and ones that will be added in the revised manuscript.