Author response to Referee #1 comments from August 30th, 2016

We are thankful to Referee #1 for his valuable comments and suggestions, which will certainly improve our manuscript. In the following, the response to the individual comments is given with some new figures at the end. The original review is quoted in italics, whereas the author response is given in normal font.

This manuscript presents results of comparing estimated terrestrial water storage (TWS) from four hydrological models with GRACE derived TWS in 31 hydrological basins. Four metrics were used in evaluating model performances. Components of TWS as well as actual and potential ET were examined in selected basins to show the impact of model physics on estimated TWS.

The results and discussions are generally well presented and justified. But I think the paper can be further improved in a few areas. For instance, the fact that three of the four models do not model groundwater, which may contribute significantly to TWS changes, is not explicitly mentioned and discussed in the paper. In addition, the four metrics used in evaluation may be good for summarizing the differences but they do not necessarily reflect the actual discrepancies between modeled and GRACE derived TWS. For instance, the amplitude and phase differences may not be important if TWS exhibits strong inter-annual variability.

-Except the annual amplitude and phase differences, we also show the explained variance with the seasonality removed to evaluate the agreement of models with GRACE in terms of inter-annual variability.

Additional comments: Page 3, data set. Please emphasize the fact that three of the four models do not simulate groundwater and discuss its potential impacts on model estimated TWS in the result section. Also, do these models account for anthropogenic impacts such as groundwater abstraction? If not, how would this affect the comparison with TWS from GRACE which does detect changes associated with groundwater withdrawals?

-Thanks for the suggestion, and this is indeed an important information. In line 7 on page 3 of the paper we propose to include the following sentences: Some of the main characteristics of the four numerical models are presented in Table 1, which provide more information on how models are different with each other. For instance, although soil moisture and snow water are included in all models, surface water and groundwater are simulated differently. JSBACH is the only model which does not include surface water. Groundwater is simulated by WGHM, where the anthropogenic impact such as groundwater abstraction is also considered. JSBACH does not include groundwater explicitly. However, soil moisture in deep layers below the root zone is simulated and buffers extreme soil moisture conditions in the layers above. Thus, some of the characteristics of real groundwater are considered. We use the term subsurface water for both soil moisture and groundwater. But the impact from consideration of groundwater to TWS variations in WGHM will be investigated in the following discussion.

A Fig. 1 will be added in the manuscript showing the differences of explained variances from WGHM with and without groundwater. The positive values indicate that WGHM with groundwater exhibits better agreement with GRACE than the one without groundwater. The large impact mainly locates at basins such as Toscantins, Niger, Huang He, Mekong and Mississippi. Only in three basins: Lena, Indus and Yukon, the differences are negative.

Page 4, Line 8: I understand why you removed the trend but the ability to predict trend is also an important part of the models. Can you provide a scatter plot comparing trends from the models and those from GRACE in the 31 basins?

-A scatter plot comparing trends from the models and those from GRACE in the 31 basins is shown in Fig 2, which will also be included into the final paper . The TWS trends from various models do perform quite differently among each other and with GRACE.

Page 5 Line 10, should the second "GRACE" be TWS?

-Yes, and it has been changed accordingly.

Page 6. Line 20, I don't think it is appropriate to compare GRACE errors with the RMSE since the former represents instrument and post-processing errors and has nothing to do with how well models perform.

-The GRACE errors are calculated to indicate the TWS uncertainties from GRACE, which can be applied to indicate about where GRACE might be suitable as a validation tool for models and where not. Special attention should be paid to basins with

large GRACE errors, as the large discrepancies could be related to large GRACE TWS uncertainties, but not to model differences. We would like to stress that we do not directly compare GRACE errors with RMSE, but that we use the GRACE errors only as indicator of observation uncertainty. The way to estimate GRACE errors is introduced in Zhang et al. (2016). The error estimation is also investigated through an end-to-end simulation performed by Flechtner et al. (2016). We thus believe that the errors we calculated are plausible.

In addition, basin-scale GRACE errors are smaller than the gridded errors which are spatially correlated. Did you consider spatial correlation of errors (in both modeled and GRACE TWS) when calculating basin-scale RMSEs between the model and GRACE? Either way, I think it only makes sense to compare RMSEs among the models.

-The correlation between the gridded errors from GRACE is much larger than the one from models and is considered by using the squared exponential covariance function to estimate the statistical covariance between two grids as proposed by Landerer and Swenson (2012). The error estimates from the gridded data set also show consistent results with the ones derived directly from Stokes coefficients.

The statistics in Table 2 shows the models generally did not performed well in the tropical climate. Why is that? Does it have something to do with runoff estimates as ET is energy limited in this type of climate? You don't necessarily need to collect in situ stream flow data, but some discussions and plots on runoff may be needed to explain this result.

-The large RMSE values in tropical regions are partly related to the fact that the TWS variability in this region is comparably large. Besides, the runoff comparison is shown for three basins affected by the tropical climate. It is seen that the bad performance of a certain model is connected with its differently simulated runoff. At Amazon basin, the positive runoff simulated from MPI-HM also leads to comparably small variability in TWS. At Zaire basin, the large inter-annual variations in TWS from LSDM are just corresponding to its runoff in an opposite way. At Mekong basin, the much larger amplitude in TWS from JSBACH compared with GRACE is related to the apparent large amplitude in its runoff.

Page 7. Line 5 to 12. As you pointed out that AET does not have a significant impact on TWS in humid areas, then what is the purpose of including three basins from that climate in Fig. 7? I think including more basins from drier climate is more useful here.

-This is true. However, as we add the figures of runoff comparison, three basins in tropical zone and three basins in dry climate are chosen.

Page 7. Line 13-22, I didn't learn anything from this paragraph and it can be removed. As you correctly pointed out that AET may be significantly different from PET which does not help much in explaining the result. Again, I think presenting runoff estimates is more useful.

-The paragraph will be shortened and we will replace the PET figures with runoff comparison.

Figs. 2-5: It would be very helpful if you provide time series of TWS for a basin(s) with the largest deviation from GRACE in either of these metrics. For instance, it's hard to visualize how significant a 45 degree difference in phase is. In addition, two of these metrics measure differences in seasonality which may not mean much when the interannual variability of TWS is much stronger. So providing actual TWS time series along with some discussions will be helpful for readers to understand the usefulness and limitation of these metrics.

-Fig. 4 showing the time series of TWS for basins with the largest deviation from GRACE will be added in the manuscript along with some discussions.

Some sentences will be added: As each metric usually focuses only on one specific property of statistical performance and has its own limitations, the time series of TWS are shown for some basins with the largest deviation between GRACE and the model. The TWS time series are shown for Yukon basin, where both WGHM and JSBACH exhibit the largest deviation of annual amplitudes from GRACE. Although the annual amplitude is simulated better by LSDM and MPI-HM, apparent negative phase differences are shown. Amur basin is also shown, as LSDM, WGHM and MPI-HM all have the largest negative phase differences with GRACE here. Models generally capture the inter-annual signals but perform quite differently among each other and with GRACE in terms of seasonality. Almost opposite phase differences are shown for these models. The smallest

explained variance for MPI-HM happens at St. Lawrence basin, where a much larger amplitude and a negative phase difference compared with GRACE are shown. When the annual signal is removed, models perform differently in terms of the explained variance. In Nile basin, large inter-annual variations simulated by LSDM lead to even negative explained variance compared with other models.

Fig. 7. I think including runoff instead of PET is more appropriate here. Also, please try to use the same y-axis range for all plots which makes it to compare the magnitude of TWS and ET.

-The y-axis range will be changed and the runoff comparion is also shown (Fig. 3).

Fig. 8, "Subsurface water" should be soil moisture + groundwater storage for WGHM and soil moisture for all other models. Again, please use the same range for all y-axis if possible. In the caption, snow water content should be snow water equivalent.

-This will be changed accordingly.

References

Flechtner, F., Neumayer, K.-H., Dahle, C., Dobslaw, H., Fagiolini, E., Raimondo, J.-C., and Güntner, A.: What Can be Expected from the GRACE-FO Laser Ranging Interferometer for Earth Science Applications?, Surveys in Geophysics, 37, 453–470, doi:10.1007/s10712-015-9338-y, http://dx.doi.org/10.1007/s10712-015-9338-y, 2016.

Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resour. Res., 48, doi:10.1029/2011WR011453, W04531, 2012.

Zhang, L., Dobslaw, H., and Thomas, M.: Globally gridded terrestrial water storage variations from GRACE satellite gravimetry for hydrometeorological applications, Geophys J Int., 206, 368–378, doi:10.1093/gji/ggw153, 2016.

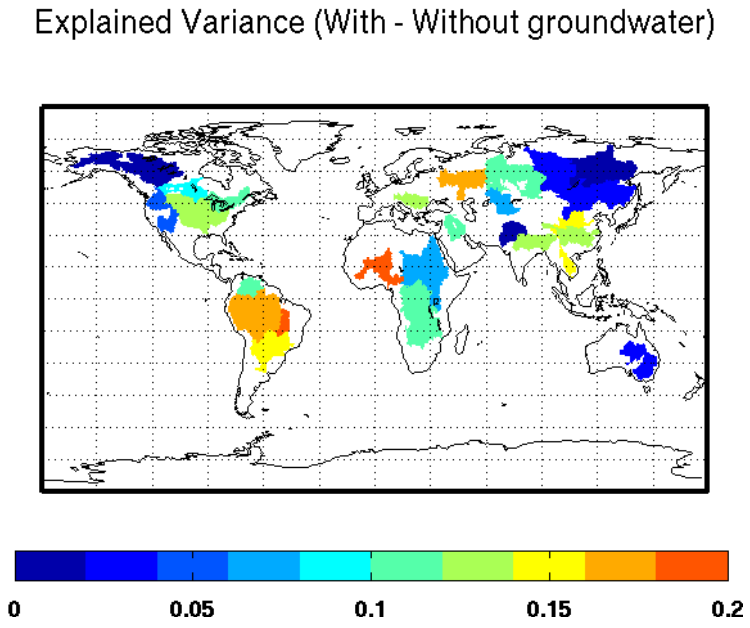

Figure 1. The differences between the explained variance values from WGHM with and without groundwater.

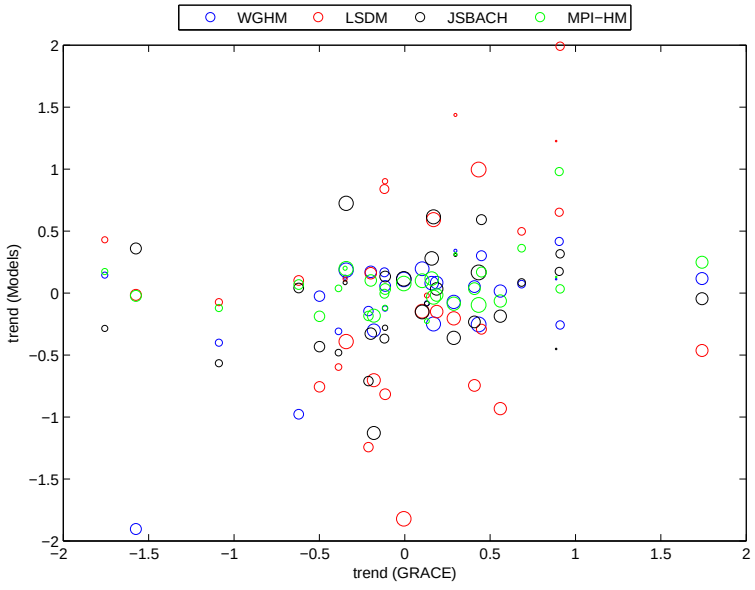

Figure 2. The scatter plot comparing trends from the models and those from GRACE in the 31 basins. Symbol size varies with rive basin area.

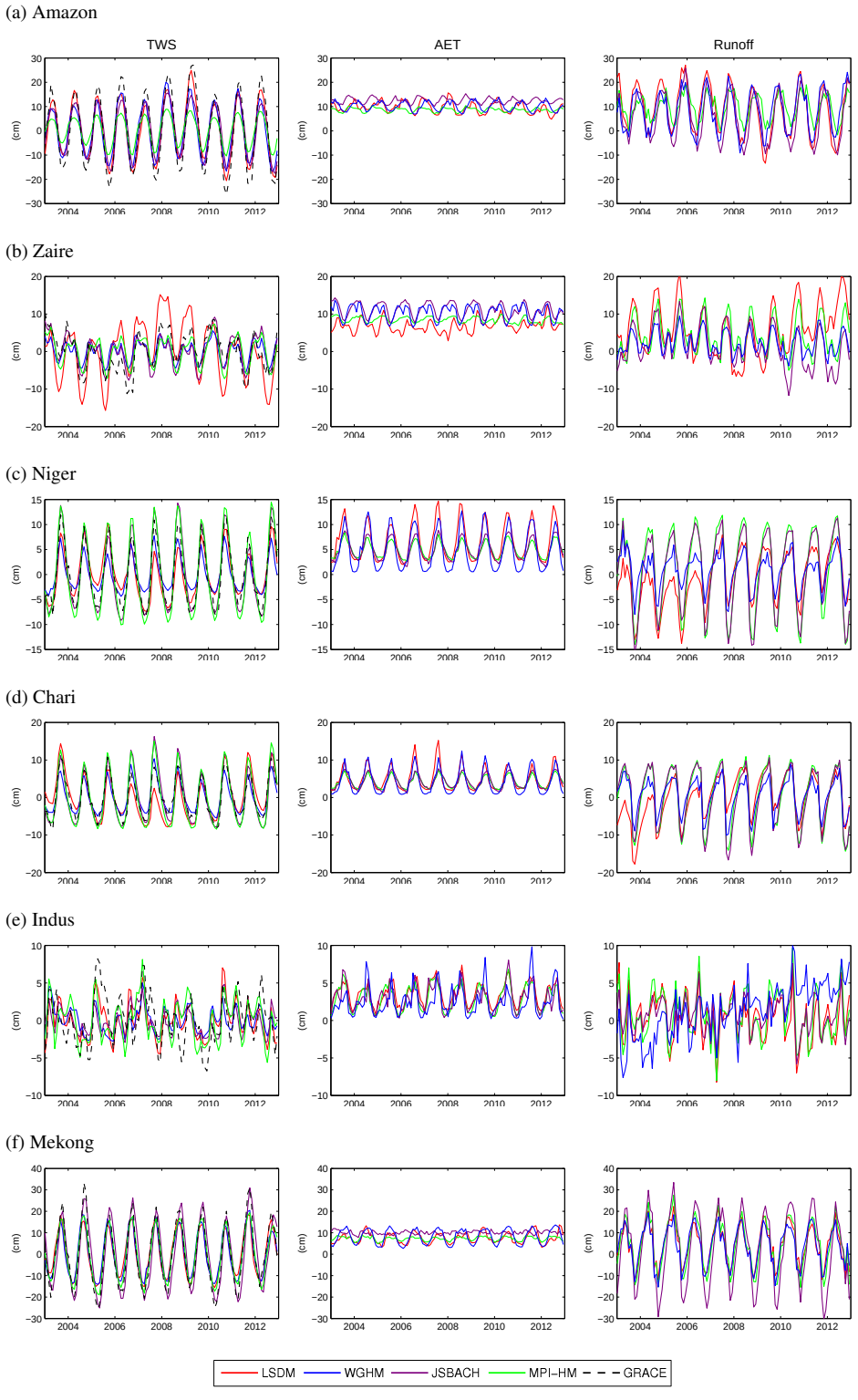

Figure 3. Time series of TWS (left column) from GRACE and models, model simulated AET time series (second column) and model simulated runoff time series (third column); each for six different catchments: Amazon, Zaire, and Mekong in tropical zone; Niger, Chari and Indus in dry zone.

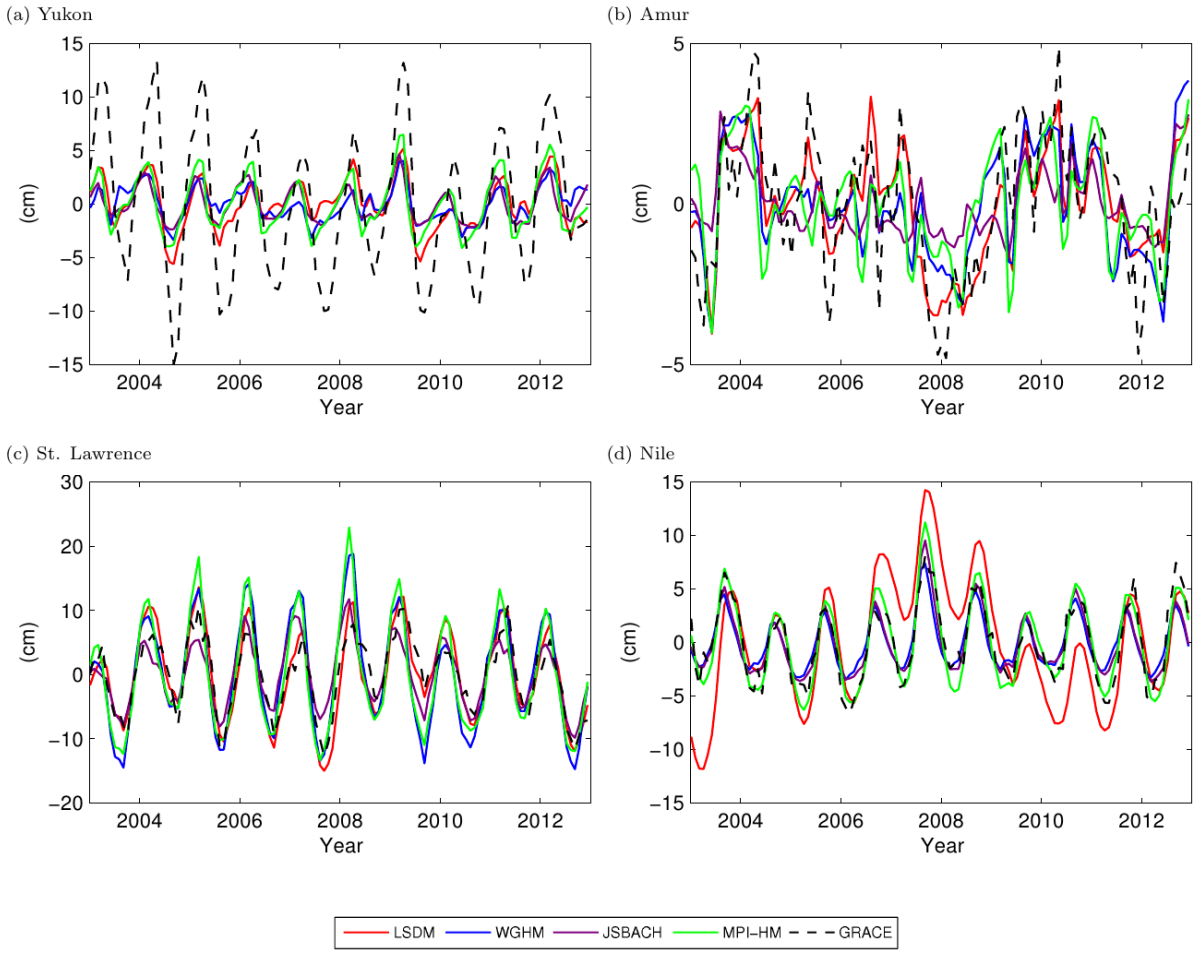

Figure 4. Examples of monthly TWS time series from GRACE and models for the basins with the largest deviation between model and GRACE in each of the four metrics: Relative amplitude differences (Yukon), phase differences (Amur), explained variance (St. Lawrence) and explained variance with annual signal removed (Nile).