Review#3

The article evaluates S2S TWS forecasts produced from FLDAS over Africa using gravity observations. I think the article reads well and I think it stretches the surface of a relatively unexplored area. That is, it highlights the importance of improving model physics of groundwater as well as it relevance for S2S forecasts. I want to also say that the authors motivate the GRACE community to reduce latency on their products, as GRACE-DA could be beneficial to improve the forecast. I'd encourage the authors to add some of these cavetas in the conclusions section. Beside this "major" comment, I, here, list only minor suggestions for the authors.

Thank you for your supportive comments. We revised the final section extensively to highlight the importance of improving groundwater simulation for enhancing TWS forecasts. We also emphasize the need for reducing GRACE data latency to benefit TWS forecast in the last section.

2.2. > it is unclear to me what NMME models are, at around line 111, please add a brief broad description of why they are needed here. Also unclear what and why the downscaling is needed. Table 1> what variables of these models are used?

Background information on NMME models and why they are needed have been provided in section 2.2 as the following: "*To generate TWS hindcasts, atmospheric forcing fields must be obtained from hindcast products to properly represent forecast uncertainty, rather than from reanalysis which cannot predict future weather events. Unlike reanalysis, meteorological hindcasts are produced by climate models without constraints of observations and therefore, are subject to larger uncertainties. FLDAS-Forecast employs a suite of NMME models developed by multiple institutions to provide S2S precipitation (and temperature which is not currently used by FLDAS-Forecast) forecasts (Table 1). The ensemble approach not only enables uncertainty quantification but also generally yields higher predictive skill than any single model (Wood et al., 2002; Kirtman et al., 2014)*".

Downscaling is needed because of the coarse spatial resolutions of the hindcasts, "*NMME precipitation hindcasts are provided as monthly data on a 1° global grid, while non-precipitation GEOS hindcasts are provided at 0.5° in latitude by 0.625° spatial resolution. All meteorological hindcasts are bias-corrected and spatially downscaled to the 0.25° resolution using CHIRPS and MERRA-2 data, respectively, and further temporally disaggregated using LIS built-in functions (Arsenault et al., 2020; Hazra et al., 2023)*".

Only the precipitation field from NMME is used in the FLDAS forecast system. We updated the table caption to clarify this.

Line 138: Typo CLMS

Corrected.

Line 186: are the percentiles computed using seasonal mean? Please clarify in manuscript

Yes. This has been clarified in section 2.5 where an equation for computing percentiles has been added.

Fig 4. Why is the correlation so small already at 1-month lag time? How significant are these statistics?

Thanks for the question. As shown in the spatial correlation maps, there are regions of high correlations (see below). However, low and even negative correlations in certain areas (due to the opposite trends in simulated and GRACE observed TWS) suppressed domain-average correlations. We have updated the correlation and other maps using grey to mask out groundwater depletion regions and a triangle is used for the low end of the label bar (see below) to highlight that any values lower than -0.1 are shown in white.

The manuscript has also been updated to say, *"In CAR and South Sudan, TWS hindcasts from both models show near and below zero correlations with GRACE/FO data, due to the opposite trends between reanalysis TWS and GRACE/FO data in that region (Supplementary Fig.S2)"*.

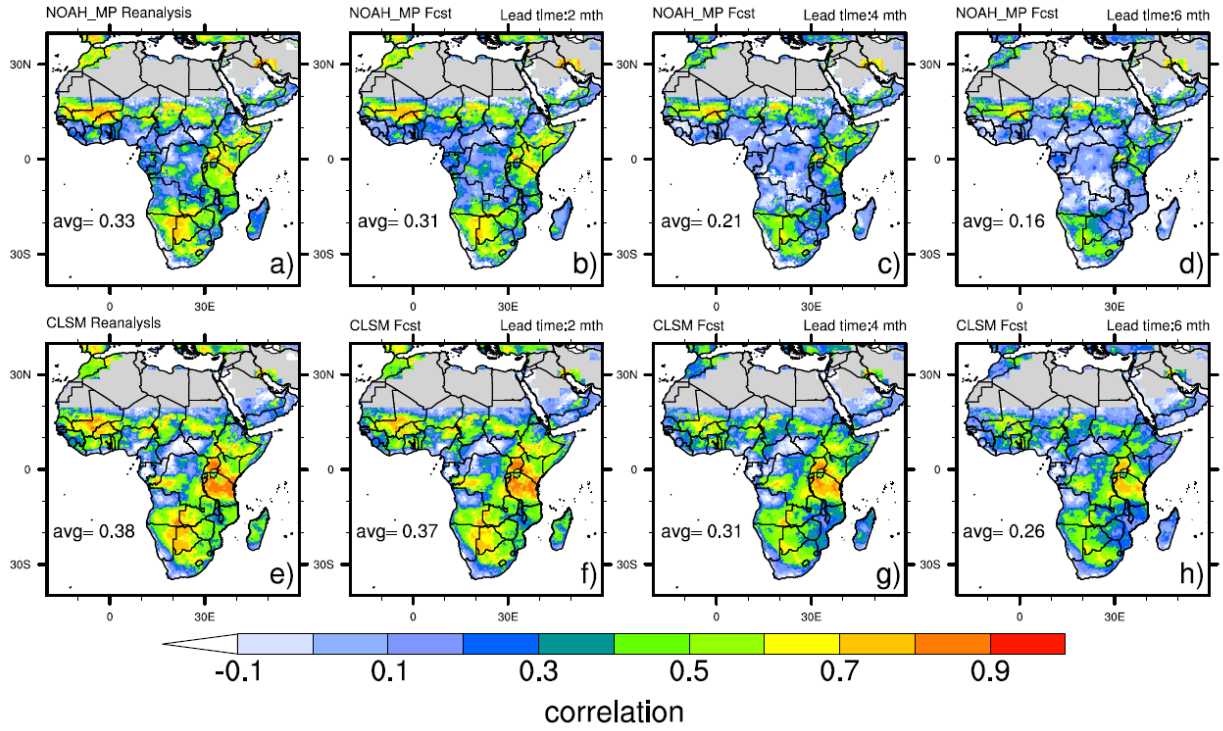

Fig.4 Correlation between non-seasonal reanalysis TWS and ensemble mean TWS forecasts of all NMME models at three lead times, and GRACE TWS observations for Noah-MP (top row) and CLSM (bottom row). Domain average correlations are shown in inset text.

We didn't do significance test as the purpose was to compare the correlation between the two models. Significance test also requires temporal independence of the data, which is not the case for TWS.

Line 308 – 324: it is unclear in my option what this analysis is really telling us. What is ROC and why is it computed only on the lower tercile (drier forecasts?)? Please add some general background on the metric and its interpretation.

Before discussing ROC results, we added the reason why we use ROC scores, "*While RMSEs and correlation quantify the magnitude of discrepancies and the temporal consistency between two time series, they do not directly assess the ability to accurately forecast wetter and drier conditions. Therefore, we use ROC scores to evaluate the performance of Noah-MP and CLSM in predicting terciles, corresponding to below-normal, near-normal and above-normal conditions*".

In addition, we describe ROC in the data section (section 2.6), "*Additionally, skill in forecasting terciles is assessed using the relative operating characteristic (ROC) score, a commonly used evaluation metric measuring the ratio of hit rates to false alarm rates (Met Office). A ROC score of 1 indicates a perfect forecast. ROC scores below 0.5 suggest no*

skill, while scores above 0.6 indicate predictive skill (Met Office). High ROC scores and strong correlation are commonly interpreted as indication of skillful forecasts (e.g., Yuan and Zhu, 2018)".

The spatial map for upper terciles is provided in the supplementary file (see Line 316 of the manuscript) as it is very similar to that of lower terciles.

Fig 6 > wouldn't be useful to also show difference maps of the first two rows wrt GRACE (bottom)?

Thanks for the suggestion. We now show differences in persistence (see below image) and added the following text to the manuscript," *Differences between simulated and GRACE/FO TWS persistence exhibit spatial patterns similar to those of mean annual precipitation, reflecting the strong influence of precipitation and associated uncertainty on persistence (Fig.7, bottom two rows). However, the two models often exhibit contrasting performances. Compared to GRACE/FO TWS, Noah-MP underestimates persistence in central Africa and overestimates it elsewhere. In contrast, CLSM overestimates persistence in central Africa while underestimating it in other regions, a discrepancy that is more pronounced at the 2- and 4-month lags. In wetter central Africa, strong interannual variability in CLSM TWS helps retain past wetness conditions and contributes to its enhanced persistence. On the other hand, the underestimation of persistence by CLSM in drier regions may be linked to the model's tendency to overestimate ET, which acts as continuous disruption to soil moisture states, thus leading to low persistence. For both models, discrepancies between reanalysis and GRACE/FO increase with increasing lags (Fig.7, bottom two rows), reflecting cumulated differences in their abilities to retain past states".*

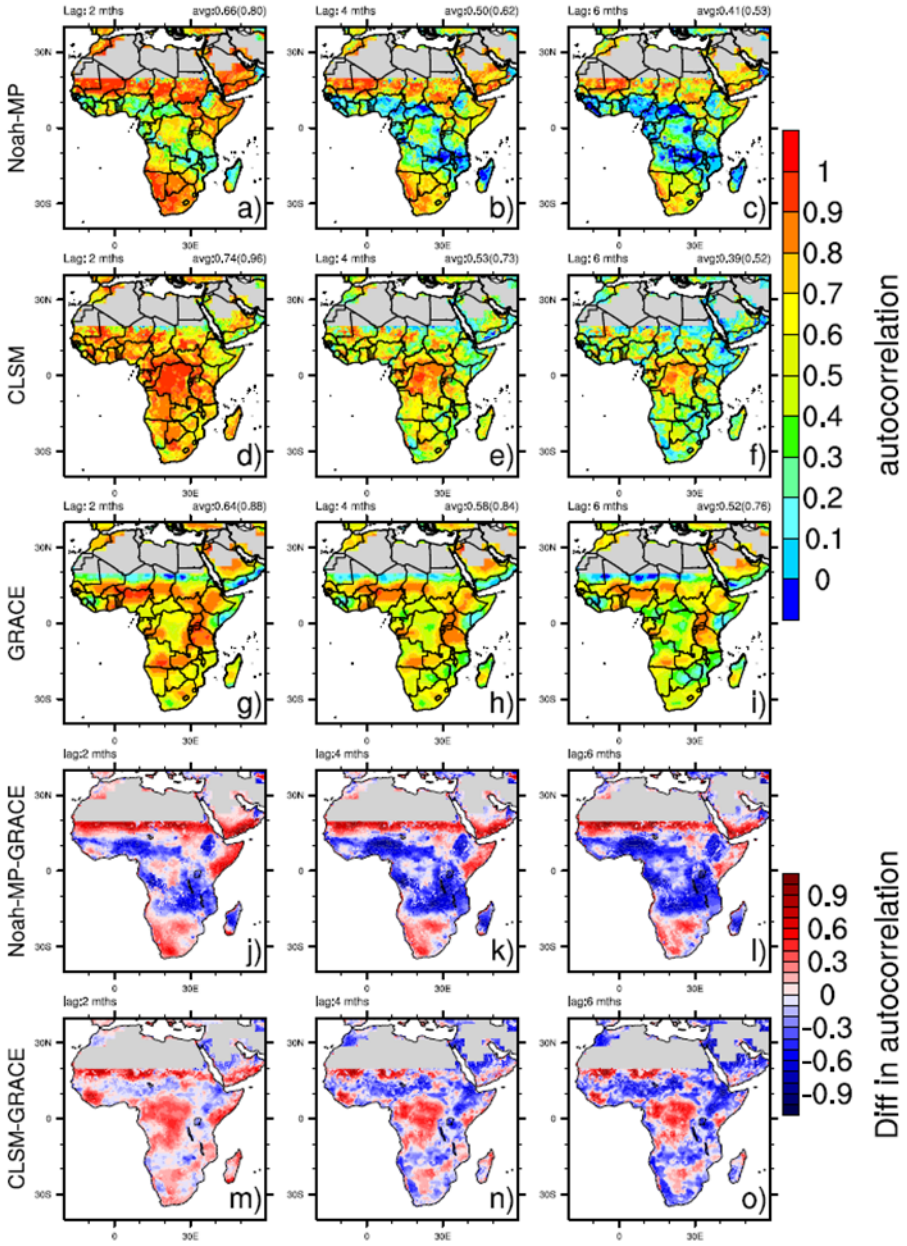

Fig.7 Autocorrelation of Noah-MP and CLSM reanalysis TWS (top two rows), and GRACE/FO data (third row) at three lags. The fourth and fifth rows show differences in autocorrelations between the reanalysis and GRACE/FO data. Upper right text in the top three rows shows average autocorrelation and fraction of area with autocorrelation>0.37 (in parentheses).

Fig 8 . Can the authors make it clear that top left figure is the IC and everything else is forecasts? At first I thought top raw was initialization, while the bottom was the forecasts.

The layout of this figure has been improved, and percentile maps of Noah-MP forecasts and GRACE data have been added. IC and forecast are now clearly labeled in the top two rows. Please see our response to Review#1 on update of this figure.