Review #1

Summary: the authors offer an evaluation of FEWSNET S2S terrestrial water storage forecasts for Africa. The manuscript focuses on differences between the two land surface models included in the FEWSNET forecast ensemble--CLSM and Noah-MP--and offers commentary on the performance of each. Overall, they conclude that CLSM offers advantages when simulating and forecasting TWS. Results also show how various NMME meteorological S2S forecasts compare, but these results are not emphasized in the discussion. The primary source of evaluation data in the main text is GRACE, while information on precipitation forecasts is contained in supplementary material and is addressed only briefly in the text.

I find the results presented in the manuscript to be interesting, and the explanation of these results is generally quite clear and useful. I did find myself a bit confused at times, when the authors bounced between comparing hindcasts to reanalysis and comparing hindcasts to GRACE observations, and when some of the explanation of geographic patterns seemed to me to be speculative. But these are minor points, and I have only a few questions that I would like to see addressed before the paper is published in final form.

Thanks for your supportive comments. We agree that our initial submission overlooked some important findings related to precipitation forecasts and we have since revised the relevant paragraph in the Summary section to highlight those findings: "*Consistent with the above discussion, TWS forecasts are highly sensitive to inter-annual variability of precipitation forecasts, which differs substantially across NMME models. TWS forecasts driven by precipitation forecasts with larger interannual variability (e.g., GEOSv2) showed lower correlation and higher RMSEs with respect to GRACE/FO observations, whereas those driven by precipitation forecasts with lower interannual variability (e.g., GFDL and CSM5) yielded more accurate TWS forecasts. Performance of TWS forecasts also responds to changing interannual variability of NMME precipitation forecasts with lead time. In most cases, precipitation forecasts exhibit decreasing interannual variability with increasing lead time, likely reflecting reduced forecast skill at long leads when prediction reverts toward climatology (Zhang et al., 2021). This decrease in variability leads to contrasting model behaviors, with CLSM TWS forecasts showing reduced RMSEs at longer lead times, whereas Noah-MP forecasts exhibit increasing RMSEs*".

In addition, we created a new section, section 3.3, to present statistical results relative to the reanalysis to avoid confusion between the two sets of evaluation statistics. Both the

Results and Summary and discussions sections have been extensively revised to improve accuracy and reduce redundancy.

Specific comments:

Line 204: isn't the 1m CLSM "soil depth" a choice that was made by the authors? This implementation of the model might output 1m soil moisture, but the model also has an implicit soil water profile that could be used to extract an estimate of total soil moisture integrated to any depth. Similarly (and maybe more easily) the authors could have used 1m soil moisture from Noah-MP rather than the full 2m column. Why not compare 1m CLSM to 1m Noah-MP, or 2m CLSM to 2m Noah-MP?

The 1 m CLSM root zone depth is prescribed by model developers, not a choice made in this study. The model only has three subsurface states, a 2 cm surface layer, a 1 m root zone and the total profile. Indeed, the profile soil water includes deeper soil moisture. But because it does not explicitly simulate groundwater, which is also included in the total profile soil water, there is no other way to extract deeper soil moisture.

Comparing 1 m CLSM with the 1 m root zone soil moisture from Noah-MP would yield more comparable dynamics. However, as the purpose of this analysis is to assess the relative contributions of soil moisture and groundwater to TWS dynamics, we want to present soil moisture in the entire unsaturated zone.

We clarified this issue in section 2.3 as follows: "*Although CLSM does not explicitly model groundwater, groundwater variation is included in the total profile soil moisture; thus, CLSM groundwater storage is obtained by subtracting water storage in the root zone from that of the total soil profile, following previous studies (e.g., Li et al., 2019b). Compared to Noah-MP, CLSM groundwater contains soil moisture from the 1-m depth to the implicit water table. Despite this diagnostic approximation, CLSM groundwater has been shown to compare well with in situ groundwater in different climates (Xia et al., 2017; Li et al., 2019b)*".

Lines 234-249: In Figure 2, the reanalysis errors look almost identical to the forecast errors for both Noah-MP and CLSM. Yet the authors invoke NMME uncertainties when explaining some aspects of model errors. Given that the patterns and magnitude of error appear to be very similar in reanalysis and in forecasts at all lead times, aren't these errors more about model bias than about forecasts? Even the explanations that invoke interannual climate variability seem like they'd need more evidence in their support, since we'd want to know that errors in interannual meteorological variability are seen in a similar way in both CHIRPS (or MERRA-2) and in the NMME models.

Thanks for this comment. The similarity in RMSEs between the forecasts and re-analysis was discussed in Line 250, but it may have been overlooked. We agree that invoking interannual variability in NMME precipitation here is inappropriate and that uncertainties in model physics contribute substantially to those large RMSEs. However, given that there is also similar RMSE patterns between the two land surface models such as the large RMSEs in southern Zambia and Angola, precipitation errors also likely played a role in those spatial patterns of RMSEs.

We revised the opening paragraph of section 3.2 as: "*RMSEs of the ensemble mean TWS hindcasts of all NMME models, with respect to GRACE/FO data, exhibit distinct spatial patterns (Fig.3). Large RMSEs are observed in the interior western Sahel, a large region across Lake Victoria, Lake Tanganyika, and Lake Volta as well as southern Zambia and Angola, for both models. As the models do not simulate surface water which is detected by GRACE/FO satellites, unresolved surface water dynamics and water management activities may have contributed to errors in lake areas. In addition, uncertainties in precipitation forcing data, for both reanalysis and hindcasts, especially under a changing climate, may further amplify errors in simulated TWS. As discussed earlier, the East African Rift, which includes Lake Victoria, has seen increased precipitation variability (Boergens et al., 2024); similarly, Southern Africa including southern Angola has been experiencing erratic precipitation patterns and more severe meteorological droughts in recent years (Trisos et al., 2022; Correia et al., 2025). However, considering that the reanalysis exhibits similar spatial patterns and magnitudes of RMSEs as the hindcasts (Figs.3a,e), deficiencies in model physics are likely the dominant contributor to RMSEs in TWS hindcasts*".

Line 285: If these results compare model forecasts to their own reanalysis, can we really say that degradation of Noah-MP forecasts is due to an "inability" to simulate long-term TWS variability? Couldn't we just as easily say that the persistence of CLSM forecasts is due to that model's "inability" to simulate rapid runoff and drainage? Without an independent evaluation dataset (for this specific result) it's not possible to know which model's behavior is better. That said, the subsequent results that *do* offer comparison with GRACE make a more convincing case. I would recommend that the authors avoid making statements about the quality of model performance when using the retrospective simulations as the truth. (In fact, they might consider moving these statements out of this section, as I admit that I was confused on my first reading about which statements had an observational basis and which were about simulation comparisons.)

Your points are well taken. We moved the statistical results relative to reanalysis to a new section, section 3.3. We also eliminated words like “inability” when discussing results using the reanalysis as reference. We revised this section extensively to simply describe the results without speculation. We also added overestimation of surface runoff as an additional factor impacting groundwater dynamics in the Summary and discussions section.

Section 3.4: Why aren't any GRACE comparisons offered in this section? It seems odd to show the forecast without any evaluation.

Noah-MP and GRACE based percentile maps have been added in Fig.9 (previously Fig.8, see below image) and corresponding discussions have been included in this section.

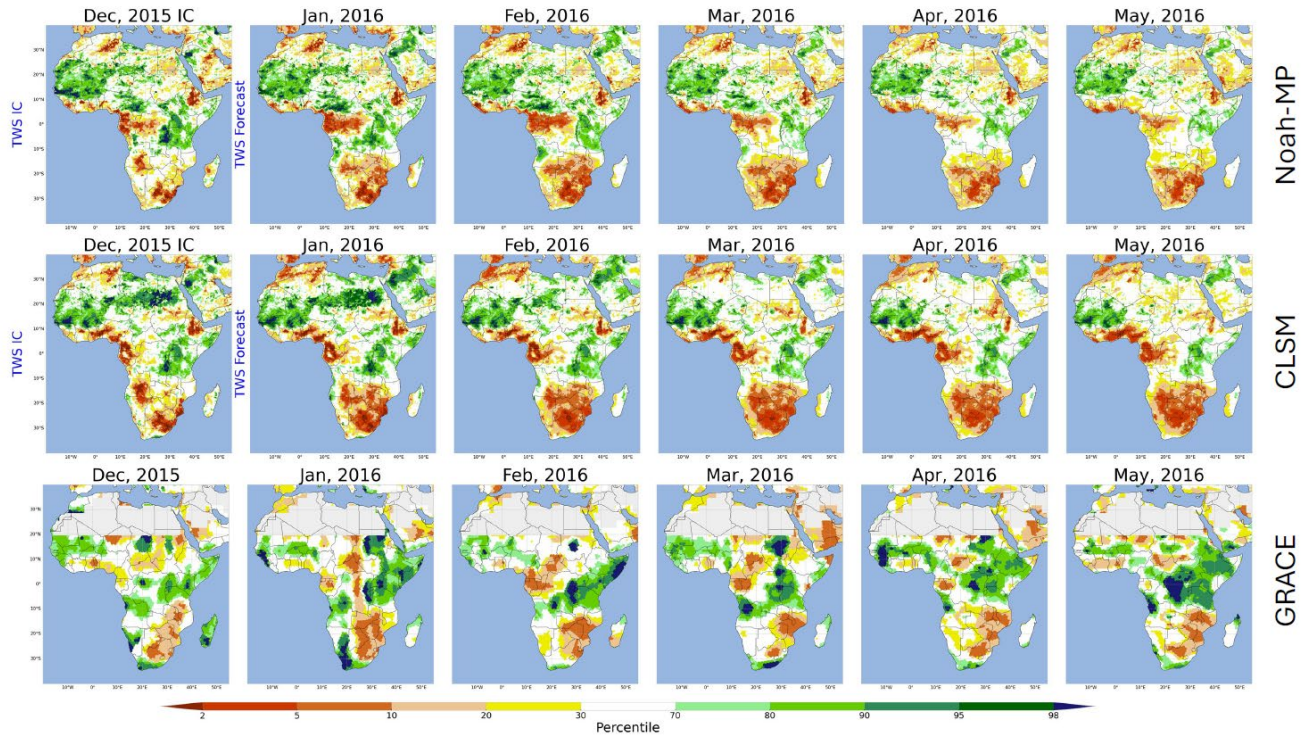

Fig.9 TWS percentile maps derived from Noah-MP and CLSM mean TWS forecasts (top two rows) of all NMME models, initialized in December 2015, and corresponding maps for GRACE/FO data (bottom row).