Generally, the contribution treats a well known topic: the shortcoming of hydrological models in representing long term trends and decadal changes of TWS. This is a relevant topic also in the scope of better understanding the occurrence of extreme events. This contribution goes beyond the investigations of Scanlon et al. (2018) (https://www.pnas.org/doi/10.1073/pnas.1704665115), in particular by accomplishing detailed regional analyses and by looking not only at trends but also at decadal changes. Unfortunately, the study is limited to the time period 2003 to 2016 because of the availability of WGHM data. Due to this 6 years of GRACE/GRACE-FO data cannot be considered, which is quite a significant part of the 20 year time span. With respect to the title and the focus of the paper, I wonder whether it would not make sense to look at some other hydrological models (e.g. the GLDAS models also evaluated by Scanlon et al. (2018)) in order to exploit the whole GRACE time span and to give more value to the message that should be conveyed. However, please explain why you decided to focus your study on WGHM and ISBA.

We thank the Reviewer 3 for his/her comments, that helped considerably improve the manuscript. We focused on global hydrological models rather than land surface models, allowing a more detailed representation of hydrological processes across continental areas. In particular, land surface models of the Global Land Data Assimilation System (GLDAS) such as VIC or NOAH , only take into account vertical fluxes and do not explicitly represent surface water storage and aquifers. While land surface models, such as NOAH, present the undeniable advantage of availability in near real time, allowing a longer period overlap with GRACE and GRACE-FO missions, their accuracy is lesser than those of global hydrological models such as ISBA-CTRIP or WGHM. We added the justification of our choice in the introduction (L76-82). We also provide in supplementary material the comparison of GRACE-based and NOAH-based TWS, which confirms the lower accuracy (larger residuals and lower R^2 over most continental areas) of NOAH in comparison to WGHM or ISBA-CTRIP.

Overall, the manuscript is well written and easy to follow. In particular the part on the regional analyses is extremely interesting and provides new insights into possible reasons for shortcomings of the models. In some parts information should be more concise as commented below.

Changes to the manuscript have been made as follows.

Abstract:
- l14: changes with respect to what? **“with respect to the temporal average” added**
- l19: well correlated (please be more concise), how can differences in TWS be correlated with precipitation? **Rephrased as “consistent with precipitation” (i.e. a drop (rise) in precipitation is consistent with a drop (rise) in TWS)**

- l21: you should mention that this issue is well known and has already be investigated a lot. **This is discussed in the introduction at L54-69.**

Introduction:

- l. 54 onwards: I do not understand why this part about the water mass budget is relevant for this paper. **The paragraph has been removed from the introduction.**

2 Methods:

-l60: why do you truncate at degree 60? There is still some signal contained in the higher degree coefficients. **The signal contained in the higher degree coefficients is overpowered by noise, requiring specific mitigation only available in dedicated filters.**

-l157: do you average the three products? Please be more specific how you "estimate" precipitation. **We use two distinct precipitation products: GPCC and IMERG. GPCC is based on rain gauges measurements. IMERG is based on TRMM and GPM satellite measurements. We rephrased for clarity. We apply the same processing to all time series, as detailed in section 2.5.**

-l178: please define CL at its first occurrence. **Done (CL stands for Confidence Level).**

3 Results / 4 Discussion

- Fig. 1: please consider a different colorbar for Fig 1b. In the other subplots red is larger than yellow/white. It is confusing for interpretation. **The colorbar has been changed for Fig. 1b.**

- l 207: is it possible that the model variability is damped too strongly by the filter that you applied? **We use the same filter (radius at 250 km) on all fields to be able to compare them.**

- Fig. 3: very interesting figure! How should sub-annual to decadal signals in the Sahara region be interpreted? **As no significant hydrological signal is expected in the Sahara, this may be interpreted as the noise signature of the TWS residuals, which contains a high and a low frequency component. A sentence has been added at L262-265.**

- l273: what kind of anthropogenic influences and which kind of climate variability? Please be more specific and cite relevant studies. **These two categories include a variety of processes that are discussed in section 4. Among anthropogenic influences, irrigation has the most impact on TWS changes. Among climate influences, precipitation (droughts/excess rainfall) has the most impact on TWS changes. These examples were added at L272-275. Relevant studies are abundantly cited in the discussion section.**

-l545: could the negative trend in the Black Sea Catchment be related to uncertainties in the water mass correction for lakes? **The lake correction is only added to WGHM and ISBA, not to GRACE. It cannot be responsible for the decreasing trend observed in GRACE-based TWS.**

5 Conclusion

-l587/589 parameterisation → parameterization. **Corrected.**

- l529: joint calibration against discharge and TWSA has been applied e.g. by Werth et al. 2009. **The reference has been added to the conclusion.**

- general: you could also indicate the potential benefit from GRACE data assimilation

We added this information in the conclusion (L596-602).