Response to Anonymous Referee #1

Authors: The authors warmly thank the reviewer for their careful review of the paper and positive comments on the proposed study. These comments are all valuable and helpful for revising and improving our paper, and we have studied them carefully. We respond to the reviewer's comments below and we detail how we plan on improving the paper.

General comment

The paper presents a method for a non-static temporal and spatial validation of the downscaled GRACE (Gravity Recovery and Climate Experiment) data at a resolution deemed appropriate for assessing groundwater storage and irrigation. The authors have combined in situ measurements (e.g groundwater level (GWL) with data-driven (e.g Random Forest and ML) models within an extensive validation framework. Their motivation was driven by the lack of comprehensive dynamic validation strategies for GRACE-derived downscaled products in both time and space to cope with changing hydrological processes through the seasons. In general, their results show that the bias-corrected ML and RF improved the correlation with in situ measurement as compared to the to the LR reference. However, the scaling factor method (SF) degraded the performances and cannot be used at 0.5° resolution as a valid downscaling approach. They also highlighted the flaws of static GRACE dowscaling methods in catchments or areas with hydrological processes varying across the year.

I found the paper very nice to read and has some novelty in both philosophy and methodology dealing with downscaling the GRCAE product which are of interest to the audience of HESS. It is well written and structured in coherent sections with appropriate content.

- By start reading the paper the sentence L 6-8 in the abstract " The point is that the performance of GWS downscaling methods may vary in time due to changes in the dominant hydrological processes through the seasons. To fill the gap, this study investigates the dynamic performance of GWS downscaling by developing a new metric for estimating the downscaling gain (new validation) against non-downscaled GWS" draw my attention. This is one of the main motivations behind this work. However, I was not able to see an explicit consideration of the variability of the dominant hydrological processes in the proposed methodology nor in the results and discussions. If this has been done within the GLEAM model to simulate the soil moisture (SM) storage then this deserves better description and thorough discussion in the results and discussion section.

Authors: We acknowledge that this issue was unclear in the previous version. To address this concern, more detailed and specific discussions will be inserted in section 4.3 of the revised manuscript. In particular, we better describe the dominant hydrological processes throughout the year and how they are (or are not) represented by the model. We also modify the figure 8, by adding two new complementary graph (8b and 8c in the revised).

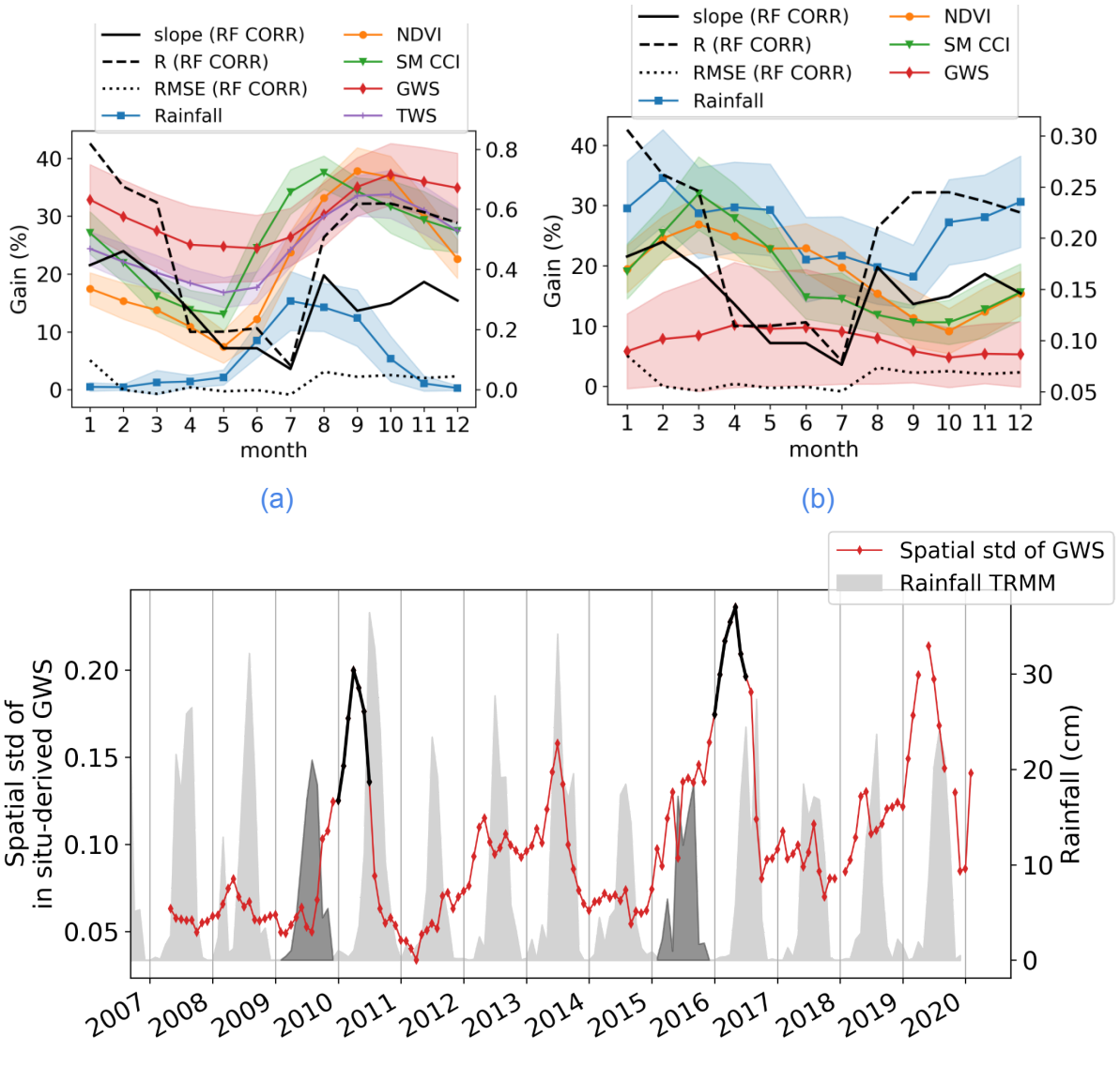
New paragraphs (section 4.3 of the revised):

- "*The periodicity in downscaling performances is due to the capacity or incapacity of the model trained at LR to restitute the spatial variability of some intermittent processes.* In this paper, the tested disaggregation methods are empirical, as the majority of existing methods (Ali et al., 2021; Jyolsna et al., 2021; Ning et al., 2014; Zhang et al., 2021; Zuo et al., 2021). Therefore, we are not able to represent explicitly the underlying hydrological processes that explain (in the downscaling procedure) the spatial variability of GWS at a given time. However, the performance of the disaggregation methods essentially relies on their capability to represent implicitly the discharge and recharge of the aquifer at the 0.5° resolution. This is the reason why the temporal variability in the disaggregation performance can be interpreted in terms of taking into account the dominant hydrological processes and their seasonal dynamics."

- "In the Telangana State, the hydrological year can be divided into several periods given their dominant hydrological processes. The month of August marks the start of aquifer recharge by the rainfall that occurs two to three months after the beginning of the monsoon (which lasts generally from August to October, see figure 8a). It is also the beginning of the growing season (which typically lasts from July to November) when the monsoon rainfall stored at the surface and in the aquifer are used for irrigation. The higher spatial gains on slope, R and RMSE during this period shows that the recharge process in space is correctly represented with the precipitation data at 0.5° (having mainly a North-South gradient). The period between January and March, during the dry season, is marked by the heavy pumping and use of surface water for crop irrigation. This process is relatively well represented by the downscaling model from SM CCI and NDVI data, which provide indirect information on irrigation and crop stage, respectively. During this period, the spatial variability of both predictors (illustrated by figure 8b that represents the interannual average of the monthly spatial variability) is relatively large, accounting for the differences in crop fraction and type that highly depend on surface water availability. By April-May, irrigation stops, and groundwater reaches its lowest level. The downscaling gains obtained at that time of year are relatively low. The model probably fails to restitute the diversity of HR GWS when the water availability and thus water exchanges are very scarce and hardly inferable from the chosen predictors. At the beginning of the monsoon in June-July, heavy rainfall occurs and fills rivers and reservoirs. However, at this early stage of the monsoon, rainfall has not reached the aquifers and GWS remains low as in April and May. Also, surface water is an important component of the water column at this time of year (up to 24% of the annual fluctuation of TWS, see Sect. 2.2.3.), yet runoff is not directly modeled by any of the variables of the RF model, which could also mislead the model into attributing surface water stocks to groundwater."

- "The use of a spatial gain also highlights the difficulties of state-of-the art "static" downscaling methods (calibrated with constant parameters) to restitute an interannual variability. This is illustrated by figure 8b that represents the interannual average (curve) and variability (envelope) of the monthly spatial variability of GWS. The interannual variability of GWS, which is lowest from August to January and

highest from April to July, is inversely proportional to the downscaling performance. This result indicates that this kind of method is unable to represent the interannual variation of the dominant hydrological processes. Such a difference in interannual variability during the end of the dry season can be explained by the succession of dryer and wetter periods dictated by El Nino and La Nina phenomena (Asoka et al., 2017; Vissa et al., 2019). This involves differences of yearly rainfall cumulation that determine the type of crops according to their water needs. During the driest years in particular, differences in water availability widen the gap between 0.5° regions, explaining higher spatial variabilities of GWS. This is illustrated in figure 8c by the abnormally high GWS spatial variability following the dry monsoons of 2009 and 2015."



New figure 8 (in the revised manuscript): Monthly medians of spatial gains (black curves) on slope, R and RMSE for downscaled GWS with the RF CORR model (left axis) with, on the right axis :

(a) average ± standard deviation (std) of low-resolution rainfall, NDVI, SM from the CCI dataset, in situ-derived GWS and TWS scaled between 0 and 1, and
(b) interannual average ± std of the monthly spatial variability (std) divided by the grid maximum of rainfall, NDVI, SM from the CCI dataset and in situ-derived GWS, scaled between 0 and 1.
(c) Time series of the monthly spatial variability (std) of in situ-derived GWS (red curve) with monthly rainfall cumulation (grey). The abnormally dry monsoons of 2009 and 2015 and the high GWS spatial variability during the following dry season are highlighted in black.

Detailed comments

- It may be better to add the native resolution of the SM CCI product to the text (L. 145). Sure, the information exists in the Table 2.

Authors: This information (0.25°) will be added to the main text of the revised version (section 2.2.2).

- Why there is a need to check whether the downscaled product fits to the validation data better than the LR (original GRACE) product?

Authors: Thanks to the reviewer for highlighting this point. Comparing the fit to the validation data of both downscaled GWS and LR GWS allows to assess whether the downscaling process is necessary at all. Indeed, there is no need to use a downscaled product over the LR product if it gives poorer performances at the fine resolution when compared to validation data. What we regret in state-of-the-art validation methods is that performance metrics between downscaled and validation data are qualitatively considered satisfying or unsatisfying. Using a “reference hypothesis” (here “non-downscaled” case) allows to have a reference and to quantitatively judge whether the downscaled GWS is better or worse in terms of accuracy at the targeted (fine) resolution. This will be better explained in the manuscript in Section 3.1.1:

“As highlighted in the introduction, a lack in the majority of publications on GRACE downscaling is the comparison of the downscaled GWS with a null hypothesis. In particular, current evaluation methods check whether metrics fall within an acceptable range that is qualitatively defined. Using a “reference hypothesis” (here the “non-downscaled” case) allows to quantitatively judge whether the downscaled GWS is better or worse in terms of accuracy at the targeted (fine) resolution, and to evaluate if the downscaling process is efficient.”

- In Fig. 2. How the uncertainty envelope was calculated? Can you add this to the text in the appropriate section?

Authors: The uncertainty envelope is the average of the mascon uncertainty resampled at 0.5° that is provided with GRACE data. This information will be added in the legend of Figure 2a in the revised manuscript.

- In L. 276 you reported that " …revealing that the RF suffers from overfitting". Firstly, can you add the R2 value for the test set in the RF? Secondly, do you think the data quality is responsible for the overfitting of the RF during the test phase?

Authors: The R2 for test and train sets for LM and RF models are :
RF : R2 test = 0.93 ; R2 train = 0.98
LM : R2 test = 0.91 ; R2 train = 0.89
We agree with the reviewer that overfitting can be due to data quality, as the model learns data noise, but overfitting also partly comes from the small amount of data available. In fact, the model is trained on an ensemble of 139 points, which is relatively small compared to the complexity of the RF model, resulting in poor generalization.
The R2 values obtained for train and test sets will be added to the revised manuscript in section 4.1. and in table 4:

"*The RF model has a better R2 than the ML model (0.97 against 0.90), yet the RMSE on the test set is way larger than on the train set (4.6 cm against 1.9 cm)*. This reveals that the RF model suffers from overfitting due to data quality and the small amount of data (139 points) used to train the model, resulting in poor generalization."

L. 280 "…already revealing the uncertainty induced by the deconvolution with GLEAM RZSM". I don't understand how the lower performance as compared to in situ is attributed to the uncertainty? This needs to be clarified. In addition, I think that there is a need for better developing the uncertainty issue in this paragraph. This deserves better discussion here although a section on other uncertainty sources in validation already exists in the discussion.
Authors: We thank the reviewer for raising this issue. What we meant is that we cannot expect a high performance when comparing satellite data (or modeled from satellite data) to in situ data, because of (i) the inherent uncertainties of the data, (ii) the interpolation of in situ data and more generally (iii) the diversity of data sources. All those uncertainty sources also apply to the TWS predicted by models at both low and high resolutions. For clarity, the sentence at Line 280 will replaced in the revised by:

"However, the performance is lower when compared to in situ data. As an example, the R2 between in situ-derived TWS (sum of GWS-OW and RZSM GLEAM) aggregated at LR and GRACE TWS is 0.80. This shows that only limited agreement can be expected between satellite data (or modeled from satellite data) to in situ data, because of (i) the inherent uncertainties of the data, (ii) the interpolation of in situ data and more generally (iii) the diversity of data sources. All those uncertainty sources also apply to the TWS predicted by models at both low and high resolutions."

Bibliography

Ali, S., Liu, D., Fu, Q., Cheema, M.J.M., Pham, Q.B., Rahaman, M.M., Dang, T.D., Anh, D.T., 2021. Improving the Resolution of GRACE Data for Spatio-Temporal Groundwater Storage Assessment. Remote Sens. 13, 3513. https://doi.org/10.3390/rs13173513

Asoka, A., Gleeson, T., Wada, Y., Mishra, V., 2017. Relative contribution of monsoon precipitation and pumping to changes in groundwater storage in India. Nat. Geosci. 10, 109–117. https://doi.org/10.1038/ngeo2869

Jyolsna, P.J., Kambhammettu, B.V.N.P., Gorugantula, S., 2021. Application of random forest

and multi-linear regression methods in downscaling GRACE derived groundwater storage changes. Hydrol. Sci. J. 66, 874–887. https://doi.org/10.1080/02626667.2021.1896719
Ning, S., Ishidaira, H., Wang, J., 2014. Statistical Downscaling of Grace-Derived Terrestrial Water Storage Using SaTELlite and Gldas Products. J. Jpn. Soc. Civ. Eng. Ser. B1 Hydraul. Eng. 70, I_133-I_138. https://doi.org/10.2208/jscejhe.70.I_133
Vissa, N.K., Anandh, P.C., Behera, M.M., Mishra, S., 2019. ENSO-induced groundwater changes in India derived from GRACE and GLDAS. J. Earth Syst. Sci. 128, 115. https://doi.org/10.1007/s12040-019-1148-z
Zhang, J., Liu, K., Wang, M., 2021. Downscaling Groundwater Storage Data in China to a 1-km Resolution Using Machine Learning Methods. Remote Sens. 13, 523. https://doi.org/10.3390/rs13030523
Zuo, J., Xu, J., Chen, Y., Li, W., 2021. Downscaling simulation of groundwater storage in the Tarim River basin in northwest China based on GRACE data. Phys. Chem. Earth Parts ABC 123, 103042. https://doi.org/10.1016/j.pce.2021.103042