**Response to Review Comments**

Manuscript Number: HYDROL43309

**Response of terrestrial water storage and its change to climate change in the endorheic Tibetan Plateau**

Dear Editors and Reviewers:

We really appreciate the editors and reviewers for providing invaluable comments and suggestions which have helped us improve the quality of the manuscript substantially. We provide point-to-point responses (in blue colored texts) to those comments and hope that we have addressed all the concerns in the response and revised manuscript.

We are looking forward to hearing back from you soon.

Xingong Li

(on behalf of all the authors)

## Reviewer #1:

Terrestrial water storage (TWS) are good indicators of the response of hydrological system to climate change in TP. This study applied a random forest method to extend the time series of TWSA and TWSC and based on this datasets, trend and contribution analysis were made. Since this is a revised version (although for me, it is a completely new one), I have read through all comments and responses from experts and authors. I fully agree with the comments of Reviewer 1 and also found that comments from other two experts are very positive. Therefore, I reread the manuscript carefully and made additional comments. I leave the decision to the editor.

Thank you very much for reading our manuscript and providing valuable comments and suggestions, which have helped us improve the manuscript substantially.

**Comments of mine:**  
1) In general, after reading this manuscript, I didn't find novelty, which is in agreement with Reviewer 1. But I agree with the authors that there are some contributions such as the inclusion of lake effect on TWS change in TP. Although the methods are old, some new data sets are used and the results are useful.

Thanks for your comment. We agree with you and the reviewer 1 in the first round that this research doesn’t have very significant breakthroughs and method novelty. And we are also pleased that you do agree with us regarding our contributions in terms of the new data sets used and results. For the convenience of the reviewer and editors, we re-summarized these incremental contributions and new findings as follows:

1. We compared, for the first time in the region, the contribution rates of different TWS components estimated from the random forest (RF) model with those from the traditional mass balance approach. We evaluated the RF method on estimating both TWSA and TWSC and found that there are some performance differences among them, while most of the existing studies only focused on TWSA. The reconstructed long-term (1989-2019) TWSA time series data, to the best of our knowledge, could be the most comprehensive (long time and 8 sub-regions) TWS data in the region.

2. We analyzed the lake effect on TWS change, for the first time in the region, using a newly developed long-term lake volume data (all the lakes greater than 1 km2 in 1989-2019). We found that, in general, the effect of lake water storage change (LWSC) on TWSC has increased from P1 to P2 in the IB, but with significant spatial and temporal heterogeneity.

3. We analyzed the contribution rates of different storage components to TWSC under different hydrological conditions (i.e. storage surplus vs. deficit) within sub-regions and revealed the spatial and temporal variations in hydrological cycle in the region. Those findings include the difference between eastern and western Inner Basin, the sign change of TWSA trend from pre-mutation period to post-mutation period, and the difference between TWSA and TWSC.

2) The feasibility and interpretability of the RF method in estimating TWSA and TWSC were not fully explained, in particular the interpretability, which the authors regard this as a merit distinguished from other black box models.

We are sorry that the feasibility and interpretability of the RF were not fully explained in the manuscript, and will explain it here and in revised manuscript.

Similar to other related studies (*Chakraborty et al., 2021; Zhou et al., 2017*), the feasibility in this study mainly refers to that whether the RF method can be used to estimate and reconstruct regional TWSA and TWSC time series. Specifically, whether the method can reproduce the magnitude, inter- and intra-annual variation, regional differences, temporal trend of TWSA and other characteristics. Its feasibility was validated in our study by comparing model estimates with available GRACE TWSA and TWSC observations, which indicated that the calibrated RF models with optimal forcing variables and parameters can provide promising estimates. Thus, the RF method is feasible in reconstructing TWSA and TWSC time series in the region. We also found that the RF models have different performance in different regions and discussed possible reasons for this difference (see also response to Comments 12). In revised manuscript, we explained this feasibility more clearly (Section 5.2, Page 52-53, Line 841-871).

The interpretability of the RF method, which is the unique merit distinguishing from other machine learning methods (e.g., artificial neural network, and support vector machine), mainly refers to its ability to quantify the importance (i.e., contribution rate in our study) of different features (e.g., different forcing variables in RF regression) in machine learning in general (*Ribeiro et al., 2016; Gilpin et al., 2018*) and its applications in hydrology (*Chakraborty et al., 2021; Shortridge et al., 2016; Xu et al., 2021*) and related disciplines (*Vystavna et al., 2021; Evans and Cushman, 2009*). In this study, to further understand and verify the interpretability, we compared the contribution rates from the RF method with those estimated by the conventional mass balance approach, which have clear physical meanings (see also response to Comments 5), in four selected regions (Section 5.2, Table 7). The comparison indicates that the orders of the magnitudes of the component change contribution rates to TWSC estimated by the two methods are basically the same (i.e., similar dominate components), while their magnitudes are somewhat different.

Algorithmically, feature importance in RF is calculated based on the modeling error difference for out of bag samples that with and without noise for all tress (*Breiman, 2001*), and the variable with a greater modeling error difference has a higher importance. As the RF trees are built based on random selection of variables and samples, the method has a non-linear model structure. On the contrary, the mass balance approach acts like a general linear model where each storage component has the same weight of 1 (i.e., TWSC=SWC+SWEC+LWSC+GPIC). So, the RF method could provide a more reliable and comprehensive estimation on the contribution of each variable. This is the main merit of the RF method compared to the mass balance approach, especially for long-term time series data (see also response to Comments 5). However, as a data driven method, the RF method requires more data to train the model, and its result may not be so reliable with a small amount of training data (e.g., a few days, months or years). And the method cannot reveal the negative or positive effects of different storage components as it lacks physical meaning.

In summary, in the revised manuscript, we added more details on the interpretability, including the definition of interpretability (Page 53, Line 872-880), its algorithmic basis (Page 54, Line 881-887), and the merits and limits of the RF method (Page 54, Line 889-897).

3) If possible, the authors can add reference to the division of TP into eight sub-regions. If the authors did the classification themselves, more details are needed.

Thanks for the suggestion. We divided the endorheic TP into the sub-regions by ourselves. We have added more details about the division in the revised manuscript (Section 2.2, Page 11-12, Line 208-224), which are also summarized below for your convenience.

Firstly, the whole endorheic region of the Tibetan Plateau is divided into two large endorheic basins, Inner Basin (IB) and Qaidam Basin (QB). The boundary of the two basins is determined by their hydrological dis-connectivity, which is similar as *Zhang et al. (2013)*.

The two basins were further divided into eight sub-regions based on aridity index, catchment distribution, permafrost type, and climate type. Specifically, we first calculated the aridity index for each of the 477 catchments in the two basins (see Fig. 2 in the manuscript). Then, we divide IB into northern and southern parts based on the aridity index threshold of 0.1 and permafrost types (i.e., continuous permafrost in northern IB, and island permafrost and seasonally frozen ground in southern IB) (see Fig. 1b in the manuscript) of their constituent catchments. In the southern IB, climate types (see Fig. 1c in the manuscript) and an aridity index threshold of 0.2 (division of arid and semi-arid) were further used to divide it into the western and eastern sub-regions (i.e., S06 and S04). In the northern IB, climate type gradually transforms from arid-desert-cold to polar-tundra (see Fig. 1c in the manuscript), and aridity index gradually increases from 0.03 to 0.43 (see Fig. 2 in the manuscript). Aridity index threshold of 0.1 and 0.2 were used to further divide the northern IB into western sub-region (S02) and central-eastern region which was further divided into sub-region S03 and S05 based on the dominate climate types of their constituent catchments (see Fig. 1c in the manuscript).

In the QB, we divided it as two sub-regions (S07 and S08) using aridity index threshold of 0.1. Because of the coarse spatial resolution of the data used, some small catchments with lowest aridity index (< 0.03, hyper arid) in the hinterland and with higher aridity index (> 0.2, semi-arid) in the northeastern QB were aggregated into their surrounding sub-regions, respectively.

4) Why did the authors use CMFD? Why not CGDPA?

We utilized the CMFD data based on the following reasons:

1. The CMFD data has been validated and widely used (e.g., *Ding et al., 2018; Guo et al., 2019; He at al., 2020; Ma et al., 2019; Wang et al., 2020; Yang et al., 2010*), especially in the Tibetan Plateau which has very limited ground observations. *He at al., (2020)* compared station observations of China Meteorological Administration (CMA) with several satellite remote sensing based data, and found that CMFD has a better accuracy. Other validations with several independent observations in the Tibetan Plateau (*Guo et al., 2019;* *He at al., 2020; Ma et al., 2016; Wang et al., 2020*) also show a higher accuracy of the CMFD data. We also compared the CMFD data with only three CMA station observations available in the Inner Basin (IB, also known as the Qiangtang Plateau). The results also showed that it has a higher accuracy (see the Figure below). Additionally, we are more familiar with the CMFD data than the CGDPA (China Gauge-Based Daily Precipitation Analysis) data as we also have been used the CMFD data in our other studies.

2. The CGDPA data was produced based on ground station observations using an optimal interpolation method (*Shen and Xiong, 2015*). Its accuracy is largely dependent on the number of stations, and its uncertainty is very high with sparse observations. Unfortunately, stations are very sparse in the Tibetan Plateau (*Shen and Xiong, 2015*) and there are only three stations in the Inner Basin with a total area of 102.36 ×104 km2. The CMFD precipitation data is obtained by fusing remote sensing data (TRMM), reanalysis data (GLDAS) and in-situ data (*He at al., 2020; Yang et al., 2010*). As such, the CFMD data tends to capture the realistic distributions of precipitation at high spatiotemporal resolution than the interpolation based CGDPA data in the data limited Tibetan Plateau.

3. The CMFD data has a higher spatial resolution (0.1°) than the CGDPA data (0.25° or 0.5°). We selected the CMFD data as we have small sub-regions in this study (e.g., S01, S06 and S07).

4. Additionally, we checked the link provided by the data producer of CGDPA (<http://cdc.nmic.cn/sksj.do?method=ssrjscp>, *Shen and Xiong, 2015*) and in several other literature (e.g., <http://data.cma.cn/>), but we just could not find the data.



**Figure.** Comparison of annual precipitation between CMFD data with CMA station observations at Shenzha (a), Gaize (b), and Bange (c) during the period of 1985-2013.

5) Line 390-391: This is a very simple index for contribution analysis. Any reference? Discuss its merits and dis-merits in discussion.

Thanks for the suggestion. We have added several references (Section 3.2, Page 23, Line 410-411) and provided more discussions on the index in the revised manuscript (Section 5.2, Page 54-55, Line 882-892), which are also summarized below for your convenience.

The main merit of the index is its simplicity and a clear physical meaning as it is derived from the universal mass balance equation. The index (i.e., the relative contribution rate) reflects how much the proportional change in a certain water storage component leads to a proportional change in total water storage (*Wang et al., 2021; Zhang et al., 2017a; Zhang et al., 2017b*). The second merit is that the index can be used even with small amount of data sets (e.g., a few days, months, or years), which is more reliable than data-driven methods (e.g., random forest) when data amount is limited. The third merit is that the index can reflect the negative or positive effect of each water storage component change on total water storage change, which is also the main limit of the random forest method. Positive values indicate that changes in the storage components result in water storage surplus (i.e., increase of total water storage), and vice versa for negative values.

The main limit is that the index may not be suitable for long-term time series data (e.g., multi-annual) in some cases. In a long time period, the positive change of a storage component in some years could be offset by its negative change in other years, resulting in a weak effect on total water storage change. As such, the dominant component identified by the index could be biased or wrong. For example, during P1 in S04 (Fig. 11f in the manuscript), while lake water storage change (LWSC) has a higher effect on net TWSC than soil water storage change (SWC) for the period, SWC dominates TWSC during storage surplus (P1+) and deficit (P1-) periods. This is the main reason why we also calculated the contribution rates for different hydrological scenarios. Another limit is that the index may smooth out temporal variation because of its averaging approach and therefore, may not guarantee consistent results between short-term and long-term data.

6) From Table 4, we can see that some variables are not independent, such as Temp and ET. Can RF handle such dependences among input variables? How will these dependences play a role?

Unlike the traditional linear regression method, random forest (RF) method is not sensitive to the dependence among different input variables (i.e., the multicollinearity problem) (*Breiman, 2001*). The method considers the interactions between different input variables, and does not require independent input variables (*Cutler et al., 2007*). Algorithmically, RF fits many regression trees to a data set, and the final modeling result is the combination from all the trees. Each tree is constructed using randomly selected variables and bootstrap training samples (*Breiman, 2001*). For each time of the bootstrap sampling, only two-thirds of the training sample is utilized for training the regression tree, and the remaining one-third constitutes the out-of-bag sample is used for validation the tree *(Breiman, 2001; Cutler et al., 2007)*. The built-in validation and random selection of variables enable RF to handle the dependence among input variables (*Zhou et al., 2017; Rodriguez-Galiano et al., 2014*).

As the terms of “dependent variable” and “independent variable” could be confusing, we have changed them to “response variable” and “forcing variable”, respectively, in the revised manuscript (Tables 3 and 4).

7) If possible, shorten Section 4.3.1-4.3.3 or 4.3.4

Thanks for the suggestion. We have condensed and shortened these sections, and these changes can be found in the following sections in revised manuscript:

Section 4.3.2 Page 36-37, Line 599-600, Line 604-607; Page 39, Line 621-623, Line 626-627.

Section 4.3.3 Page 39, Line 630-631, Line 636-638; Page 40, Line 647-649.

Section 4.3.4 Page 42, Line 669-671.

8) Provide more details about different scenarios (storage surplus scenarios and storage deficit scenarios)

Thanks for the suggestion. We have added more details about different scenarios in the revised manuscript (Section 4.4, Page 44, Line 709-713), which are also summarized below for your convenience:

Based on Pettitt mutation test (Table 5), the study period of 1989-2019 (i.e., PA) is divided into the pre-mutation (i.e., P1) and post-mutation periods (P2). So, scenario “P1”, “P2” and “PA” (Section 4.4, and Fig. 11) mean component contributions during the three time periods. In addition, years with negative TWSC (i.e., storage deficit and referred to as P1-, P2-, and PA-) and years with positive TWSC (i.e., storage surplus) (referred to as P1+, P2+, and PA+) within the three periods were also selected as scenarios. For example, in the IB, P1 and P2 are the hydrological years from 1990 to 2004 and from 2005 to 2019 respectively, and P1+ are the hydrological years of 1993, 1996-1998 and 2000-2001, while P1- includes the hydrological years of 1990-1992, 1994-1995, 1999, and 2002-2004 (see Fig. 11 in the manuscript). The mean TWSC (Table 6 in the manuscript), SWC, SWC, LWSC, and GPIC under these scenarios (i.e., these consecutive or discontinuous years) were calculated and used for component contribution calculation.

9) There are too many contents in Discussion.

Thanks for the suggestion. We have condensed and shortened the Discussion, and part of the original contents in Discussion (Page 59-60, Line 980-1000) have been moved into Supplementary Materials. These changes can be found in the following sections in revised manuscript:

Section 5.1 Page 49-50, Line 772-775, Line 781-782, Line 788-795.

Section 5.2 Page 55-56, Line 892-912.

Section 5.3 Page 57, Line 924-928; Page 59, Line 983-986.

Section 5.4 Page 60, Line 989-1009; Page 62, Line 1036-1050.

10) Put the uncertainty results into Results. Discussion about uncertainty can still be in Discussions.

Thanks for the suggestion. As some of uncertainty results have been presented in the Results (mainly in the Fig. 4g-i, Fig. 8 and Fig. 9g-i in the manuscript), we moved them into the supplementary materials to make the manuscript more concise (Table S3).

11) Line 743-744 the difference of ERA-5 and GLDAS-Noah can be put in supplementary materials.

Thanks for the suggestion. We have moved the difference of ERA-5 and GLDAS-Noah into Supplementary Materials. And other related details have also been moved over there (e.g., Page 16, Line 275-280).

12) Line 793: the authors mention the difference performance of RF models on data with trends and fluctuation. More interpretability about this point is welcome.

Thanks for the suggestion. Based on the different performance of the RF models on estimating TWSA and TWSC in different regions, combined with the temporal variation characteristics (e.g., significant trend or fluctuation) of TWSA in these regions, we speculate that one possible reason is that the RF method might be better at capturing fluctuations. But it must be pointed out that this is just a speculation of us without any test.

On the other hand, the selection of input forcing variable tends to have a higher effect on model performance. Based on the relationship between TWSA and TWSC (*TWSAt* = *TWSAt-1* + *TWSCt*), TWSA at current time not only includes the information at the current time (TWSCt), but also contains the information at the previous time (TWSAt-1). In other words, it reflects the memory effect of a hydrological system to some extent. But most of the forcing variables do not reflect the memory effect, except for soil water storage and snow water equivalent with slim possibility. (Section 5.2, Page 52-53, Line 835-853).

13) Check the references carefully. There are still some mistakes. One such example is Line 1179, 1185, 1195. The list of authors is not complete.

Thanks, we have carefully checked and corrected the references.

### References:

Breiman, L., 2001. Random forests. Machine Learning, 45(1): 5-32.

Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. Expert Systems with Applications, 170, 114498.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783-2792.

Ding, J., Zhang, Y., Guo, Y., & Ma, N. (2018). Quantitative comparison of river inflows to a rapidly expanding lake in central Tibetan Plateau. Hydrological Processes, 32(21), 3241-3253.

Evans, J. S., & Cushman, S. A. (2009). Gradient modeling of conifer species using random forests. Landscape ecology, 24(5), 673-683.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.

Guo, Y., Zhang, Y., Ma, N., Xu, J., & Zhang, T. (2019). Long-term changes in evaporation over Siling Co Lake on the Tibetan Plateau and its impact on recent rapid lake expansion. Atmospheric research, 216, 141-150.

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing dataset for land process studies over China. Scientific Data, 7(1), 1-11.

Ma, N., Szilagyi, J., Niu, G. Y., Zhang, Y., Zhang, T., Wang, B., & Wu, Y. (2016). Evaporation variability of Nam Co Lake in the Tibetan Plateau and its role in recent rapid lake expansion. Journal of Hydrology, 537, 27-35.

Ma, N., Szilagyi, J., Zhang, Y., & Liu, W. (2019). Complementary‐relationship‐based modeling of terrestrial evapotranspiration across China during 1982–2012: Validations and spatiotemporal analyses. Journal of Geophysical Research: Atmospheres, 124(8), 4326-4351.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Model-agnostic interpretability of machine learning. arXiv preprint arXiv:1606.05386.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment, 476, 189-206.

Shen, Y., & Xiong, A. (2016). Validation and comparison of a new gauge‐based precipitation analysis over mainland China. International journal of climatology, 36(1), 252-265.

Shortridge, J. E., Guikema, S. D., & Zaitchik, B. F. (2016). Machine learning methods for empirical streamflow simulation: a comparison of model accuracy, interpretability, and uncertainty in seasonal watersheds. Hydrology and Earth System Sciences, 20(7), 2611-2628.

Vystavna, Y., Harjung, A., Monteiro, L. R., Matiatos, I., & Wassenaar, L. I. (2021). Stable isotopes in global lakes integrate catchment and climatic controls on evaporation. Nature Communications, 12(1), 1-7.

Wang, B., Ma, Y., Su, Z., Wang, Y., & Ma, W. (2020). Quantifying the evaporation amounts of 75 high-elevation large dimictic lakes on the Tibetan Plateau. Science advances, 6(26), eaay8558.

Wang, W., Zhang, Y., Geng, X., & Tang, Q. (2021). Impact classification of future land use and climate changes on flow regimes in the Yellow River Source Region, China. Journal of Geophysical Research: Atmospheres, 126(13), e2020JD034064.

Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. Wiley Interdisciplinary Reviews: Water, e1533.

Yang, K., He, J., Tang, W., Qin, J., & Cheng, C. C. (2010). On downward shortwave and longwave radiations over high altitude regions: Observation and modeling in the Tibetan Plateau. Agricultural and Forest Meteorology, 150(1), 38-46.

Zhang, G., Yao, T., Shum, C. K., Yi, S., Yang, K., Xie, H., ... & Yu, J. (2017a). Lake volume and groundwater storage variations in Tibetan Plateau's endorheic basin. Geophysical Research Letters, 44(11), 5550-5560.

Zhang, G., Yao, T., Xie, H., Kang, S., & Lei, Y. (2013). Increased mass over the Tibetan Plateau: From lakes or glaciers?. Geophysical Research Letters, 40(10), 2125-2130.

Zhang, Q., Liu, J., Singh, V. P., Shi, P., & Sun, P. (2017b). Hydrological responses to climatic changes in the Yellow River basin, China: Climatic elasticity and streamflow prediction. Journal of Hydrology, 554, 635-645.

Zhou, J., Shi, X., Du, K., Qiu, X., Li, X., & Mitri, H. S. (2017). Feasibility of random-forest approach for prediction of ground settlements induced by the construction of a shield-driven tunnel. International Journal of Geomechanics, 17(6), 04016129.

## Reviewer #2:

This study examined changes in terrestrial water storage (TWS) over the endorheic basin in the Tibetan Plateau using GRACE satellite data along with other reanalysis data and products before and after mutation. This is a very interesting study. I enjoyed reading the manuscript. Nevertheless, it needs some further improvements:

Thank you very much for your comments and suggestions!

1：Line 30: 1989-2019

Thanks, and we have corrected it. (Page 2, Line 30)

2: Line 40-43: The overall increase of TWSA in the IB (2.20 mm/a, P < 0.05) during PA was mainly attributed to its increase in the north-eastern area during P2, with a gentle variation in P1 (-0.65 mm/a) and even TWSA reduction in western (-0.14 mm/a) and two southern sub-regions (-0.62 mm/a and -0.27 mm/a) over PA. The conclusion is confusing. I cannot clearly know what you what to express, please write it again.

Thanks, we have rewritten it as “Significant annual TWSA increase (2.20 mm/a, P < 0.05) in the IB comes from a higher increase during P2 (3.62 mm/a, P < 0.05) and a gentle variation during P1 (-0.65m/a, P > 0.1). In space, the increase attributes to the higher increase trend (> 4.50 mm/a, P < 0.05) in north-eastern IB during PA and P2, and slight decrease in western and southern IB during PA”. (Page 1-2, Line 43-47)

3: Line 970 to 977: In According to the conclusion, the contribution rates are quite different before and after PA (Line 47 to Line 51). Please explain the reason.

We are sorry that, in this section, we may not clarify that “PA” refers to the entire analysis period (1989-2019), and “P1” and “P2” refer to the pre- and post-mutation periods, respectively. And in Conclusion and Abstract, we only discussed the difference between Inner Basin (IB) and Qaidam Basin (QB) during the entire period (i.e., PA).

The contribution rate difference between IB and QB is mainly caused by the difference in environmental factors (e.g., precipitation, air temperature, evapotranspiration and permafrost type) and in water storage components between them. As shown in our results, there are large differences in the quantity, variation trend, and variation amplitude of these factors between the two regions. And the contribution rates difference also reflects the different hydrological processes in them to some extent. For example, in the IB with widely distributed lakes, lake water storage has a more significant increase trend and a higher effect on regional water storage change. And in the QB, due to the lack of continuous permafrost, the exchange of soil water and groundwater is stronger than that in IB. As such, change in soil water and groundwater have higher contribution rates to regional water storage change in the QB.

For different periods (i.e., P1 and P2), there are also some differences in the contribution rates and dominant components. Similar to the spatial difference between IB and QB and their sub-regions, these temporal differences are also caused by the difference in related hydrological and climatic variables during different periods. In Section 5.3 (Page 56-59, Line 918-982), we further discussed the possible causes of those spatial and temporal differences.

4: In the 5.4 (Line 904 to 960), you referred to the difference between TWSA and TWSC. You analyzed the contribution of different components to TWSC before and after mutation for TWSC. If the difference dominant components exist for TWSA before and after mutation. And what is the different between TWSA and TWSC.

The dominant components for TWSA are also different during the pre- and post-mutation period, as shown in the Figure below. The contribution rates for TWSA are quite different from those for terrestrial water storage (TWS) and TWSC and they do not have a clear physical meaning but just the result of mathematical calculation. The main reason for this is that TWSA is the anomaly of the TWS, which has a different magnitude from the TWS and only reflects the relative variation of the TWS. In contrast, TWSC calculated from TWSA (i.e., TWSAt – TWSAt-1) is equal to what can be calculated from TWS (i.e., TWSt – TWSt-1). So, the contribution rates and dominate components for TWSC clearly and truly reflect which component changes dominate regional water storage change.



**Figure.** Contribution rate of different water storage component anomalies to TWSA during the pre-mutation period (P1) **(a)**, post-mutation period (P2) **(b)**, and whole period (PA, 1989-2019) **(c)** in the Inner Basin (IB), Qaidam Basin (QB) and their sub-regions (S01-S08).

Based on our results and literature (*Deng et al., 2018; Lv et al., 2019; 2020; Wang et al., 2020*), the main differences between TWSA and TWSC are summarized below:

1. The magnitude and sign (i.e., negative or positive) of TWSA and TWSC have different meanings. The higher the TWSA, the greater the regional water storage, although we don’t know what the actual water storage is. The higher the absolute value of TWSC (i.e., a larger positive value or a smaller negative value), the faster the exchange of water storage components and regional hydrological cycle. Positive and negative TWSA indicate that whether the TWS during a certain time periods (e.g., day, month, and year) is higher or lower than its time series mean. Positive or negative TWSC means water storage surplus or deficit in a certain period compared to previous time period.
2. The time scale of the information in TWSA and TWSC is different. TWSA includes not only the information at current time (i.e., *TWSAt* = *TWSAt-1* + *TWSCt* where *TWSCt* = *Prect* – *ETt* – *Rt*), but also the information in previous time. The equation *TWSAt* = *TWSAt-1* + *TWSCt* can be further expended recursively by using TWSAt-2, TWSAt-3, …, and TWSA1 to replace TWSAt-1 which leads to *TWSAt* = *TWSCt* + *TWSCt-1* + *TWSCt-2* + … + *TWSC1*. In other words, TWSA reflects the accumulative effect of TWSC in time, while TWSC itself is only determined by the input (e.g. Prec) and output (e.g., ET and R) at current time (*TWSCt* = *Prect* – *ETt* – *Rt*).
3. As *TWSCt* = *Prect* – *ETt* – *Rt*, the water balance term is TWSC rather than TWSA (*Budyko, 1974; Lv et al., 2019; 2020; Wang et al., 2020*). As such, TWSC tends to be more closely related to other water balance components (Prec, ET, and R) at the current time than TWSA is. On the other hand, TWSA at the current time is more closely related to cumulative water balance term in previous time. If Prec meets the water consumption (e.g., ET in the endorheic basins) and we have a surplus in water balance term, TWSA will increase (*Deng et al., 2018*). Generally, TWSC is more closely related to the temporal variation of Prec while TWSA is more closely related to the magnitude of Prec.
4. The long-term trends of annual TWSA and TWSC also have different meanings and magnitude. TWSA trend is the same to TWS trend, but is different from TWSC trend. A positive TWSA trend indicates increasing regional water storage, and vice versa for a negative trend. A positive trend slope of absolute TWSC (i.e., |TWSC|) means accelerated hydrological cycle. The contribution rates of different water storage components’ change (e.g., SWC, SWEC, LWSC, and GPIC) to TWSC are quite different from those of components’ anomaly (e.g., SWA, SWEA, LWSA, and GIPA) to TWSA. And the latter is also different from the contribution rates of different water storage components (e.g., SW, SWE, and LWS) to TWS.

We have clarified these differences in the revised manuscript (Section 5.4, Page 61-62, Line 1010-1035).

5. By Random Forest, you extend the TWSA dataset by dividing the TWSA dataset from 2003 to 2020 into training sets and test sets at a ratio of 7:3 for the RF model. The optimal model (with its variable combinations) of each region was used to extend the time span of the GRACE (FO) TWSA and TWSC data from 2002 back to 1989. If you can verify the extended data by the local site data.

We totally agree with the reviewer that we should verify the extended data with ground observations. However, with the remote and harsh natural environment in the Tibetan Plateau and especially in the endorheic regions, ground observation data for hydrological and climatic variables are very scarce. For example, there are only three national meteorological stations in the Inner Basin with a total area of 102.36 ×104 km2，and ground observations for soil water, groundwater and other water storage components are even scarcer, which makes the verification impossible.

Because of limited observation data on storage components, to indirectly validate TWSA, groundwater level data from observation wells have often been used to validate the groundwater storages estimated from mass balance approach with the GRACE TWSA data (e.g., *Bibi et al., 2019; Gao et al., 2020; Tang et al., 2017*). These studies are mainly conducted in regions where groundwater observations are available and groundwater plays a significant role in regional water storage change, such as the northern China plain (*Gao et al., 2020*) and northern Pakistan (*Tang et al., 2017*). But in the endorheic Tibetan Plateau, due to data limitation, we cannot distinguish groundwater storages from glacier and permafrost water storages (i.e., GPIA and GPIC in the manuscript) even if there were groundwater level observations for verification.

In summary, the lack of available observational data in the region make it impossible for us to verify the extended data. This is also one of the limitations of this study as we pointed out in Section 5.1 (Page 51, Line 815-816) in the manuscript.

### References:

Bibi, S., Wang, L., Li, X., Zhang, X., & Chen, D. (2019). Response of groundwater storage and recharge in the Qaidam Basin (Tibetan Plateau) to climate variations from 2002 to 2016. Journal of Geophysical Research: Atmospheres, 124(17-18), 9918-9934.

Budyko, M. I. (1974). Climate and life. Int. Geophys. Ser., vol. 18, 508 pp., Academic Press, New York.

Deng, H., Pepin, N. C., Liu, Q., & Chen, Y. (2018). Understanding the spatial differences in terrestrial water storage variations in the Tibetan Plateau from 2002 to 2016. *Climatic Change*, *151*(3), 379-393.

Gao, F., Wang, H., & Liu, C. (2020). Long-term assessment of groundwater resources carrying capacity using GRACE data and Budyko model. Journal of Hydrology, 588, 125042.

Lv, M., Ma, Z., Li, M., & Zheng, Z. (2019). Quantitative analysis of terrestrial water storage changes under the Grain for Green program in the Yellow River basin. Journal of Geophysical Research: Atmospheres, 124(3), 1336-1351.

Lv, M., Ma, Z., & Yuan, N. (2021). Attributing terrestrial water storage variations across China to changes in groundwater and human water use. Journal of Hydrometeorology, 22(1), 3-21.

Tang, Y., Hooshyar, M., Zhu, T., Ringler, C., Sun, A. Y., Long, D., & Wang, D. (2017). Reconstructing annual groundwater storage changes in a large-scale irrigation region using GRACE data and Budyko model. Journal of Hydrology, 551, 397-406.

Wang, J., Chen, X., Hu, Q., & Liu, J. (2020). Responses of terrestrial water storage to climate variation in the Tibetan Plateau. Journal of Hydrology, 584, 124652.