

## Anonymous referee #1

**RC 1.0:** The authors have satisfactorily revised the manuscript, improved readability, and clarified the raised questions. I have a few more minor remarks that should be addressed before final publication.

**Response:** We thank the reviewer for their positive assessment of the revised manuscript and for acknowledging the improvements in readability and clarity. We sincerely appreciate the reviewer’s ongoing participation and are committed to addressing the remaining minor comments to finalize the manuscript.

**RC 1.1:** The findings of the uncertainty assessment (Sec. 4.8) should be shortly reflected in the conclusions as well as in the abstract.

**Response:** We thank the reviewer for this valuable suggestion. In the revised manuscript, the findings of the uncertainty assessment have been included in the conclusion section (P: 43, L:612-614), as follows:

“Furthermore, the uncertainty associated with BNML\_TWSA is assessed for each grid cell in the form of standard error ( $\sigma_\varepsilon$ ). The results showed that the standard error of the BNML\_TWSA exhibits a smaller magnitude in grid cells located in arid regions compared to those in other regions.”

This information is also included in the abstract section (P:1, L:18-19).

**RC 1.2:** Fig. 3: Do you only consider CCs > 0? Otherwise you may find positive differences also by deteriorating from a negative value toward zero.

**Response:** We thank the reviewer for this suggestion. In the revised manuscript, we have excluded the grid cells that have negative CC values for the leading ML model. Specifically, 1,535 grid cells ( 2.65% of the total grid cells considered in this study) have CC values less than zero, and we have excluded these when calculating the difference. Although the improvement in terms of CC value difference (between the leader model and the worst model) is not large for all grid cells globally, more than 14.4% of grid cells show improvements greater than 0.2, while an additional 15.7% of grid cells exhibit improvements between 0.1 and 0.2 (P:15,L:330-335). The updated figure is also shown below.

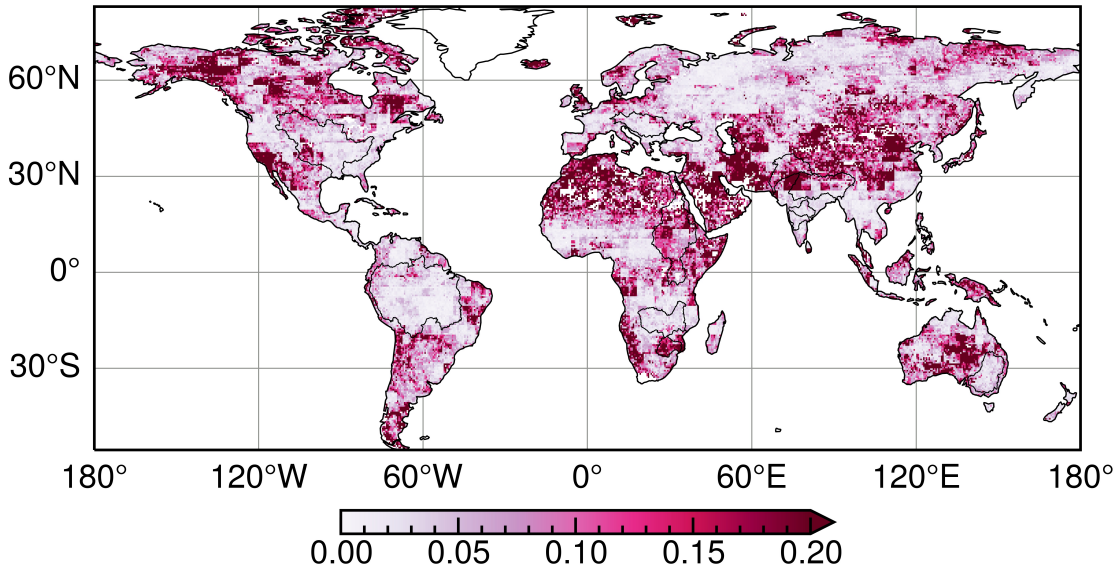


Figure 1: Difference between the correlation coefficient (CC) values obtained from the leader ML model (excluding negative CC values) and the worst performing ML algorithm in each grid cell during the test period.

**RC 1.3:** It seems very strange to me that your KGE values are always negative while you get positive NSE values in the range of 0.5 to 0.8. I suggest to reassess the computation of KGE or to provide some discussion why these negative values would be feasible. Possibly, KGE' can be used as an alternative to KGE. Please provide the reference for KGE.

**Response:** We sincerely thank the reviewer for highlighting the unusual combination of negative KGE values alongside positive NSE values (0.5 to 0.8). We agree that this required further explanation, and we appreciate you bringing it to our attention.

In the previous version of the manuscript, we employed a modified Kling-Gupta Efficiency (KGE') as defined below (Kling et al., 2012; Gupta et al., 2009)

$$KGE = 1 - \sqrt{(CC - 1)^2 + (\beta - 1)^2 + (\alpha - 1)^2}, KGE \in (-\infty, 1] \quad (1)$$

where  $\beta$  and  $\alpha$  represents bias error and variability error respectively.

$$\beta = \text{bias ratio} = \frac{\bar{S}}{\bar{O}} \quad (2)$$

$$\alpha = \text{variability ratio} = \frac{\sigma_S / \bar{S}}{\sigma_O / \bar{O}} \quad (3)$$

In this revised manuscript, we have now calculated the KGE according to the original equation presented by Gupta et al. (2009):

$$KGE = 1 - \sqrt{(CC - 1)^2 + (\frac{\bar{S}}{\bar{O}} - 1)^2 + (\frac{\sigma_S}{\sigma_O} - 1)^2}, KGE \in (-\infty, 1] \quad (4)$$

where  $\sigma_O$  and  $\sigma_S$  represent the standard deviations of observations and simulations, respectively. This correction has resolved the issue of the negative KGE values.

We have also added the appropriate citation for the original KGE formulation (Gupta et al., 2009) to the revised manuscript (P:12-13). Thank you again for your valuable feedback.

## Additional editorial remarks:

**RC 1.4:** L7: I suggest to change grid cell-based to “raster-based”

**Response:** We have updated the Abstract in the revised manuscript to reflect this change (P:1, L:7).

**RC 1.5:** Table 3 & 4: Captions should not be capitalized

**Response:** We are thankful to the reviewer for bringing it our notice. In the revised manuscript, we have rectified these mistakes as follows:

### Caption of Table 3

“Details of basins and streamflow observation locations for six global river basins. Sources: Global Runoff Data Centre (GRDC; <https://portal.grdc.bafg.de>) and Central Water Commission (CWC; <https://indiawris.gov.in>), India”

### Caption of Table 4

“Overview of global precipitation, evapotranspiration and storage change data products utilized for streamflow calculations”

**RC 1.6:** Figure 4: Units are missing for the axes.

**Response:** We have updated this figure in the revised manuscript and the updated figures is also presented below.



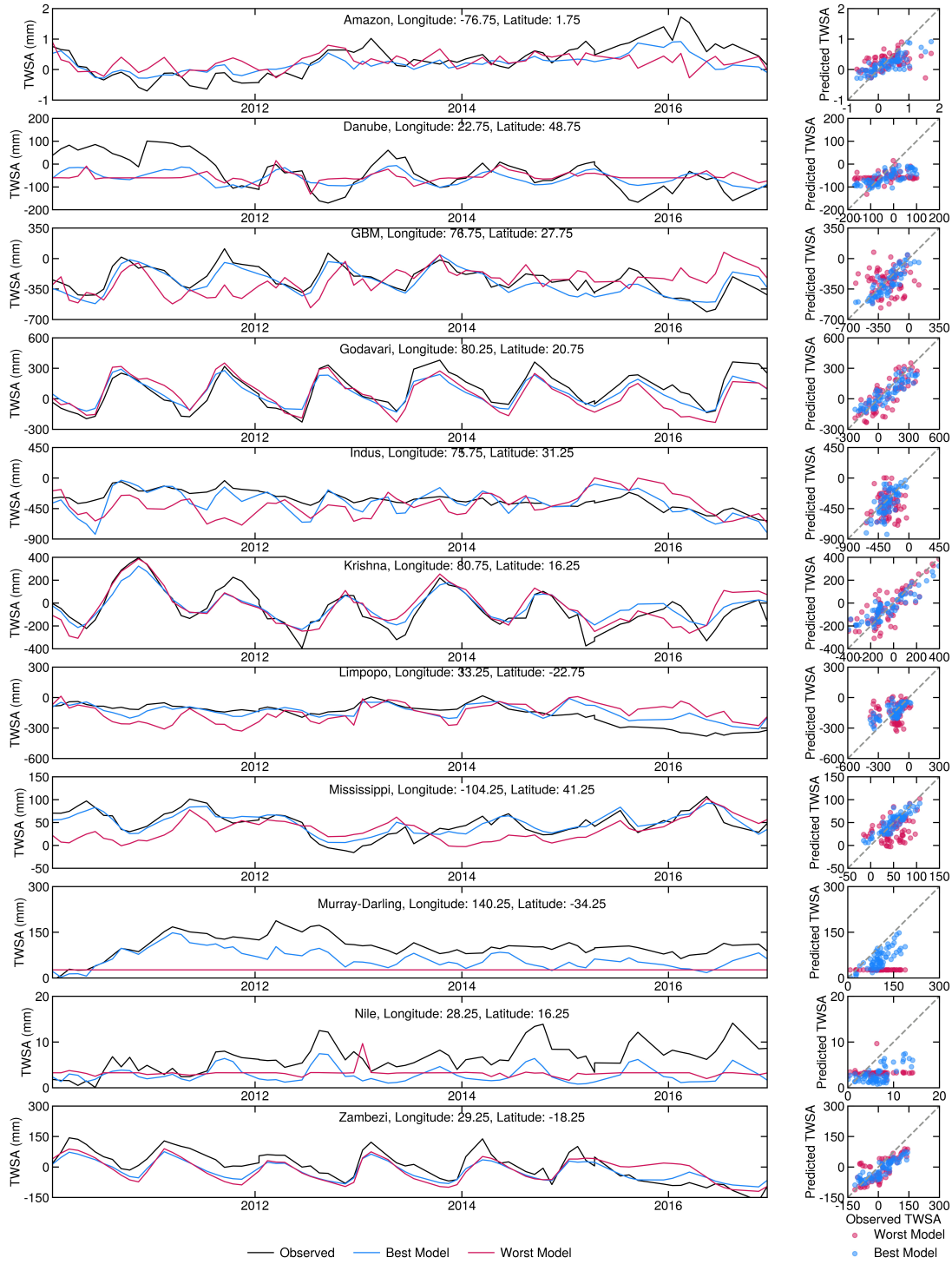


Figure 2: Time series (left columns) for grid cells showing the maximum improvement within the basins considered in this study, including observed TWSA and TWSA predicted by the best and worst models. Scatter plots (right columns) compare the TWSA predicted by the best and worst models with the observed TWSA.

**RC 1.7:** Figure 8: Y-Axis labels should be put on the left side, next to the numbers.

**Response:** This figure has been updated in the revised manuscript according to the reviewer's suggestion. Please find the revised figure below.

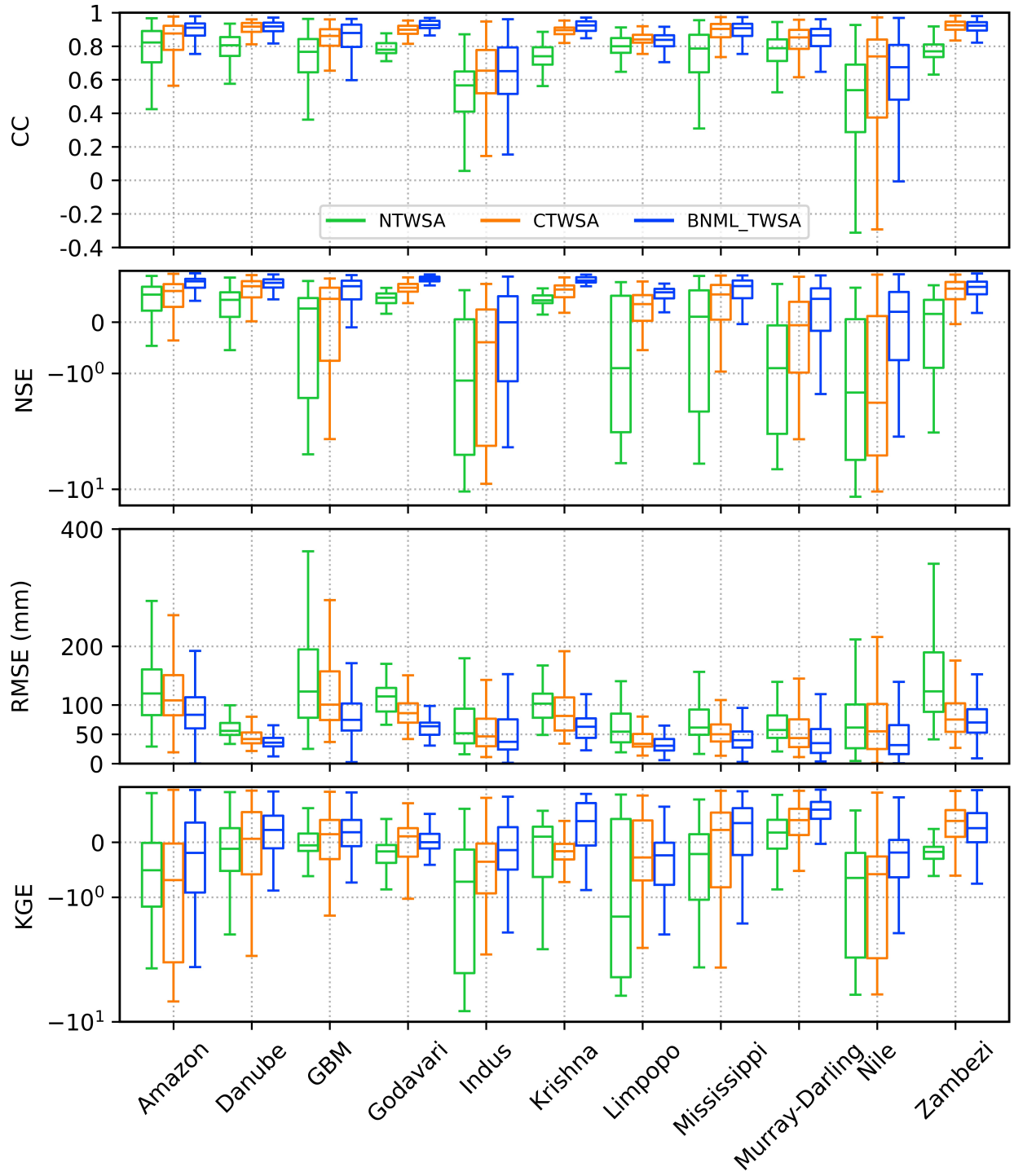


Figure 3: Box plot of CC, NSE, RMSE and KGE values for the grid cells of each basin, excluding the outliers.

**RC 1.8:** Figure 9, 11, 19: Units (mm) are missing for the y-axes.

**Response:** We thank the reviewer for bringing it to our notice. In the revised manuscript, we have rectified these figures and have shown the updated versions below.

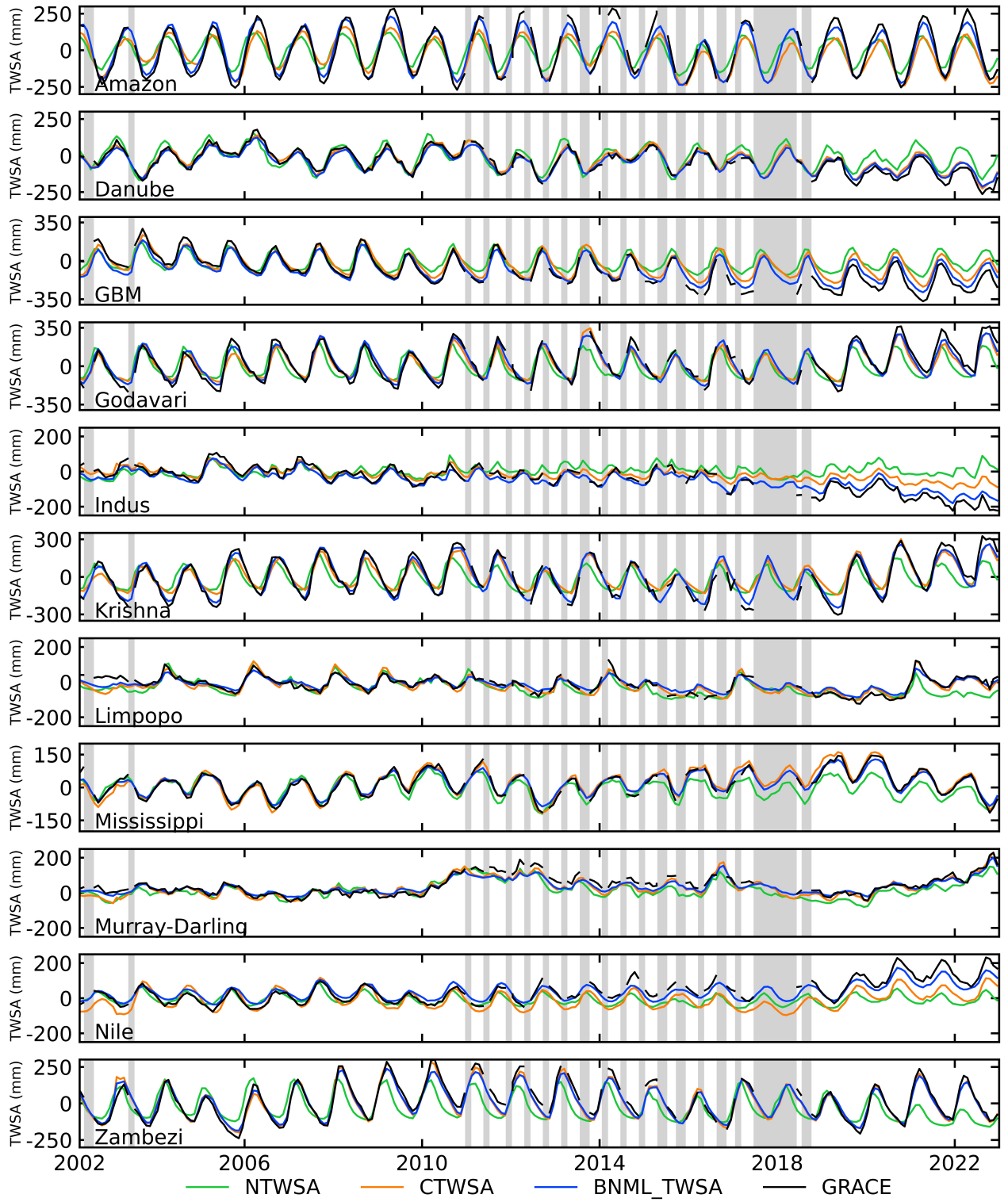


Figure 4: Comparison of TWSA time series from April 2002 to December 2022 (GRACE period). Vertical gray bars indicate missing GRACE observations.

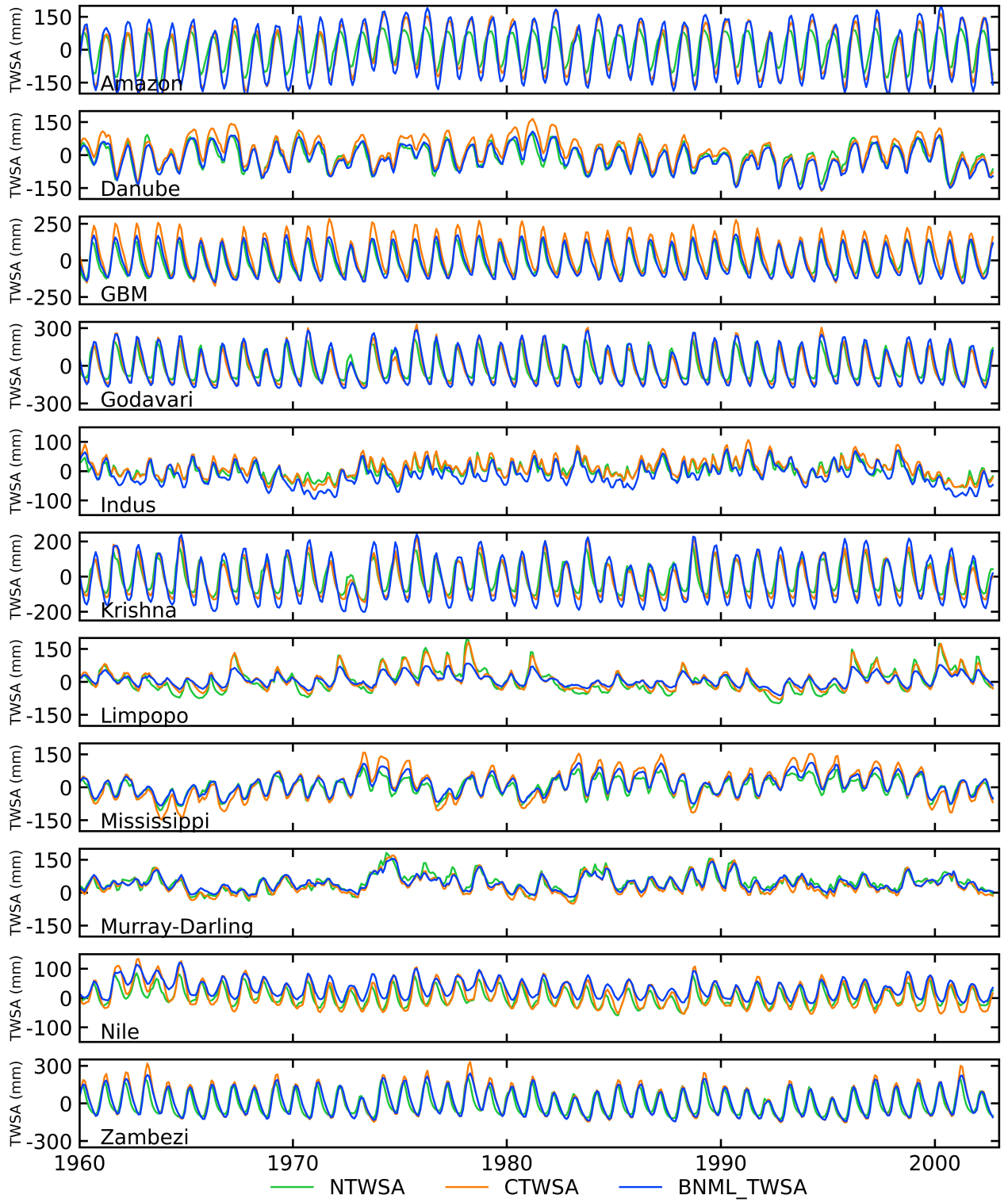


Figure 5: BNML\_TWSA during the pre-GRACE period (Jan 1960 - Mar 2002)



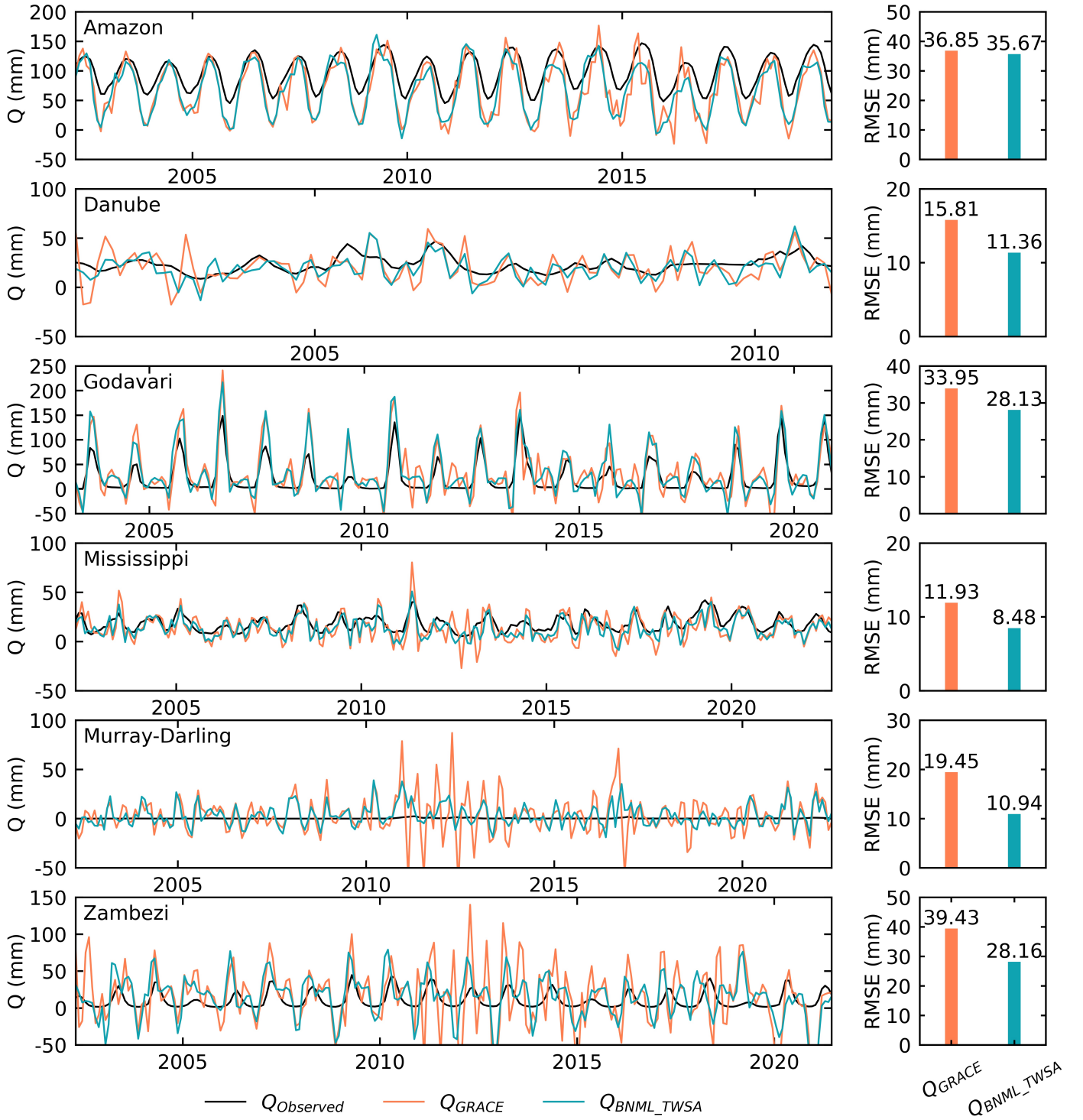


Figure 6: Comparison of observed streamflow ( $Q_{Observed}$ ),  $Q$  obtained from water balance using GRACE TWS data ( $Q_{GRACE}$ ) and  $Q$  obtained from water balance using BNML-TWSA TWS data ( $Q_{BNML\_TWSA}$ ). RMSE values (right columns) obtained for  $Q_{GRACE}$  and  $Q_{BNML\_TWSA}$  against  $Q_{Observed}$ .

**RC 1.9:** Figure 18: Notation (Label) for CC is missing for the color-scale

**Response:** We appreciate the reviewer's constructive feedback. The label "CC" has now been added to the color scale in Figure 18 in the revised manuscript.

## Data product:

**RC 1.10:** Once you publish the global data set, I encourage you to provide also the global grids as netcdf files – instead of only catchment based portions. But probably this is your intention anyway. I think it would be good to provide also the Model Input and BN Predictors data as netCDF grids. For the BN Predictors, you could use an integer based classification. With your current method, you will have to generate a huge number of text-files for the global grid. Importing these data into models will be much easier based on netCDF. Please check also the CF compliance of your netCDF files. Lat and lon seem to have issues here so that cdo does not recognize the grid type (latlong).

**Response:** After acceptance of the manuscript, we will upload the netcdf files for BNML\_TWSA products, Model Input, and BN Predictors using an integer-based classification. Additionally, we will check the CF compliance of our generated NetCDF files. We sincerely apologize for the issue regarding opening the NetCDF files in CDO. While publishing the global dataset, we will ensure it has CF compliance.



## Anonymous referee #2

**RC 2.0:** I appreciate efforts by the authors to address my comments substantially. Now I only have some minor comments.

**Response:** Thank you for acknowledging our efforts to address your comments. We appreciate your thorough review and are committed to addressing the remaining minor comments.

### Minor and technical comments:

**RC 2.1:** L37-46: I suggest to add a topic sentence at the head of the paragraph, like in the next paragraph. Currently, the paragraph is just a list of example studies and the reader should guess what the paragraph is about based on them.

**Response:** We appreciate the suggestion and in the revised manuscript (P:2,L:39-41), we have added the following at the head of the paragraph:

“It is imperative that a reliable, long-term continuous TWSA dataset is valuable for various aspects of the functioning of the Earth system including the assessment of basin-scale water balance and local hydrological extremes. Different studies have employed various techniques to reconstruct long-term TWSA beyond the GRACE period.”

**RC 2.2:** L66-87 (A) and L88-109 (B): Each of two paragraphs deals with two topics, so it is less clear what each paragraph wants to talk about. Paragraph A deals with the importance of 1) testing multiple ML algorithms and 2) the optimal feature selection. Paragraph B deals with the importance of the optimal feature selection and debriefs the whole study. I would suggest merge the parts about the optimal feature selection as a new paragraph so that there are three paragraphs about multiple ML algorithms, the optimal feature selection, and debriefing the study.

**Response:** We appreciate the suggestion and have updated the concerned portion of the manuscript in the revised version (P:3-4, L:70-115).

**RC 2.3:** Equation 1: I think Eq. 1 can be missed in the current form and a reader may regard that Noah considers surface water storage. Sun et al. (2019) used the equation to show the TWS calculation in general, not for a specific product or model. However, in this study, Equation 1 is to introduce how TWS is calculated in Noah, which does not calculate SWS. So, I suggest to replace CSWS with CWS.

**Response:** In the revised manuscript, we have incorporated the reviewer’s suggestion (P:7, L:167-171). The updated content is also shown below.

$$TWS = SnWE + SMC + CWS \quad (5)$$

where SnWE represents snow depth water equivalent, SMC is soil moisture content, and CWS is canopy water storage. We have also updated Figure 1 and Table 1 in the revised manuscript to incorporate this information.

**RC 2.4:** Figure 4: The six grid cells highlighted are all neighboring to each other. I understand the logic to select the grid cells based on the differences in CC, but using the resulting sample group is less informative for the purpose of the figure, I would say. I suggest to consider selecting other grid cells from other regions, too.

**Response:** We appreciate the reviewer’s suggestion to select grid cells from other regions. In the revised manuscript, we have updated this figure to include a grid cell showing the maximum improvement within a river basin for each of the 11 river basins considered in this study (P: 15, L:335-338). The updated figure is also provided in Figure 2 of this response document.

**RC 2.5:** L358-360: Two last sentences are duplicated in terms of the message.

**Response:** We sincerely appreciate the reviewer’s effort in highlighting these duplicate sentences. In the revised manuscript (P:19, L:364-365), we have corrected this issue as follows: “The results indicate the superior performance of BNML-TWSA compared to the other methods, which is evident from these plots on a global scale.”

**RC 2.6:** L417-418: I think it’s wrong to say like “the proposed model accurately reproduced the TWSA series during the pre-GRACE period” by comparing three model simulations.

**Response:** In the revised manuscript, we have updated this sentence as follows (P:26, L:423-424): ‘The proposed model simulated the TWSA series during the pre-GRACE period and identified recorded climate extreme events that occurred in the hindcast period.’

**RC 2.7:** L435-436: I suggest to specify the period.

**Response:** We sincerely appreciate the reviewer’s suggestion. In the revised manuscript (P:29, L:441-442 and P:31, L:449-450), we have specified the period of comparison as follows:

P:29, L:441-442: “Specifically, the period spans from April 2002 to July 2019 for JPL\_ERA5 and from April 2002 to December 2016 for JPL\_MSWEF.”

P:31, L:449-450: “...for the period from April 2002 to July 2018...”

**RC 2.8:** L515-516: It’s hard to agree. Both BNML\_TWSA and GRACE struggle to get the observation in rather drier basins, and they behave similarly in general in other basins. One can at least highlight the comparable behavior of  $Q_{BNML\_TWSA}$ , or there would need to be additional information (e.g., metrics) to say that  $Q_{BNML\_TWSA}$  performs better clearly.

**Response:** We appreciate the reviewer’s observation regarding the need for proper justification of the improved performance of  $Q_{BNML\_TWSA}$ . In the revised manuscript, we have added RMSE values obtained against observed streamflow for both  $Q_{BNML\_TWSA}$  and  $Q_{GRACE}$  and depicted them alongside the time series plot (Figure 6 of this response document). Additionally, we have updated this result in the “**Comparison with streamflow measurements based on basin-scale water balance**” sub-section of the revised manuscript (P:36, L:522-524).

**RC 2.9:** L521-524: Is it the uncertainty of GRACE as an assumed ground truth that would affect the product evaluation, or is the processing errors in GRACE linked to the uncertainty of BNML\_TWSA? If the former, it would be better to specify it. If it is the latter, BNML\_TWSA is still subject to uncertainties of mascon solutions, although they perform better compared to the spherical harmonic solutions. I would indicate both aspects.

**Response:** We appreciate the reviewer’s valuable suggestion. In the revised manuscript, we have mentioned both aspects as follows (P:39, L:529-536):

“There are various sources that contribute to the uncertainties in reconstructed TWSA. The primary source of uncertainties arises from the measurement errors, inherent processing errors, leakage errors and model assumptions associated with the original GRACE data, as documented by Boergens et al. (2022) and Gao et al. (2023). Nevertheless, this issue is effectively mitigated by utilizing the mascon solution, which demonstrates clear superiority over the spherical harmonics data (Kalu et al., 2024). The JPL mascon solution employs a Coastline Resolution Improvement (CRI) filter to minimize leakage errors across land/ocean boundaries. Additionally, gain factors are utilized to further mitigate these leakage errors. Moreover, a Bayesian framework is implemented to more effectively eliminate correlated errors compared to traditional empirical filters (Wiese et al., 2016).”

**RC 2.10:** L526-527: The improvement by multiple MLs is majorly for (semi-)arid regions (Figure 3). This means that, every ML algorithm perform greatly for wetter regions or none of them could not improve the simulation there. So, for wetter regions, I suspect there are still rooms to improve (e.g., the KGE map in Figure 6).

**Response:** We appreciate the reviewer’s observation regarding the lower KGE values observed in river basins within wet climatic zones, such as the Amazon and Godavari. We concur that this highlights the potential for improvement. As the reviewer suggests, and as we also note in the manuscript (P: 41-42, L: 576-581), integrating machine learning (ML) models with physical models offers a promising avenue for achieving such improvements.

**RC 2.11:** L556-557: The sigma in arid regions may be small by nature because of the small variance of TWS there. However, I agree that testing multiple ML algorithms partly contributed to the improvement in performance (and small sigma values) in arid regions, referring to Figure 3.

**Response:** We sincerely thank the reviewer for this insightful observation regarding the improvement in performance in arid regions. Indeed, testing multiple ML algorithms contributes to performance enhancement globally. However, this improvement is particularly significant in arid regions.

**RC 2.12:** Please make sure that the uncertainty data from Figure 20b is included when the published datasets are updated.

**Response:** We will upload the uncertainty data in terms of standard error as a netCDF file when the published datasets are updated. This information is also included in the Data Availability section of the revised manuscript.

## References

- Boergens, E., Kvas, A., Eicker, A., Dobslaw, H., Schawohl, L., Dahle, C., Murböck, M., and Flechtner, F.: Uncertainties of GRACE-Based Terrestrial Water Storage Anomalies for Arbitrary Averaging Regions, *Journal of Geophysical Research: Solid Earth*, 127, e2021JB022081, <https://doi.org/https://doi.org/10.1029/2021JB022081>, e2021JB022081 2021JB022081, 2022.
- Gao, S., Hao, W., Fan, Y., Li, F., and Wang, J.: A Multi-Source GRACE Fusion Solution via Uncertainty Quantification of GRACE-Derived Terrestrial Water Storage (TWS) Change, *Journal of Geophysical Research: Solid Earth*, 128, e2023JB026908, <https://doi.org/https://doi.org/10.1029/2023JB026908>, e2023JB026908 2023JB026908, 2023.
- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, 377, 80–91, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Kalu, I., Ndehedehe, C. E., Ferreira, V. G., and Kennard, M. J.: Machine learning assessment of hydrological model performance under localized water storage changes through downscaling, *Journal of Hydrology*, 628, 130–157, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2023.130597>, 2024.
- Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, *Journal of Hydrology*, 424–425, 264–277, <https://doi.org/https://doi.org/10.1016/j.jhydrol.2012.01.011>, 2012.
- Wiese, D. N., Landerer, F. W., and Watkins, M. M.: Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution, *Water Resources Research*, 52, 7490–7502, <https://doi.org/https://doi.org/10.1002/2016WR019344>, 2016.