https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

Benchmarking multimodel terrestrial water storage seasonal cycle against GRACE observations over major global river basins

Sadia Bibi[1], Tingju Zhu[1*], Ashraf Rateb[2], Bridget R. Scanlon[2], Muhammad Aqeel Kamran[3], Abdelrazek Elnashar[4], Ali Bennour[5], Ci Li[1]

[1]ZJU-UIUC Institute, International Campus, Zhejiang University, China

[2]Bureau of Economic Geology, Jackson School of Geosciences, University of Texas at Austin, Austin, TX, USA,

[3]Department of Environmental and Resource Sciences, Zhejiang University, China

[4]Department of Natural Resources, Faculty of African Postgraduate Studies, Cairo University, Giza 12613, Egypt.

[5]State Key Laboratory of Remote Sensing Sciences, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100101

Correspondence to: Tingju Zhu (tingjuzhu@intl.zju.edu.cn)

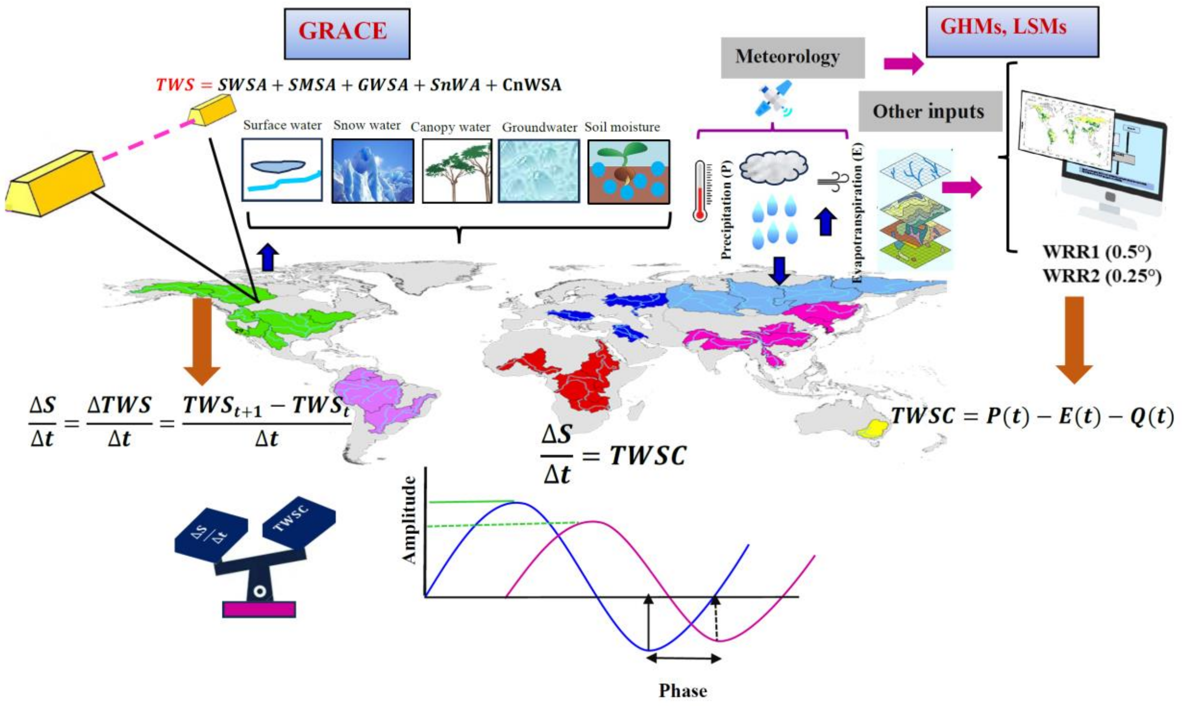
Abstract

The increasing reliance on global models for evaluating climate and human-induced impacts on the hydrological cycle underscores the importance of assessing their reliability. Hydrological models provide valuable data on ungagged river basins or basins with limited gauge networks. The objective of this study was to evaluate the reliability of 13 global models using the Gravity Recovery and Climate Experiment (GRACE) satellites total water storage (TWS) seasonal cycle for 29 river basins in different climate zones. Results show that the simulated seasonal total water storage change (TWSC) does not compare well with GRACE even in basins within the same climate zone. The models overestimated the seasonal amplitude in most boreal basins and underestimated it in tropical, arid, and temperate zones. In cold basins, the modeled phase of TWSC precedes that of GRACE by up to 2-3 months. However, it lags the GRACE phase by one month over temperate, arid to semi-arid basins. There was good agreement between GRACE and model amplitudes in the tropical zone. With the findings and analysis, we concluded that R2 models with optimized parametrizations have a better correlation with GRACE than the reverse scenario. This signifies an enhancement in the predictive capability of models regarding the variability of TWSC. The seasonal

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

amplitude and phase-difference analysis in this study provide new insights into the future improvement of large-scale hydrological models and TWS investigations.

Keywords: Global hydrological models, Land surface models, GRACE, hydrological system, total water storage, seasonal cycle



1. Introduction

In the face of global climate change, there has been a growing focus on total water storage (TWS) as a crucial metric of the global hydrological cycle (Bolaños Chavarría et al., 2022). TWS serves as a comprehensive indicator of water availability, encapsulating various components of water storage, including canopy water, lakes, rivers, snow and ice, soil moisture, and groundwater. It regulates biogeochemical fluxes and energy in the climate system (e.g., the amount and rate of carbon dioxide (CO_2) flux between the land surface and the atmosphere (Pokhrel et al., 2021). Moreover, TWS is associated with flood and drought forecasts and has substantial repercussions for water resources, social safety, and global food security (Tapley et al.,

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

2019). Therefore, monitoring TWS variations is crucial for quantifying water resource availability and improving the understanding of global water, energy, and carbon cycles and their interplay with climate change (Famiglietti, 2004). Irrespective of its hold over numerous processes and mechanisms in Earth's system, integrated TWS measurements are obscure due to poor gauging networks and complex river basin hydrology (Hassan and Jin, 2016).

Hydrological models are forced by precipitation (P) and various climatic parameters to anticipate the storage and flow of water on continents, along with the control of other Earth subsystems, for instance, the oceans and atmosphere via processes such as runoff (Q) and evaporation (E) respectively. Changes in the water budget $\left(\frac{ds}{dt} = P - E - Q\right)$ of specific regions, such as major river basins, play an important role in the accurate monitoring of the stability and dynamical behavior of the water cycle (Werth and Güntner, 2010). For hydrological modeling, a reliable depiction of the continental hydrological cycle and its components are critical. Nevertheless, variations in TWS, on the other hand, become a fundamentally important independent source of information in evaluating large-scale models (Güntner & Güntner, 2008). There are two types of hydrological models at the global scale: Land Surface Models (LSMs) and Global Hydrology Models (GHMs). LSMs have been developed to simulate fluxes between the land and the atmosphere (Bierkens, 2015). LSMs may not produce a reliable estimate of changes in TWS because of their emphasis on energy-flow simulations (Scanlon et al., 2018). The hydrological community has developed GHMs for streamflow modeling at catchment outlets and solving the water balance equation to deal with global water scarcity (Bolaños Chavarría et al., 2022). In contrast to LSMs, GHMs have a more realistic water budget scheme and simulate human interventions, such as water usage and infrastructure for water resources (Veldkamp et al., 2018). GHMs and LSMs perform differently in simulating the TWS owing to different physics and model structures, atmospheric forcing data, parameterization, and land-surface processes (Zhang et al., 2017). The differences between the models vary according to climatic conditions and basin geography, with notable disparities in tropical, snow-dominated, and monsoonal regions (Milly & Shmakin, 2002; Schellekens et al., 2017).

Furthermore, little is known about the geographical significance and features of certain storage processes**.** The lack of global comprehensive independent benchmarks hinders comparing and validating these models. For instance, many LSMs do not account for surface water storage or deeper groundwater (Güntner, 2008). In this regard, large-scale hydrological studies greatly benefit from the Gravity Recovery and Climate Experiment (GRACE) satellites, which were launched in March 2002

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

and have been incredibly helpful for the assessment of hydrological models (e.g., Lo et al., 2010 Schellekens et al., 2017 Trautmann et al., 2018) as well as understanding global hydrological processes (Li et al., 2019; Eicker et al., 2014) and water storages (e.g., Kim et al., 2009). GRACE measurements have been applied to calculate model parameters and to evaluate model simulations at regional (Lo et al., 2010), continental (Trautmann et al., 2018), and global (Kraft et al., 2022; Trautmann et al., 2022) scales. Compared to GRACE-derived TWS trends, Scanlon et al. (2018) revealed that the TWS trends of GHMs were either underestimated or had the opposite sign over basins across the globe due to human intervention and climate change respectively. Other studies focused on the seasonal cycle of TWSC to identify disparities between models and GRACE for instance Zhang et al (2017) validated TWSC simulations from four hydrological models and found that model runs generally agree with observations only to a very limited extent. Discrepancies among the models were not solely attributable to uncertainties in meteorological forcing but rather to the model structure, parametrization, and representation of discrete storage components with dissimilar spatial features. In their comparison of basin average TWSC from GRACE with seven hydrological models over a seasonal time frame, Scanlon et al (2019) emphasized the implication of water storage components in addition to water fluxes to enhance model performance. They discovered that changes in modeled fluxes overestimate seasonal TWSC in northern basins while underestimating storage capacities in tropical basins due to a lack of storage compartments (such as surface water and groundwater). Nevertheless, the phase difference between GRACE and the modeled TWSC seasonal cycle was not generally covered.

In this study, we take advantage of Water Resource Reanalysis tier-1 and tier-2 products which provide a large set of LSMs and GHMs (Schellekens et al., 2017). We investigate the performance of 13 models in simulating the amplitude and phase at seasonal cycle relative to the latest release (RL06) of GRACE TWS for 29 major river basins under different climates.

Unique aspects of this study include:

1. Benchmark seasonal TWSC amplitudes and phase based on 13 GHMs and LSMS against GRACE.
2. Compare high-resolution forcing and more optimized structured R2 models against R1 models and access their ability to simulate TWSC variability and replicate water storage against the latest release (RL06) of GRACE TWSC.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

2. Materials and Methods

2.1. Global River Basins

We selected 29 major global river basins (Fig. S1) with drainage areas of ≥ 500,000 km^2 (Table 1). According to the Köppen–Geiger climate (KGClim) classification scheme for 1984-2013 (Cui et al., 2021) (Fig. S2), these basins cover five climate zones: polar, boreal, temperate, arid, and tropical. The dataset referred to as KGClim is publicly available at 1km spatial resolution and can be downloaded at https://doi.org/10.5281/zenodo.5347837.

2.2. GRACE data

We used release 6 (RL06) mascon solutions from the University of Texas Center for Space Research (CSR-M), and the Jet Propulsion Laboratory (JPL-M) water storage data (2003-2014), of equivalent water thickness. The data over the study period were sufficient to accommodate the average changes in the seasonal cycle of land water storage. Mascon solutions are great improvements over traditional spherical harmonics. Unlike spherical harmonics, mascon solutions do not require a postprocessing filter (Watkins et al., 2015; Save et al., 2016) and are more applicable to regional and global scales. JPL-M applies a coastline filter to attenuate the leakage between the ocean and land, and scale factors were applied at a grid scale to strengthen the signal smaller than three degrees. The CSR-M uses a finer hexagonal at a quarter-grid degree for coastline filters. The missing months in the GRACE record were filled using linear interpolation (Xiao et al., 2015; Liesch & Ohmer, 2016). GRACE data can be accessed through these websites, https://podaac.jpl.nasa.gov/dataset/TELLUS_GRACE_MASCON_CRI_GRID_RL06_V1 / http://www2.csr.utexas.edu/grace/RL06_mascons.html.

2.3. Earth2Observe global water resources reanalysis data

We evaluated 13 hydrological models based on the global Water Resources Reanalysis (WRR). WRR datasets are large collections of LSMs and GHMs, developed by the eartH2Observe (E2O) (Schellekens et al., 2017) and contain the outputs of models at two spatial resolutions represented as WRR1 and WRR2 (0.5° and 0.25° respectively). The model runs generated from WRR1 are abbreviated as "R1" whereas model runs from WRR2 were abbreviated as "R2". R1 models were forced with ERA-Interim data (WFDEI) meteorological reanalysis dataset (Weedon et al., 2014) at a 0.5° spatial resolution from 1979 to

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

2012. While R2 models were forced with the Multi-Source Weighted Ensemble Precipitation (MSWEP) dataset (Beck et al., 2017), at a spatial resolution of 0.25° from 1980 to 2014, which was used to force the R2 model. In R2 models, model algorithms were improved to better represent the hydrological processes by integrating anthropogenic impacts and earth observation inclusions (Gründemann et al., 2018). A detailed description of these datasets and the improvement from R1 to R2 models can be found in Dutra et al. (2015), Dutra et al. (2017), and Schellekens et al. (2017), respectively. The models used in this study are presented in Table S1.

We investigated seasonal TWS anomalies from large-scale GHMs, including PCR-GLOBWB (R1 and R2), LISFLOOD (R1 and R2), HBV-SIMREG_R1, W3RA_R1, SWABM_R1, and WaterGAP3 (R1 and R2), and LSMs, HTESSEL (R1 and R2), JULES_R1, and Surfex-Trip (R1 and R2).

To benchmark the selected models against GRACE TWS (JPL and CSR mascon), 2003 to 2012 period was used as a common period for R1 and GRACE, and 2003 to 2014 for R2 models and GRACE TWS. E2O data can be accessed through the E2O Water Cycle Integrator portal (https://wci.earth2observe.eu/).

2.4. Assessment of model performance

The monthly total water storage anomaly (TWSA) is the sum of all continental storage as

$$TWSA = SWSA + SMSA + GWSA + SnWA + \mathrm{CnWSA} \qquad (1)$$

Where SWSA is surface water storage, SMSA is soil moisture storage; GWSA is groundwater storage; SnWA is the snow water equivalent and CnWSA is canopy water storage.

To derive the $\frac{\Delta S}{\Delta t}$ rate of change from the models we used equation (2)

$$\frac{\Delta S}{\Delta t} = \frac{\Delta TWS}{\Delta t} = \frac{TWS_{t+1} - TWS_t}{\Delta t} = TWSC = P(t) - E(t) - Q(t) \qquad (2)$$

where TWSC is the climatological change in TWS, Q is the total outflow (net surface and groundwater outflow), t is time, and P and E are totals of precipitation and actual evapotranspiration, respectively. The seasonal cycle was calculated by taking an average of each month (from January to December).

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

2.5. Statistical analysis

A Taylor diagram is a visual approach used to describe how well data (or data sets) corresponds to the observations (Karl E. Taylor, 2001). The resemblance between the two data sets was quantified using their correlation, cantered root-mean-square difference, and standard deviation (representing the amplitude of variations). Taylor diagrams are particularly helpful in assessing various statistical aspects of complicated models or in evaluating the different models. Details of correlation coefficient R and RMS difference E are given in the supplementary information.

3. Results

We used GRACE CSR_M seasonal cycle to validate the GHMs and LSMs simulated seasonal cycle. We grouped models as GHMs (R1 and R2) and LSMs (R1 and R2) and presented the average behaviour of each group against GRACE CSR-M.

3.1. Comparison of seasonal amplitude between GRACE and models

In the snow-dominated catchments, GHMs overestimate the TWSC by ~6mm to 90 mm except for Saint Lawrence, Mackenzie, Yukon, and Ob River basins (underestimated by ~6 to 25 mm) while LSMs (both R1 and R2) underestimate it by ~7 mm to 43 mm, moreover, GHM R2 models perform better against GRACE (Fig. 1). Among four Serbian basins (Ob, Yenisei, Lena, and Kolyma), over Kolyma and Lena basins, GRACE shows a seasonal amplitude of 30 mm. However, the LSMs (R1) underestimate it by 10 mm, and GHMs overestimate it by 15-26 mm. Over the Yenisei and Ob basins, GRACE amplitude was at ~50-59 mm, however, both LSMs and GHMs underestimate TWSC with LSMs underestimating it by ~33 mm and GHMs by ~20-25 mm for R1 and R2 respectively. Among GHM-R1 models, SWBM-R1 behaved differently over these two basins where TWSC amplitude was ~16-26 mm higher than GRACE.

In the European basin, the Volga River basin is one of the largest basins, spanning across central Russia. The amplitude of GRACE TWSC over this basin was 67 mm. However, the R1 and R2 LSMs underestimated the TWSC amplitude by approximately 10-20 mm. Among the R1 GHMs, SWBM, W3RA, and WaterGAP3 overestimated the amplitude by around 25-90 mm compared to GRACE. While the GHMs R2 amplitude appeared at ~63 mm, which was relatively consistent with the measurements from GRACE.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

Over the Amur basin in northeast Asia, the GRACE signals were weak, and the amplitude was recorded at 14 mm. Both LSMs underestimated the TWSC by ~ 8-5 mm (R1 and R2) though GHMs overestimated TWSC by ~6-12 mm.

In the North American basins, understanding the TWS variations in the Saint Lawrence, Yukon, and Mackenzie River basins is crucial for managing water resources in this region. Over the Saint Lawrence River basin, GRACE shows a seasonal amplitude at 30 mm although both models' (GHMs and LSMs) behaviour was very ambiguous. Generally, LSMs underestimated the seasonal amplitude by ~7 - 19 mm for R1 and R2 models. GHMs R2 models marginally underestimated it by ~6 mm. Among GHM R1 models, SWBM and WaterGAP3 amplitudes were exceptionally high and overestimated TWSC by ~23-36 mm. In the Yukon basin, the GRACE TWSC was 63 mm but LSMs underestimated it by 43 mm and 40 mm for R1 and R2 models respectively. Nevertheless, GHMs underestimated the TWSC by 24 mm and 16 mm for R1 and R2 models. In the Mackenzie River basin, GRACE seasonal amplitude was 47 mm but both LSMs (22mm and 30mm for R1 and R2) and GHMs (~10 both for R1 and R2) underestimated the TWSC against GRACE.

Over the temperate zone, all GHMs (R1) and LSMs (R1 and R2) underestimate the seasonal amplitude by ~7 mm to 118 mm. GHMs R2 models show good agreement with GRACE TWSC over the Yellow and Rio Grande basins (Fig. 2). In Australia, the GRACE TWSC amplitude over the Murray-Darling River basin was recorded at 56 mm and both LSMs and GHMs underestimated it by ~10mm.

In two Chinese basins, GRACE seasonal amplitude was at 42 mm over the Yangtze River basin while GHMs and LSMs underestimated it by ~20 mm. In the Yellow River basin, GRACE signals were very weak, and TWSC from GRACE and GHMs R1 has an identical amplitude of 10 mm. While GHMR2 and LSMs overestimate it by ~5-9mm.

Over two major river basins in Southeast Asia, the GRACE had strong signals in the Brahmaputra-Ganges River basin where the TWSC amplitude was at 157 mm, while all the GHMs and LSMs underestimated it by ~79 mm to 118 mm. Surfex-Trip-R2 performed comparatively better among all models underestimating GRACE by 60 mm. Whereas in the Indus River basin, GRACE signals were weak and TWSC amplitude appeared at 36 mm. R1 models (LSMs and GHMs) slightly underestimated the TWSC by ~7-6 mm while the GHM R2 models agreed well with GRACE (36 mm). However, LSM R2 marginally overestimated the storage and the amplitude appeared at 40 mm.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Euphrates is the largest river in western Asia, GRACE TWSC amplitude over this basin was 53 mm whereas all the GHMs and LSMs underestimated it by ~34mm and 38 mm respectively. Danube River basin is the second largest European river basin, GRACE seasonal amplitude was recorded as 79 mm while all the GHMs and LSMs underestimated it by ~48mm to 54 mm respectively.

In the North American river basins, models did not exhibit a pronounced seasonal cycle of water storage change. At Columbia and Mississippi River basins, seasonal TWS fluctuations are subject to seasonal evolution of the moisture convergence. GRACE amplitude was 114 mm though models occurred at ~46 mm to 86 mm respectively. Over the Mississippi River basin, model peaks appeared nearly flat against the GRACE seasonal amplitude of 63 mm, and GHMs and LSMs underestimated it by ~47 mm to 53 mm respectively. In the California region, GRACE maximum storage change was 76 mm and models underestimated it by ~35mm to 49 mm. Over the Rio Grande basin GRACE signals were very weak, the TWSC amplitude was at 6 mm, and all GHMs agreed well with GRACE (9 mm and 8 mm for R1 and R2 models respectively). While LSMs overestimated the storage change amplitude by ~6-18 mm.

All the GHMs and LSMs overestimated the troughs across all the temperate basins except over the Rio Grande where TWSC ditch from SWBM-R1 appeared ~19 mm lower than GRACE CSR-M and Indus River basins where all the models underestimated the low storage change.

In arid basins, all the GHMs and LSMs underestimated the TWSC amplitude by ~22 mm to 145 mm (Fig. 3). Models and GRACE responded similarly over the Niger and Nile River basins since they were located at the same latitude. Over the Niger River basin, GRACE amplitude appeared at 112 mm while models underestimated it by ~69 mm to 90 mm. Over the Niger Basin SWBM-R1 amplitude was 56 mm, which is half the height of GRACE maxima. Similarly, over the Nile River basin, GRACE TWSC was observed at 59 mm while model amplitudes were at ~29mm to 42mm below GRACE TWSC. SWBM-R1 performed relatively better than other models over the Nile River basin, with an amplitude of 37 mm (22 mm shorter than GRACE). Similarly, in the Zambezi GRACE TWSC peak was recorded at 187 mm while models substantially underestimated the storage change and model amplitude appeared at ~115 mm to 157 mm below the GRACE TWSC. In São Francisco, GRACE TWSC showed a clear climatology, seasonal amplitude was 70 mm while models underestimated it by ~31 mm to 41 mm. Similarly, over Parana GRACE amplitude was recorded at 65 mm while models underestimated by TWSC ~47 to 51 mm

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences Discussions

below the GRACE. The models' behavior was very ambiguous, especially over the Parana River basin. HTESSEL-R1 performed comparatively better among all the models and underestimated the TWSC amplitude at 145 mm GRACE (187 mm) in this basin. Over São Francisco and Prana basins, the best-performing model, PCR-GLOBWB-R1, had amplitudes of 31 mm and 42 mm, respectively.

Over the four tropical basins (Fig. 4), all models underestimated the TWSC amplitude against GRACE. In the Mekong River basin, GRACE signals were very strong and the TWSC amplitude was at 230 mm while the GHMs and LSMs greatly underestimated it and TWSC amplitude ranged between ~134 mm to 191 mm below the GRACE. Over the Congo River basin, LSMs R1 and GHMs (R1 and R2) amplitude were at ~16 mm to 18 mm below the GRACE TWSC of 40 mm, though LSM R2 showed relatively better performance where TWSC amplitude was 6 mm below the GRACE. Surfex-Trip-R2 and PCR-GLOBWB -R1 were the relatively best-performing models over this basin where seasonal amplitudes were at 44 mm and 35 mm respectively. In the Orinoco basin, the GRACE amplitude was at 178 mm while all the models underestimated it by ~112 mm to 160 mm. In the Amazon basin, the GRACE amplitude was recorded at 178 mm while models greatly underestimated it by ~149 mm to 156 mm. Generally, over the Orinoco and Amazon River basins, PCR-GLOBWB-R1 gave better estimates of TWSC amplitude as compared to other models.

3.2 Phase difference between GRACE and models

The seasonal cycles of the boreal basins show TWS peaks in spring, which are largely generated by snowmelt. In snow-dominated basins (Fig. 1) seasonal TWSC variations from models and GRACE exhibited consistency in the timing of crest except over the Saint Lawrence River basin where Surfex-Trip-R1and SWBM-R1peaks appeared one month earlier than GRACE, while troughs were inconsistent with GRACE TWSC over all the basins. The model TWS precedes GRACE by 3-4 months. The trough in GRACE for all the basins started in September (except for the Kolyma and Amur basins where they started in October) while in models trough began in June, giving models a 3-month lead. Similarly, over Yenisei and Amur basins (July) and Saint Lawrence (where most of the models showed ditch in May), models were 4 months ahead of GRACE observations.

There was no phase difference between modeled and GRACE TWSC in the temperate zone except for the Yellow River and Rio Grande River basin where GRACE peaks were ahead of modeled TWSC by one month (Fig. 2).

In arid basins modeled TWSC peaks have an identical phase with GRACE TWSC over Niger and Nile River basins. While over the Zambezi and São Francisco River basins modeled TWSC peaks appeared in April, resulting in a one-month time lag over these two basins compared to GRACE where peak storage was recorded in March (Fig. 3).

Models and GRACE TWSC phase were quite consistent with GRACE over the Orinoco and Amazon River basins in the tropical zone. However, the GRACE peak over the Congo River basin was observed earlier in April while modeled peaks were noted in May. Similarly, over the Mekong Rivers, GRACE observed peak water storage change was observed in September while the models' peak appeared in October (Fig. 4).

3.3 Evaluation of model performance

In cold basins (Fig. 5) Taylor's diagram does not clearly distinguish which of the 13 models better represents TWSC compared to GRACE. It is worth noting that correlations between the models and GRACE are weak over all the basins and it ranges from R=0.1 to 0.5. The highest correlation (R=0.5) is found for the PCR-GLOBWB _R1, HTESSEL_R1, and HTESSEL_R2 over Mackenzie, Volga, Ob, and Yenisei River basins, respectively. Almost all models have smaller standard deviations than those observed by GRACE while RMSE was very high and ranged between 25 to 90 mm.

Figure 6 demonstrated the correlation between modeled and GRACE TWS over 11 temperate river basins. All 13 models had a good correlation with GRACE over the Columbia (R~0.6) and Brahmaputra-Ganges River basins (R~0.5 to 0.6) (except LISFLOOD_R1, SWBM_R1, and WaterGAP3_R1 which had a poor correlation over the Brahmaputra-Ganges River basin). Overall, SWBM_R1 demonstrated a good agreement with GRACE over Euphrates, Columbia, and California basins while LISFLOOD_R1 showed the lowest correlation against GRACE over this region., All the models exhibited no correlation with GRACE over the Rio Grande, Yellow River, and Yangtze River basins. All models have smaller standard deviations than GRACE observations, and RMSE ranged between 25 and 120 mm.

Figure 7 shows the correlation between models and GRACE TWSC climatology over five arid river basins. All 13 models had a strong correlation with GRACE over Niger River basins, with R ranging from 0.5 to 0.74. The highest correlation was observed for SWBM_R1 (r=0.74) and the lowest for WaterGAP3_R1 (R=0.5). Furthermore, from the 8 GHMs, HBV-SIMREG_R1, LISFLOOD_R1, PCR-GLOBWB_R1, SWBM_R1, and W3RA_R1 had good correlation over Zambezi and Nile River basins while all the 5 LSMs also showed good agreement with GRACE over the above-mentioned basins. All the

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

models exhibited no correlation with GRACE over São Francisco and Prana River basins except PCR-GLOBWB_R1 which had a good correlation with GRACE over the Prana River basin. All models showed a lower standard deviation than GRACE over this region and RMSE ranged between 30 and 110 mm.

Figure 8 reveals that compared to other climatic zones; models showed a good correlation with GRACE in the tropical zone. All models had a high correlation with GRACE over the Amazon River basins with R=0.6-0.74 except LISFLOOD R1. Apart from HBV-SIMREG_R1 and W3RA_R1, other GHMs did not correlate with GRACE observations over the Congo River basin. Thus, HBV-SIMREG_R1 and W3RA_R1 were the best-performing models while WaterGAP3_R1 was the least-performing GHM that correlated with GRACE only over the Amazon basin in this region. However, all LSMs exhibited excellent performance over these basins. Furthermore, R2 GHMs and LSMs revealed an excellent performance compared to R1 models. Almost all models have smaller standard deviations than GRACE observed TWS and RMSE ranged from 35 to 150 mm.

Fig. 9-10 show the spatial relationship between the monthly time series of GRACE TWSC and the modeled TWSC (GHMs and LSMs respectively). Fig. 9 reveals a spatial correlation between GRACE and GHMs (R1 and R2) TWSC. Some models i.e., HBV_SIMREG_R1 and PCR_GLOBWB_R1 TWSC had a good correlation (≥R=0.6) with GRACE over some basins i.e., Amazon, Marry Darling, and Indus River basin. For LSMs in Fig. 10, the R2 models showed a better correlation with GRACE TWSC than the R1 model. Two R2 models HTESSE_R2 and Surfix Trip_R2 showed a good correlation with GRACE over most of the basins. However, this correlation analysis did not illustrate any evident pattern of correlation (pixel correlation) at the basin scale between GRACE and LSMs monthly time series (Fig. 10). Therefore, the seasonal analysis is a reasonable approach to access the model performance against GRACE observations TWSC. Fig. S3-4 compared seasonal maps of GRACE observations and TWSC estimated from GHMs and LSMs (FMA Spring, MJJ Summer, ASO Autumn, and NDJ Winter). The seasonal map in Fig. S5-S6 revealed that the seasonal amplitude of GRACE is higher than GHMs and LSMs except in the boreal zone.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

4 Discussion

Across a range of time scales, seasonal features are more frequently used as analytical tools. Seasonal variations in TWS play a crucial role in understanding the water dynamics of a region but they have received little attention due to a lack of independent data. We investigate multimodal seasonal TWSC considering amplitude and phase from 13 models against GRACE. We first discussed the seasonal TWSC from GHMs and LSMs benchmarked against GRACE and identified disparities in their amplitudes and timing in different climatic zones. Table 2 summarizes the performance metrics of seasonal TWSC changes computed from 13 GHMs and LSMs against GRACE TWSC where blue boxes corresponded to higher correlation and better performance while red boxes indicated lower scores and poor representation. Overall, the model performed differently in the Northern hemisphere (boreal zones) which are largely dominated by snow. When models simulate the climate patterns, they consider complex interactions between the atmosphere, land surface, and snow cover snow modeling might be the most important factor in this region (Schellekens et al., 2017). However, accurately representing these processes in models can be challenging due to the inherent complexities of the climate system and the limited observational data available. As a result, model behavior in regions dominated by snow, such as boreal zones, may exhibit some discrepancies when compared to real-world observations.

In Yukon and Mackenzie River basins in North America and Serbian basins e.g., Lena and Yenisei and Kolyma, water storage is mainly controlled by changes in snow cover. Models did not show good correlation performance except for PCR-GLOBWB_R1 and HTESSEL (R1 and R2) which exhibited good correlation with GRACE over the basins located between 120° W to 100° E. R2 GHMs (PCR-GLOBWB_R2) and LSM (HTESSEL_R2 and Surfex-Trip_R2) showed much poorer performance than the R1 models. Differences in simulations can be ascribed to the models' structure and their internal dynamics (Bolaños Chavarría et al., 2022). The poor representation of HTESSEL_R2 and Surfex-Trip_R2 could be attributed to various factors including inaccuracies in simulating snow processes, deficiencies in representing other hydrological processes, and inadequate model calibration/validation. Model complexity e.g., increased number of soil layers in HTESSEL _R2 (Table S2) needs to account for additional vertical variations in soil properties, such as moisture content, temperature, and hydraulic conductivity. This complexity introduces more parameters and requires more accurate input data for each layer. If the additional layers are not properly calibrated or the required input data is not available, it can result in increased uncertainty and poorer model performance. Similarly, Surfex-Trip_R2 has improved groundwater, surface energy and snow, flood plains, plant growth, and land use compared to the R1 model. However, if the improvements are not properly accounted for or if the model does not accurately simulate the interactions between plant growth and other hydrological processes, or if the improved vegetation parameters are not properly calibrated, they can introduce biases or inaccuracies that adversely affect the model's performance. Furthermore, improvements in R2 models generally influence reservoir storage rather than surface fluxes (Emanuel et al., 2017). Moreover, the poor performance of PCR-GLOBWB_R2 in the boreal region could be ascribed to a lack of a realistic depiction of the glacier and ice dynamics (Sutanudjaja et al., 2018). Improving the representation of glacier and ice dynamics in PCR-GLOBWB_R2 would require enhancements in the model's parameterization schemes and input data. This could involve incorporating more detailed information on glacier geometry, ice thickness, and movement patterns using remote sensing data, ground-based observations, and specialized glacier models. Additionally, considering the interactions between glaciers and climate variables, such as temperature, precipitation, and radiation, would be crucial for capturing the complex feedback mechanisms or it may just be the presence of water storage in the cold basins that models fail to simulate accurately.

In the temperate regions, all GHMs demonstrated strong agreement with GRACE over the Columbia basin. Among GHMs, HBV-SIMREG_R1, PCR-GLOBWB (R1 and R2), W3RA_R1, and WaterGAP3_R2 also showed good performance over the Barhamaputra-Ganges River basins. All the LSMs also showed excellent performance against GRACE over this basin and HTESSEL_R2 had a good correlation with GRACE (R=0.62). Similar findings were reported by Zhang et al. (2017). Disparities between GRACE and models over other temperate basins can be attributed to the structure of the models, different water storage components for TWS calculation, parameterization as well as differences in runoff simulation and evaporation scheme (Zhang et al. 2017). In our case, the best performing models are HBV-SIMREG_R1, W3RA_R1, JULES, HTESSEL (R1, R2), and Surfex-Trip (R1 and R2) which calculated the runoff by saturation and infiltration excess, and Penman-Monteith method for evapotranspiration (Table2). Nevertheless, the LISFLOOD_R1 also used the same parameterization scheme. PCR-PCR-GLOBWB _R2 and SWBM_R1 also used a similar approach for runoff generation but a different method to calculate evapotranspiration (Hamon (tier 1) or imposed as forcing for PCR-GLOBWB and inferred from net radiations in SWBM), while in WaterGAP3 evapotranspiration was calculated by Priestley–Taylor method and Beta function was used for runoff

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

calculation. To gain a more detailed understanding of why these models behave differently over different basins in the temperate region, it would be necessary to conduct a comprehensive analysis that investigates the specific aspects mentioned above for each model and basin of interest. However, the R2 models' performance was comparatively better than the R1 models in the temperate zone. This is consistent with a previous study of the medium-sized basin in Columbia (Bolaños Chavarría et al., 2022).

In arid basins where subsurface water is the chief controller of TWSC variations, GHMs, and LSMs exhibited a good correlation with GRACE observations over the Niger and Nile River basins. In the Niger River basin, the highest correlation was found for SWBM_R1, HBV-SIMREG_R1, and HTESSEL_R1. Furthermore, HBV-SIMREG_R1, LISFLOOD_R1, PCR-GLOBWB_R1, SWBM_R1, and W3RA_R1 had good correlation over Zambezi and Nile River basins while all the LSMs also showed good agreement with GRACE over the above-mentioned basins. Our results are supported by a previous study conducted over Niger and Nile River basins where JSBACH and MPI-HM models exhibited a quite similar TWSC annual cycle when compared to GRACE (Zhang et al., 2017). However, the models behaved differently over different basins regardless of the differences in the models' structure. PCR-GLOBWB_R2 and WaterGAP3_R2 were among the least-performing models. However, in a previous study of the Limpopo River basin in Southern Africa WaterGAP3_R2 demonstrated the best performance in simulating flood events (Gründemann et al., 2018). The improved routing scheme in PCR-GLOBWB_R2, incorporation of water uses and groundwater abstraction, and reservoir management can also cause significant differences between the models because the addition of more sophisticated routing schemes and the incorporation of various water management components increase the complexity of the model. With added complexity, there is an inherent risk of introducing additional uncertainties or errors into the model. The interactions between different components and processes in the model can become more intricate, making it challenging to accurately capture TWSC. To incorporate water use, groundwater abstraction, and reservoir management components into PCR-GLOBWB_R2, certain assumptions and simplifications have been made. These assumptions can introduce biases or inaccuracies in the estimation.

Over the tropical regions, modeled TWSC had a strong correlation with GRACE observations in the Amazon basin in terms of phase, but models underestimated TWSC amplitude. This indicates that the models were able to simulate the seasonal and interannual fluctuations in water storage, aligning with the observed patterns. However, the fact that the models underestimated

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

the amplitude of TWSC indicates that they did not accurately reproduce the magnitudes of water storage changes as observed by GRACE. Among other models, HBV-SIMREG_R1, PCR-GLOBWB (R1 and R2), W3RA_R1, WaterGAP3 (R1 and R2), HTESSEL_R2, and Surefex-Trip (R1 and R2) demonstrated an excellent representation of TWSC in the Amazon basin where river channel storage is the most important factor in the seasonal TWSC variations and accurate representation of its the dynamics in hydrological models is crucial. This includes accounting for river routing, floodplain dynamics, and water exchanges between the river channels and other storage components. LISFLOOD_R1 did not show any correlation against GRACE over any of the five tropical basins and our results are supported by similar findings reported in a previous study where LISFLOOD_R1 was the worst performing model over medium tropical basin (Bolaños Chavarría et al., 2022). Similar findings were reported in a previous study (Scanlon et al., 2019) where the model underestimate seasonal TWSC in the subtropical zone ~±20° near the equator where modeled medians up to ~40% less than GRACE. LISFLOOD simulates surface water dynamics, including river flow, floodplains, and surface water storage. However, the model might have inherent limitations or simplifications that affect its ability to capture the complex hydrological processes specific to the tropical basins. The model's representation of important factors such as vegetation dynamics, groundwater interactions, or human activities might be inadequate for these regions.

Furthermore, the prevailing pattern may indicate that it is associated with subsided model performance in heavily regulated channel reaches and simulation of man-made structures i.e., reservoirs remain challenging in the LISFLOOD model(van der Knijff et al., 2010). Overall, the R2 (PCR-GLOBWB_R2, WaterGAP3_R2, HTESSEL_R2, and Surefex-Trip_R2) models showed greater agreement with GRACE than the R1 models. Fig. S 5-8 exhibit the distribution of GRACE and grouped model type (GHM or LSM) and forcing resolution (R1 and R2) in four climate zones. Disparities in the seasonal signal of TWSC between GRACE and models can be caused by uncertainties in the models, in GRACE, or both (Scanlon et al., 2019). Zhang et al. (2017) used GRACE observations to validate TWSC simulations from four numerical models over 31 global river basins. They observed that over most of the basins, GRACE error was much smaller than RMS differences and concluded that model uncertainties were the primary cause of the differences. These biases can also result from the simulated storage capacity and storage compartments e.g., SW and GW in the model, uncertainties in inflows/outflows runoff generation, and human interventions in the case of GHM or its absence in the case of LSM.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

4.1 Causes of discrepancies in seasonal amplitudes and phase between models and GRACE TWSC

The differences in seasonal amplitudes and phases between GHMs and LSMs (R1 and R2) and GRACE TWSC can be attributed to several factors:

1. **Model Assumptions**: GHMs and LSMs are based on different assumptions and parameterizations of hydrological processes. They often have different representations of soil properties, vegetation dynamics, and runoff generation mechanisms. These differences can lead to variations in simulated water storage and its seasonal patterns.
2. **Inadequate representation of local hydrological processes**: Models operate at coarse spatial resolutions, which may not capture the intricate details of the hydrological processes specific to these river basins. For example, the models may not adequately simulate snowmelt, glacier dynamics, or the influence of local geological features that can affect water storage.
3. **Input Data and Forcing**: GHMs and LSMs rely on various input data and forcing datasets, such as precipitation, temperature, and land cover information. Differences in the quality, accuracy, and spatial/temporal resolution of these input datasets can influence the simulated hydrological variables, including seasonal amplitudes and phases.
4. **Model Parameterization**: GHMs and LSMs require parameterization of various processes, such as infiltration, evapotranspiration, and groundwater dynamics. The selection and calibration of these parameters can vary among different models, leading to discrepancies in simulated seasonal patterns.
5. **Uncertainty and Errors**: Both models and GRACE have inherent uncertainties and errors. Models rely on various approximations and simplifications, while GRACE measurements are affected by sources of error, such as atmospheric contamination and leakage effects. These uncertainties and errors can contribute to differences in seasonal amplitudes and phases between the models.
6. **Inaccurate representation of human activities**: Human interventions, such as dam operations, water diversions, and irrigation practices, can significantly influence water storage patterns. If these activities are not appropriately represented or accounted for in the models, it can lead to an underestimation of TWS.

7. **Changes in land cover and land use**: Models often struggle to capture changes in land cover and land use, such as deforestation, urbanization, or agricultural practices. These changes can alter the hydrological processes and subsequently impact TWS, which may not be accurately reflected in the models.
8. **Climate change dynamics**: Models may not fully capture the complex interactions between climate change and hydrological processes. Changes in precipitation patterns, temperature, and melting of glaciers and snowpack due to climate change can significantly affect TWS dynamics, potentially leading to an underestimation of TWSC.
9. **Limitations of GRACE data**: While GRACE satellite data provide valuable insights into TWSC, they also have limitations. The spatial resolution of GRACE data is relatively coarse, and they are subject to errors and uncertainties. Comparing GHMs and LSMs directly to GRACE data may introduce discrepancies due to the differences in scale and measurement methods.

It's important to note that the specific causes of differences can vary depending on the specific GHMs, LSMs, and GRACE products being compared. These are general possibilities, and the specific reasons for discrepancies may vary depending on the characteristics and complexities of each river basin and the model used.

4.2 Implications and outlook

Our multimodel seasonal TWSC comparison demonstrates the importance of using independent remote sensing data to evaluate GHMs and LSMs in diverse hydro-climatological settings. Our findings on seasonal assessments of amplitude and phase difference provide future directions for model development, emphasizing the importance of an accurate representation of water stocks and other associated processes. It is important to note that models that include a more precise description of the internal storage dynamics provide a better comparison between simulated TWSC from global models and GRACE data. Comparing TWSC calculated from the balance of precipitation, evaporation, and observed basin outflow against directly computed TWSC variability from satellite observations may assist to find models with improved structures and process representation.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

5 Conclusions

13 models (GHMs, LSMs) were evaluated using different resolutions of Water Resources Reanalysis (WRR1 and WRR2) to compare simulated Total Water Storage Change (TWSC) against GRACE observations over 29 major river basins. Model performance differs significantly across basins, even within the same climatic region. In snow-dominated basins, LSMs generally underestimate the TWSC amplitude and GHMs overestimate. Models and GRACE exhibited inconsistency in the phase with modeled TWSC preceding GRACE with 3-4 months lags. In temperate, arid, and tropical basins GHMs and LSMs generally underestimate the amplitude. However, the modeled TWSC phase is identical to those of GRACE with few exceptions.

Apart from uncertainties associated with GRACE measurements, it provides a standalone means for model assessment. The negative phase differences between models and GRACE might indicate an overall underestimating of the TWS component (e.g., groundwater), leading to an overly rapid system response. The disparity in amplitude and phase could suggest that models are either lacking stores e.g., lakes, and rivers, or the size of the stores is insufficient. There is no single model that performs best in all regions. However, performance statistics reveal that R2 models had a better correlation with GRACE than the coarse resolution R1 models. This demonstrates that optimized model structure can increase their ability to simulate TWS variability and replicate water storage observations. Seasonal TWS variations have received little attention due to a lack of independent data for evaluation. The study provides insight into the amplitude and phase difference between models and GRACE TWSC, which can potentially contribute to further improvement of GHMs and LSMs in the future.

Data availability: GRACE data used in this study can be accessed through these websites, https://podaac.jpl.nasa.gov/dataset/TELLUS_GRACE_MASCON_CRI_GRID_RL06_V1 /http://www2.csr.utexas.edu/grace/RL06_mascons.html. E2O data can be accessed through the E2O Water Cycle Integrator portal (https://wci.earth2observe.eu/). KGClim is publicly available and can be downloaded at https://doi.org/10.5281/zenodo.5347837.

Authors contribution: SB contributed to conceptualization; data curation; formal analysis, visualization, prepared the manuscript with contributions from all co-authors and review & editing. TZ contributed to conceptualization, funding acquisition, project administration, supervision, visualization, and review & editing. AR and BRS contributed to methodology, visualization, and review & editing. MAK contributed to data curation and review & editing. AE, AB and LC in formal analysis and visualization.

Competing interests. The authors declare that they have no conflict of interests.

Acknowledgments: This study was supported by the National Key Research and Development Program of China (2020YFA0608603), and the National Key Research and Development Program of China "Study on simultaneous wet or dry years in the Yangtze River and the Yellow River under changing environment and water allocation in extreme dry years" (No.2022YFC3202300). It was also partially supported by and the National Natural Science Foundation of China (51961125204).

References

Beck, H. E., van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., Miralles, D. G., Martens, B., and de Roo, A.: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data, Hydrol Earth Syst Sci, 21, 589–615, https://doi.org/10.5194/hess-21-589-2017, 2017.

Bierkens, M. F. P.: Global hydrology 2015: State, trends, and directions, Water Resour Res, 51, 4923–4947, https://doi.org/10.1002/2015WR017173, 2015.

Bolaños Chavarría, S., Werner, M., and Salazar, J. F.: Benchmarking global hydrological and land surface models against GRACE in a medium-sized tropical basin, Hydrol Earth Syst Sci, 26, 4323–4344, https://doi.org/10.5194/hess-26-4323-2022, 2022.

Cui, D., Liang, S., Wang, D., and Liu, Z.: A 1 km global dataset of historical (1979–2013) and future (2020–2100) Köppen–Geiger climate classification and bioclimatic variables, Earth Syst Sci Data, 13, 5087–5114, https://doi.org/10.5194/essd-13-5087-2021, 2021.

Dutra, E., Balsamo, G., Calvet, J., Minvielle, M., Eisner, S., Fink, G., Pessenteiner, S., Orth, R., Burke, S., van Dijk, A., et al.: Report on the current state-of-the-art Water Resources Reanalysis, 2015.

Dutra, E., Balsamo, G., Calvet, J., Munier, S., Burke, S., Fink, G., van Dijk, A., Martinez-de la Torre, A., van Beek, R., de Roo, A., et al.: Report on the improved Water Resources Reanalysis (WRR2), EartH2Observe, Report, p. 94, 2017

Eicker, A., Schumacher, M., rgen Kusche, J., Döll, P., Mü ller Schmied, H., Eicker, A., Schumacher Á J Kusche, Á. M., Schumacher, M., Kusche, J., Döll Á M Schmied, P. H., and Schmied, H. M.: Calibration/Data Assimilation Approach for Integrating GRACE Data into the WaterGAP Global Hydrology Model (WGHM) Using an Ensemble Kalman Filter: First Results, 35, 1285–1309, https://doi.org/10.1007/s10712-014-9309-8, 2014.

Emanuel, A., Burke Gabriel Fink Albert van Dijk, S., and Polcher, J.: WP5-Task 5.1-D.5.2 Report on the improved water resources reanalysis Deliverable Title D.5.2-Report on the improved Water Resources Reanalysis Filename E2O_D52.docx, 2017.

Famiglietti, J. S.: Remote sensing of terrestrial water storage, soil moisture and surface waters, 197–207, https://doi.org/10.1029/150GM16, 2004.

Gründemann, G. J., Werner, M., and Veldkamp, T. I. E.: The potential of global reanalysis datasets in identifying flood events in Southern Africa, Hydrol Earth Syst Sci, 22, 4667–4683, https://doi.org/10.5194/hess-22-4667-2018, 2018.

Güntner, A.: Improvement of Global Hydrological Models Using GRACE Data, Surv Geophys, 29, 375–397, https://doi.org/10.1007/s10712-008-9038-y, 2008.

Güntner, A. and Güntner, A.: Improvement of Global Hydrological Models Using GRACE Data, 29, 375–397, https://doi.org/10.1007/s10712-008-9038-y, 2008.

Hassan, A. and Jin, S.: Water storage changes and balances in Africa observed by GRACE and hydrologic models, Geod Geodyn, 7, 39–49, https://doi.org/10.1016/j.geog.2016.03.002, 2016.

Karl E. Taylor: Summarizing multiple aspects of model performance in a single diagram, JOURNAL OF GEOPHYSICAL RESEARCH, , 106, 7183–7192, 2001.

Kim, H., J-F Yeh, P., Oki, T., and Kanae, S.: Role of rivers in the seasonal variations of terrestrial water storage over global basins, https://doi.org/10.1029/2009GL039006, 2009.

van der Knijff, J. M., Younis, J., and de Roo, A. P. J.: LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation, International Journal of Geographical Information Science, 24, 189–212, https://doi.org/10.1080/13658810802549154, 2010.

Hydrology and Earth System Sciences
Discussions

Kraft, B., Jung, M., Körner, M., Koirala, S., and Reichstein, M.: Towards hybrid modeling of the global hydrological cycle, Hydrol Earth Syst Sci, 26, 1579–1614, https://doi.org/10.5194/hess-26-1579-2022, 2022.

Li, B., Rodell, M., Kumar, S., Beaudoing, H. K., Getirana, A., Zaitchik, B. F., Goncalves, L. G., Cossetin, C., Bhanja, S., Mukherjee, A., Tian, S., Tangdamrongsub, N., Long, D., Nanteza, J., Lee, J., Policelli, F., Goni, I. B., Daira, D., Bila, M., Lannoy, G., Mocko, D., Steele-Dunne, S. C., Save, H., and Bettadpur, S.: Global GRACE Data Assimilation for Groundwater and Drought Monitoring: Advances and Challenges, Water Resour Res, 55, 7564–7586, https://doi.org/10.1029/2018WR024618, 2019.

Liesch, T. and Ohmer, M.: Comparison of GRACE data and groundwater levels for the assessment of groundwater depletion in Jordan, Hydrogeol J, 24, 1547–1563, https://doi.org/10.1007/s10040-016-1416-9, 2016.

Lo, M.-H., Famiglietti, J. S., Yeh, J.-F., Syed, T. H., Lo, M.-H., and Famiglietti, J. S.: Click Here for Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data, https://doi.org/10.1029/2009WR007855, 2010.

Milly, P. C. D. and Shmakin, A. B.: Global Modeling of Land Water and Energy Balances. Part I: The Land Dynamics (LaD) Model, J Hydrometeorol, 3, 283–299, https://doi.org/10.1175/1525-7541(2002)003<0283: GMOLWA>2.0.CO;2, 2002.

Pokhrel, Y., Felfelani, F., Satoh, Y., Boulange, J., Burek, P., Gädeke, A., Gerten, D., Gosling, S. N., Grillakis, M., Gudmundsson, L., Hanasaki, N., Kim, H., Koutroulis, A., Liu, J., Papadimitriou, L., Schewe, J., Müller Schmied, H., Stacke, T., Telteu, C.-E., Thiery, W., Veldkamp, T., Zhao, F., and Wada, Y.: Global terrestrial water storage and drought severity under climate change, Nat Clim Chang, 11, 226–233, https://doi.org/10.1038/s41558-020-00972-w, 2021.

Save, H., Bettadpur, S., and Tapley, B. D.: High-resolution CSR GRACE RL05 mascons, J Geophys Res Solid Earth, 121, 7547–7569, https://doi.org/10.1002/2016JB013007, 2016.

Save, H., Bettadpur, S., and Tapley, B. D.: Journal of Geophysical Research: Solid Earth High-resolution CSR GRACE RL05 mascons, https://doi.org/10.1002/2016JB013007, n.d.

Scanlon, B. R., Zhang, Z., Save, H., Sun, A. Y., Schmied, H. M., van Beek, L. P. H., Wiese, D. N., Wada, Y., Long, D., Reedy, R. C., Longuevergne, L., Döll, P., and Bierkens, M. F. P.: Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data, Proc Natl Acad Sci U S A, 115, E1080–E1089, https://doi.org/10.1073/PNAS.1704665115/SUPPL_FILE/PNAS.1704665115.SAPP.PDF, 2018.

Scanlon, B. R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M. H., Müller-Schmied, H., Döll, P., Beek, R., Swenson, S., Lawrence, D., Croteau, M., and Reedy, R. C.: Tracking Seasonal Fluctuations in Land Water Storage Using Global Models and GRACE Satellites, Geophys Res Lett, 46, 5254–5264, https://doi.org/10.1029/2018GL081836, 2019.

Schellekens, J., Dutra, E., Martínez-de la Torre, A., Balsamo, G., van Dijk, A., Sperna Weiland, F., Minvielle, M., Calvet, J.-C., Decharme, B., Eisner, S., Fink, G., Flörke, M., Peßenteiner, S., van Beek, R., Polcher, J., Beck, H., Orth, R., Calton, B., Burke, S., Dorigo, W., and Weedon, G. P.: A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset, Earth Syst Sci Data, 9, 389–413, https://doi.org/10.5194/essd-9-389-2017, 2017.

Sutanudjaja, E. H., van Beek, R., Wanders, N., Wada, Y., Bosmans, J. H. C., Drost, N., van der Ent, R. J., de Graaf, I. E. M., Hoch, J. M., de Jong, K., Karssenberg, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M. W., Vannametee, E., Wisser, D., and Bierkens, M. F. P.: PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model, Geosci Model Dev, 11, 2429–2453, https://doi.org/10.5194/gmd-11-2429-2018, 2018.

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., Sasgen, I., Famiglietti, J. S., Landerer, F. W., Chambers, D. P., Reager, J. T., Gardner, A. S., Save, H., Ivins, E. R., Swenson, S. C., Boening, C., Dahle, C., Wiese, D. N., Dobslaw, H., Tamisiea, M. E., and Velicogna, I.: Contributions of GRACE to understanding climate change, Nat Clim Chang, 9, 358–369, https://doi.org/10.1038/s41558-019-0456-2, 2019.

Trautmann, T., Koirala, S., Carvalhais, N., Eicker, A., Fink, M., Niemann, C., and Jung, M.: Understanding terrestrial water storage variations in northern latitudes across scales, Hydrol Earth Syst Sci, 22, 4061–4082, https://doi.org/10.5194/hess-22-4061-2018, 2018.

Trautmann, T., Koirala, S., Carvalhais, N., Güntner, A., and Jung, M.: The importance of vegetation in understanding terrestrial water storage variations, Hydrol Earth Syst Sci, 26, 1089–1109, https://doi.org/10.5194/hess-26-1089-2022, 2022.

Veldkamp, T. I. E., Zhao, F., Ward, P. J., de Moel, H., Aerts, J. C. J. H., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., Liu, X., Satoh, Y., Gerten, D., Gosling, S. N., Zaherpour, J., and Wada, Y.: Human impact parameterizations in global

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study, Environmental Research Letters, 13, 055008, https://doi.org/10.1088/1748-9326/aab96f, 2018.

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., and Viterbo, P.: The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data, Water Resour Res, 50, 7505–7514, https://doi.org/10.1002/2014WR015638, 2014.

Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM, Hydrol Earth Syst Sci, 14, 59–78, https://doi.org/10.5194/hess-14-59-2010, 2010.

Xiao, R., He, X., Zhang, Y., Ferreira, V., and Chang, L.: Monitoring Groundwater Variations from Satellite Gravimetry and Hydrological Models: A Comparison with in-situ Measurements in the Mid-Atlantic Region of the United States, Remote Sens (Basel), 7, 686–703, https://doi.org/10.3390/rs70100686, 2015.

Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., and Thomas, M.: Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations, Hydrol Earth Syst Sci, 21, 821–837, https://doi.org/10.5194/hess-21-821-2017, 2017.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

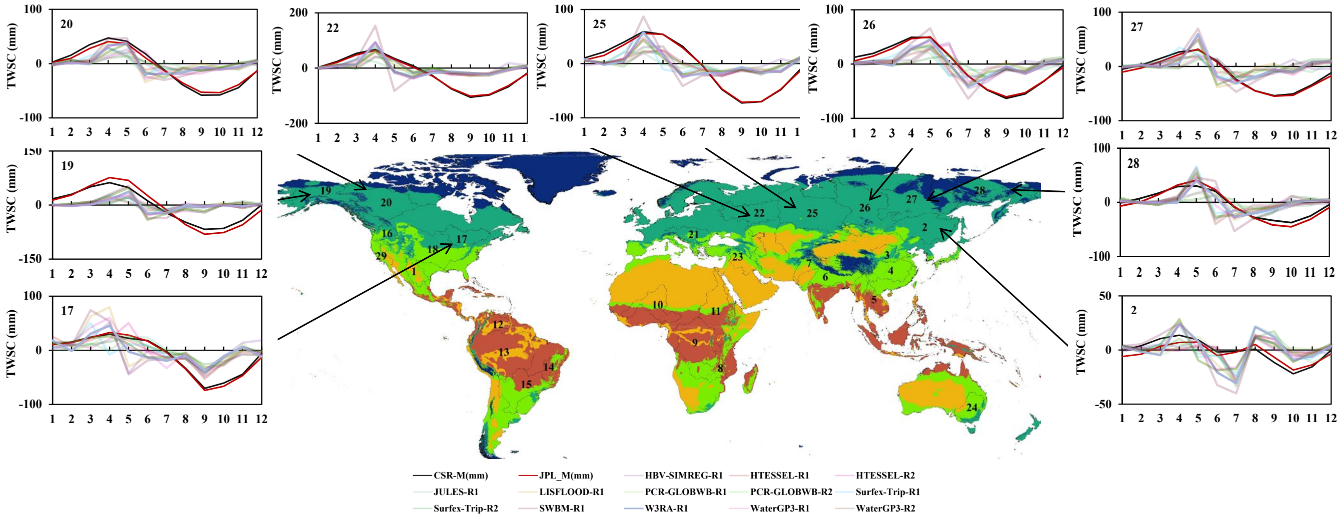


Figure 1: Seasonal TWSC in the Boreal Zones from GRACE, GHMs, and LSMs.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

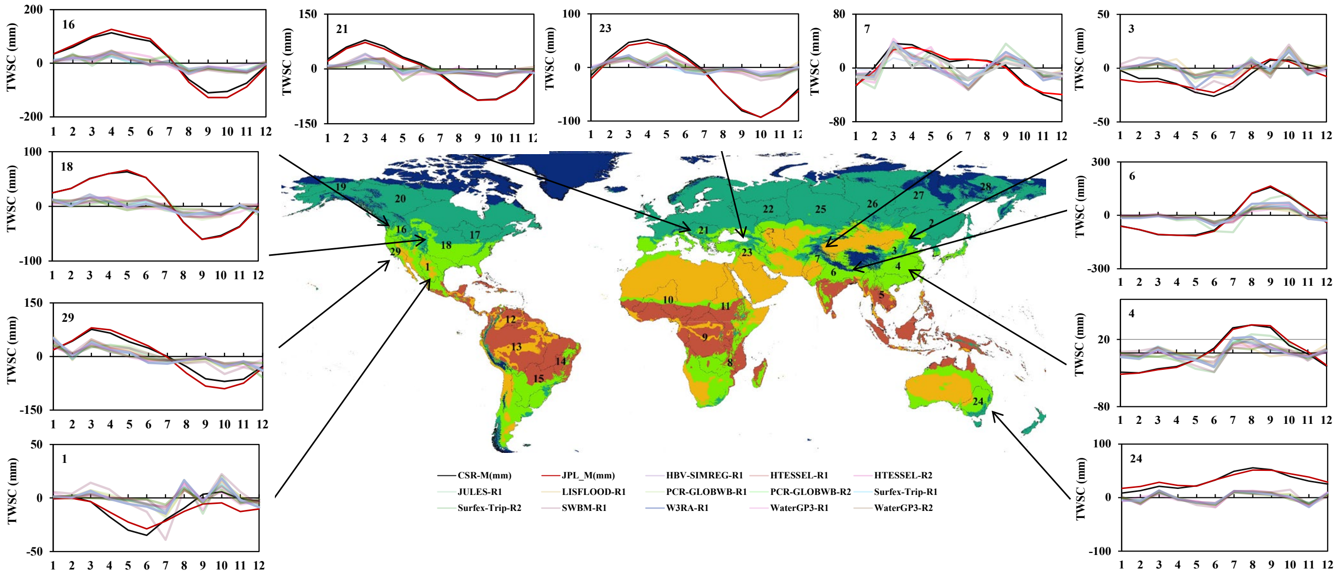


Figure 2: Seasonal TWSC in the Temperate Zones from GRACE, GHMs, and LSMs.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

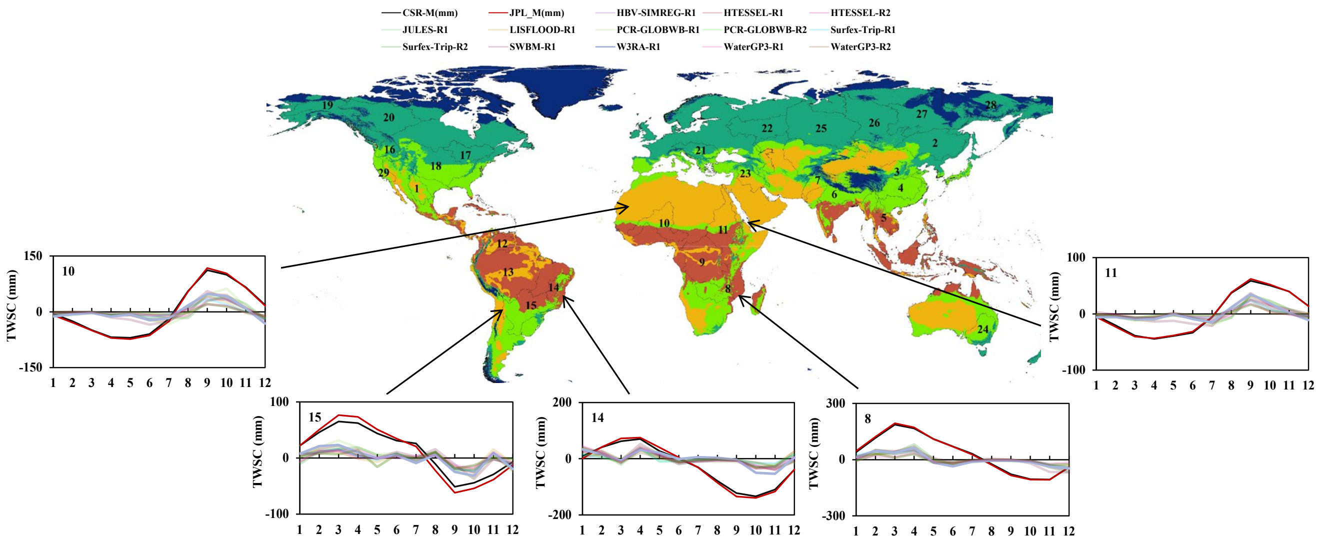


Figure 3: Seasonal TWSC in the Arid Zones from GRACE, GHMs, and LSMs.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

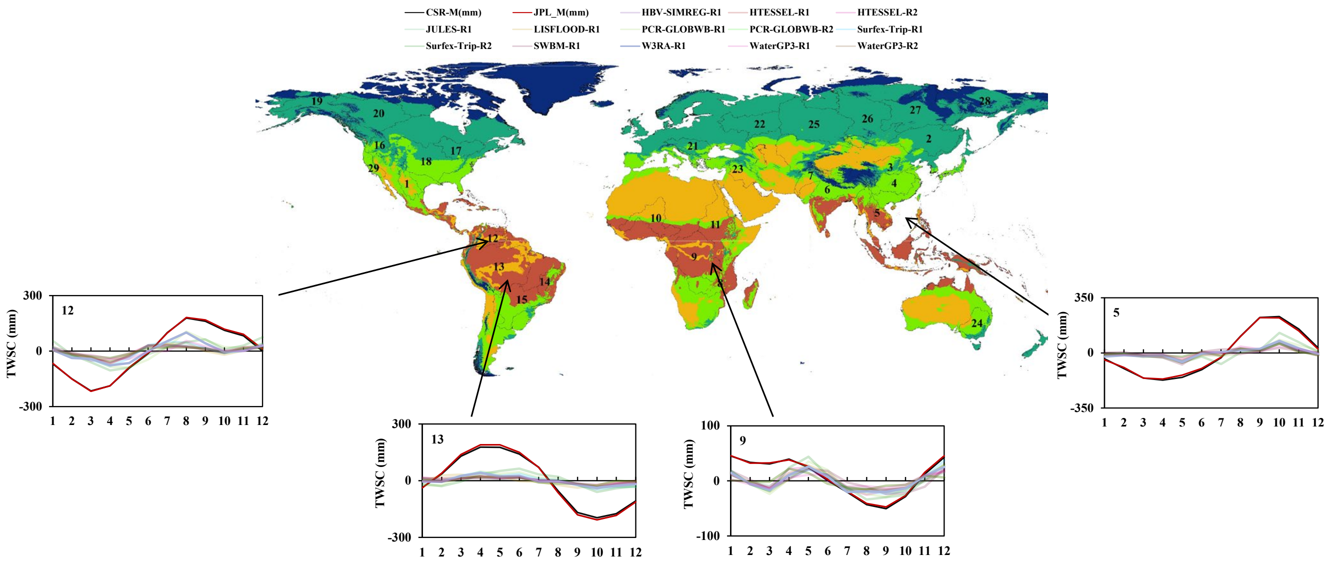


Figure 4: Seasonal TWSC in the Tropical Zones from GRACE, GHMs, and LSMs.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

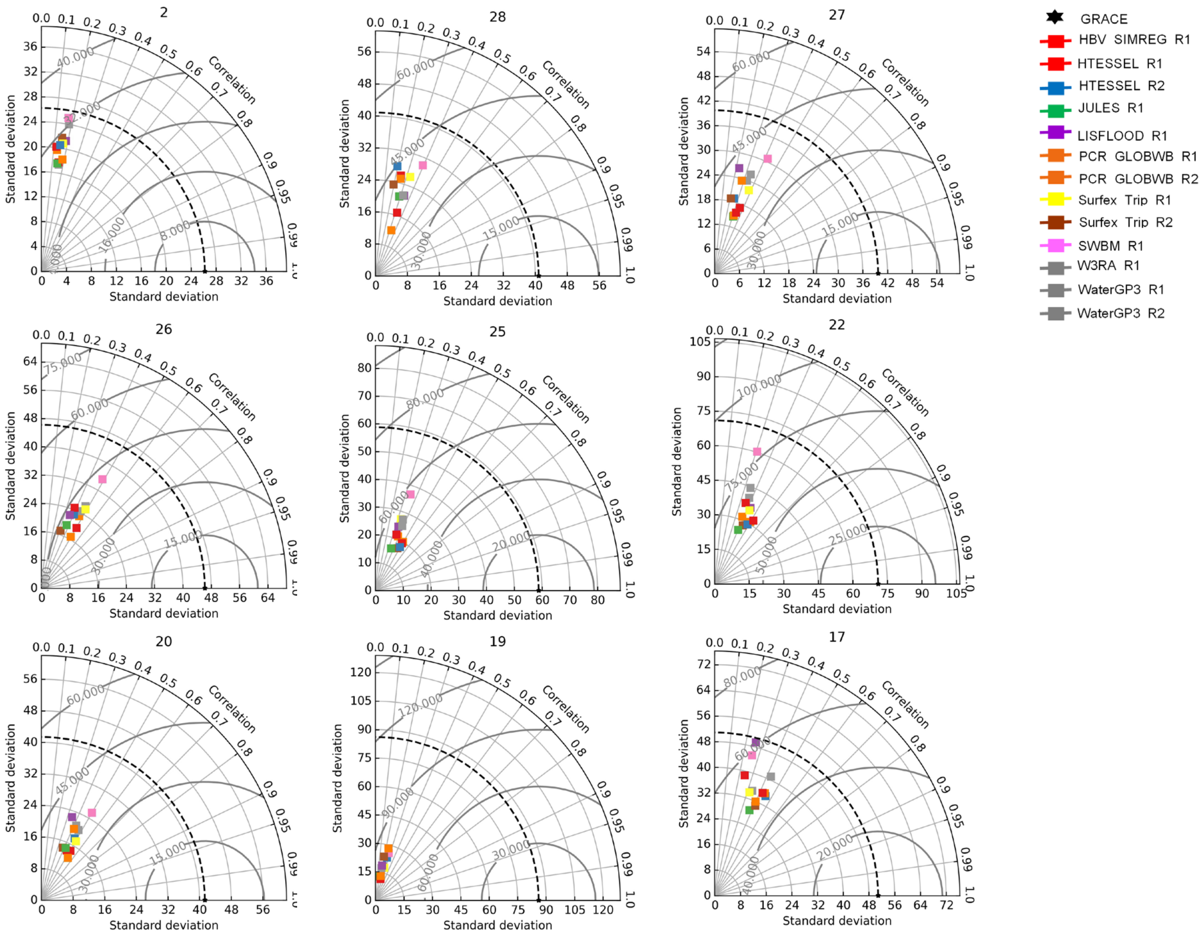


Figure 5: Taylor diagrams between GRACE observations and each model output in the Boreal Zones.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Open Access
Discussions

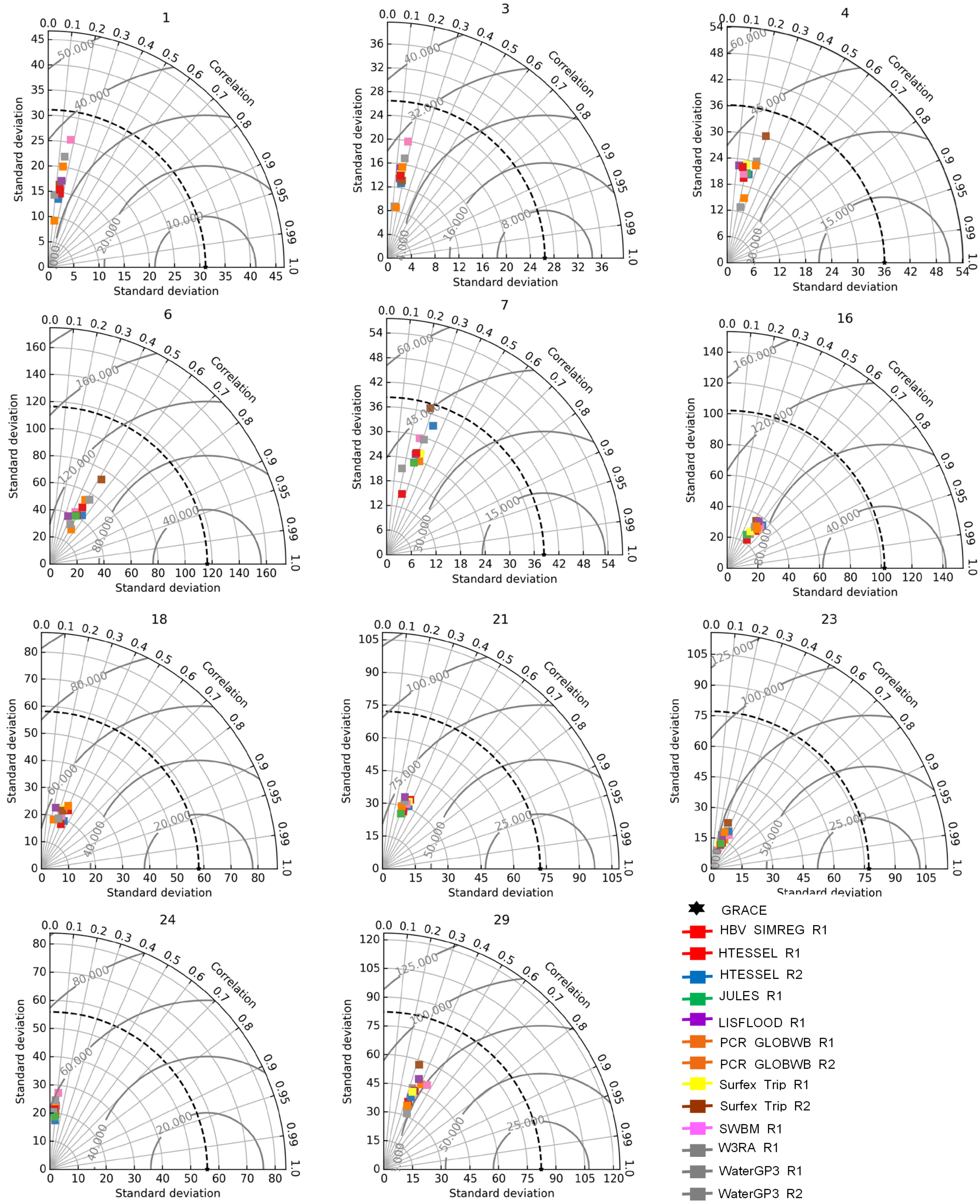


Figure 6: Taylor diagrams between GRACE observations and each model output in the Temperate Zone.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

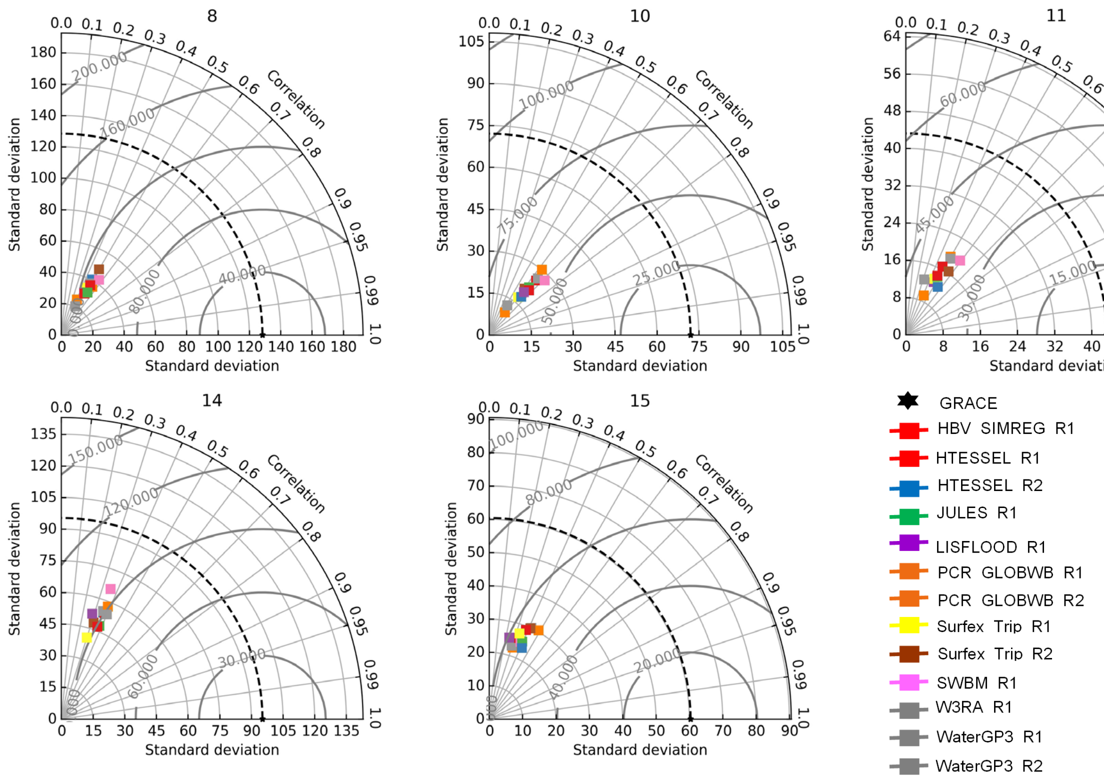


Figure 7: Taylor diagrams between GRACE observations and each model output in the Arid Zones.

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

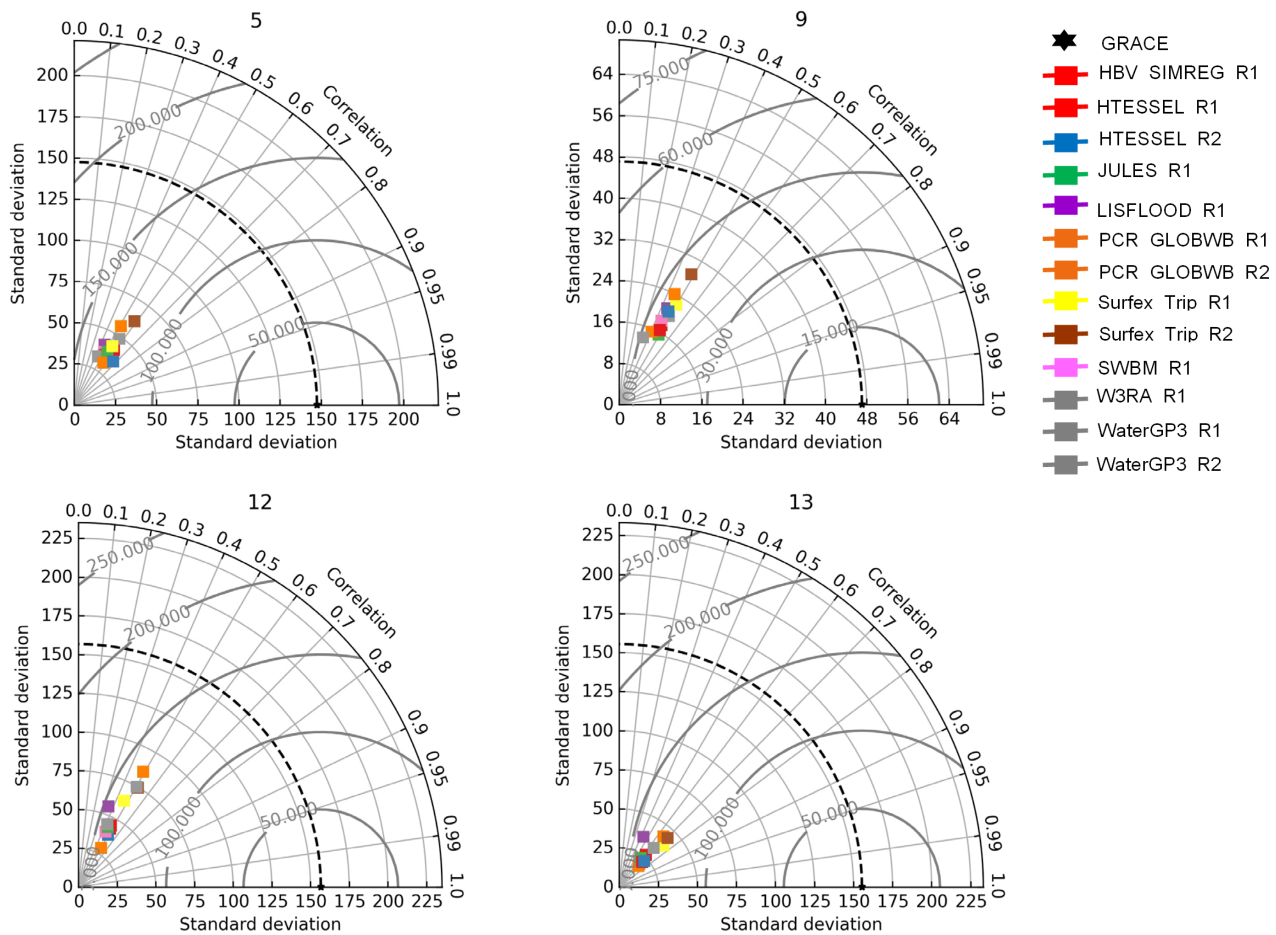


Figure 8: Taylor diagrams between GRACE observations and each model output in the Tropical Zones

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

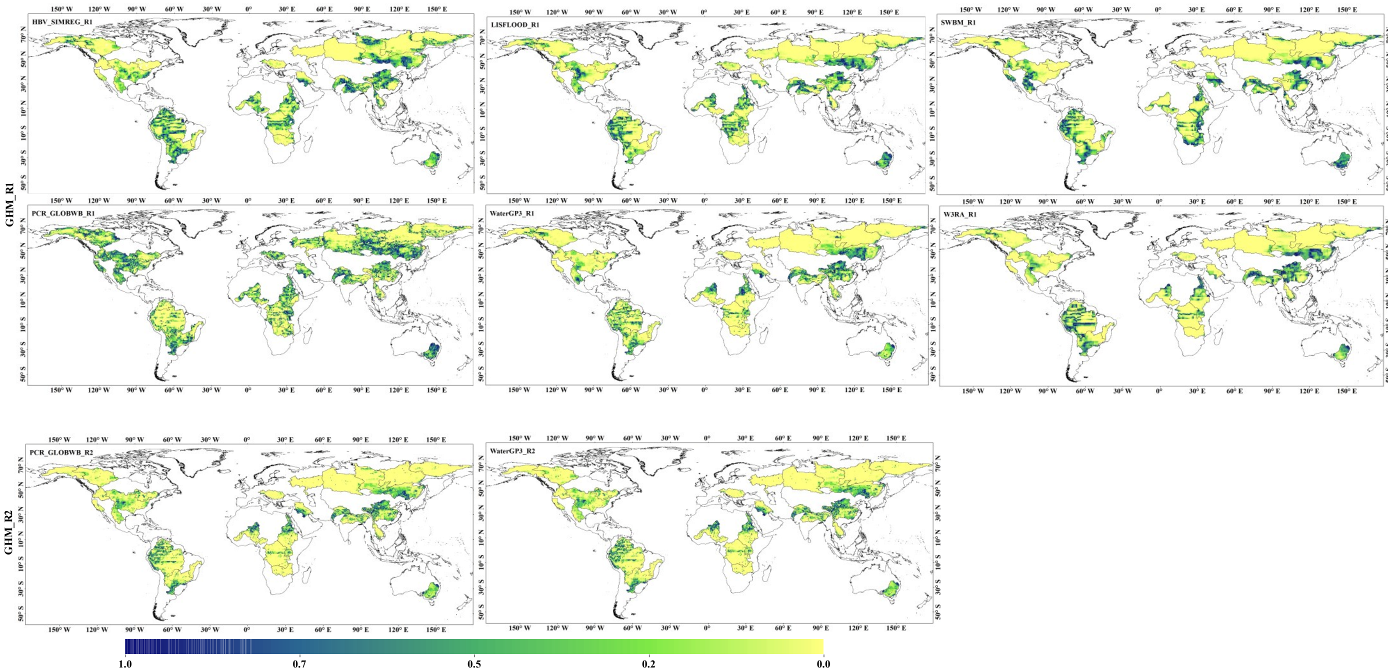


Figure 9: Spatial correlation coefficient between GRACE and GHM (R1 and R2)

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

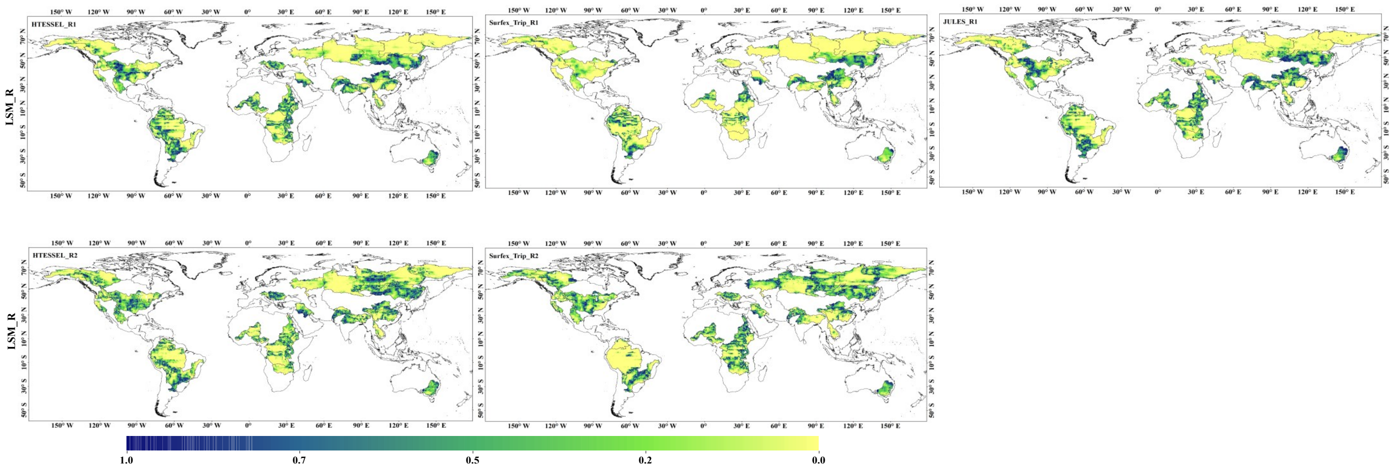


Figure 10: Spatial correlation coefficient between GRACE and LSM (R1 and R2).

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Table 1: Summary of the length, drainage area, and outflow of the selected river basins

Basin ID	River	Length in km	Drainage area in km^2	Outflow
1	Rio Grande	3,057	570,000	Gulf of Mexico
2	Amur	4,444	1,855,000	Sea of Okhotsk
3	Yellow River	5,464	745,000	Bohai Sea
4	Yangtze	6,300	1,800,000	East China Sea
5	Mekong	4,350	810,000	South China Sea
6	Brahmaputra-Ganga	3,969	1,320,000	Bay of Bengal
7	Indus	3,610	960,000	Arabian Sea
8	Zambezi	2,740	1,330,000	Mozambique Channel
9	Congo	4,700	3,680,000	Atlantic Ocean
10	Niger	4,200	2,090,000	Gulf of Guinea
11	Nile	6,650	3,254,555	Mediterranean
12	Orinoco	2,250	990,000	Atlantic Ocean
13	Amazon	6,400	7,000,000	Atlantic Ocean
14	São Francisco	3,180	610,000	Atlantic Ocean
15	Parana	4,880	2,582,672	Río de la Plata
16	Columbia	2,000	668,000	Pacific Ocean
17	Saint Lawrence	3,058	1,030,000	Gulf of Saint Lawrence
18	Mississippi	6,275	2,980,000	Gulf of Mexico
19	Yukon	3,185	328187	Bering Sea
20	Mackenzie	4,241	1,790,000	Beaufort Sea
21	Danube	2,888	817,000	Black Sea
22	Volga	3,645	1,380,000	Caspian Sea
23	Euphrates	3,596	884,000	Persian Gulf
24	Murray-Darling	3,672	1,061,000	Southern Ocean
25	Ob	5,410	2,990,000	Gulf of Ob
26	Yenisei	5,539	2,580,000	Kara Sea
27	Lena	4,400	2,490,000	Laptev Sea
28	Kolyma	2,129	647,000	East Siberian Sea
29	California	1,220		

https://doi.org/10.5194/hess-2023-163
Preprint. Discussion started: 25 July 2023

Hydrology and Earth System Sciences
Discussions

Table 2: Summary of Pearson's r of models respect to GRACE data.

		GHMs								LSMs				
	Basin ID	HBV-SIMREG-R1	LISFLOOD-R1	PCR-GLOWBWB-R1	PCR-GLOWBWB-R2	SWBM-R1	W3RA-R1	WaterGP3-R1	WaterGP3-R2	HTESSEL-R1	HTESSEL-R2	JULES-R1	Surfex-Trip-R1	Surfex-Trip-R2
Boreal	2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.2	0.2
	28	0.3	0.3	0.3	0.2	0.4	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.2
	27	0.4	0.2	0.3	0.3	0.4	0.3	0.3	0.3	0.3	0.2	0.3	0.4	0.2
	26	0.4	0.4	0.5	0.5	0.5	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.3
	25	0.3	0.3	0.5	0.4	0.3	0.4	0.4	0.4	0.5	0.5	0.3	0.3	0.5
	22	0.4	0.4	0.5	0.4	0.3	0.4	0.3	0.4	0.5	0.5	0.4	0.4	0.4
	20	0.5	0.3	0.5	0.4	0.5	0.5	0.4	0.4	0.5	0.5	0.4	0.5	0.4
	19	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2	0.3	0.2	0.3	0.2
	17	0.2	0.3	0.4	0.4	0.3	0.3	0.4	0.4	0.4	0.5	0.4	0.3	0.4
Temperate	1	0.2	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1
	3	0.2	0.1	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2
	4	0.2	0.1	0.3	0.3	0.2	0.3	0.2	0.3	0.2	0.2	0.2	0.2	0.3
	6	0.5	0.4	0.5	0.5	0.4	0.5	0.4	0.5	0.5	0.6	0.5	0.5	0.5
	7	0.2	0.3	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.3	0.3	0.3	0.3
	16	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.5	0.5	0.5
	18	0.4	0.2	0.4	0.2	0.4	0.3	0.3	0.2	0.4	0.4	0.3	0.3	0.3
	21	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.3	0.4	0.3
	23	0.4	0.3	0.4	0.3	0.5	0.3	0.3	0.3	0.4	0.4	0.4	0.3	0.3
	24	0.2	0.3	0.3	0.2	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.2	0.3
	29	0.3	0.4	0.4	0.3	0.5	0.3	0.4	0.3	0.4	0.4	0.4	0.3	0.3
Arid	8	0.5	0.5	0.5	0.4	0.6	0.5	0.4	0.4	0.5	0.5	0.5	0.5	0.5
	10	0.7	0.6	0.6	0.6	0.7	0.6	0.5	0.6	0.7	0.6	0.6	0.6	0.6
	11	0.5	0.5	0.5	0.4	0.6	0.5	0.3	0.4	0.5	0.6	0.5	0.5	0.6
	14	0.4	0.3	0.4	0.3	0.4	0.4	0.4	0.3	0.4	0.4	0.4	0.3	0.3
	15	0.3	0.2	0.5	0.3	0.4	0.4	0.3	0.3	0.4	0.4	0.4	0.3	0.4
Tropical	5	0.5	0.4	0.5	0.6	0.5	0.6	0.4	0.6	0.6	0.7	0.5	0.5	0.6
	9	0.5	0.4	0.4	0.4	0.4	0.5	0.3	0.4	0.5	0.5	0.5	0.5	0.5
	12	0.5	0.3	0.5	0.5	0.4	0.5	0.4	0.5	0.5	0.5	0.4	0.5	0.5
	13	0.7	0.4	0.7	0.7	0.6	0.7	0.7	0.7	0.6	0.7	0.6	0.7	0.7