LETTER • OPEN ACCESS

Calibrating global hydrological models with GRACE TWS: does river storage matter?

To cite this article: Tina Trautmann *et al* 2023 *Environ. Res. Commun.* **5** 081005

Environmental Research Communications

LETTER

Calibrating global hydrological models with GRACE TWS: does river storage matter?

Tina Trautmann[1,2,*], **Sujan Koirala**[1], **Andreas Guentner**[3,4], **Hyungjun Kim**[5,6,7] and **Martin Jung**[1]

[1] Department of Biogeochemical Integration, Max-Planck Institute for Biogeochemistry, -D07745 Jena, Germany
[2] Institute of Physical Geography, Goethe University Frankfurt, -D60438 Frankfurt am Main, Germany
[3] Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, -D14473 Potsdam, Germany
[4] Institute of Environmental Sciences and Geography, University of Potsdam, -D14476 Potsdam, Germany
[5] Moon Soul Graduation School of Future Strategy, Korea Advanced Institute of Science and Technology, 34141 Daejeon, Republic of Korea
[6] Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology, 34141 Daejeon, Republic of Korea
[7] Institute of Industrial Science, The University of Tokyo, 153-8505, Tokyo, Japan
[*] Author to whom any correspondence should be addressed.

E-mail: ttraut@bgc-jena.mpg.de

Keywords: GRACE, terrestrial water storage variations, river water, river routing, global hydrologic model, model calibration, model validation

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
24 November 2022

REVISED
20 April 2023

ACCEPTED FOR PUBLICATION
2 August 2023

PUBLISHED
24 August 2023

Abstract

Although river water storage contributes to Total Terrestrial Water Storage (TWS) variations obtained from GRACE satellite gravimetry, it is unclear if computationally expensive river routing schemes are required when GRACE data is used for calibration and validation in global hydrological modeling studies. Here, we investigate the role of river water storage on calibration and validation of a parsimonious global hydrological model. In a multi-criteria calibration approach, the model is constrained against either GRACE TWS or TWS from which river water storage is removed. While we find that removing river water storage changes the TWS constraint regionally and globally, there are no significant implications for model calibration and the resulting simulations. However, adding modeled river water storage a-posteriori to calibrated TWS simulations improves model validation against seasonal GRACE TWS variations globally and regionally, especially in tropics and Northern low- and wetlands. While our findings justify the exclusion of explicit river routing for global model calibration, we find that the inclusion of river water storage is relevant for model evaluation.

1. Introduction

Over the last decade, terrestrial water storage variations from GRACE and GRACE-FO satellite gravimetry provided valuable information for calibration and validation of global hydrological modeling approaches (GHMs) (Werth *et al* 2009, Döll *et al* 2014, Kumar *et al* 2016, Scanlon *et al* 2016, Mostafaie *et al* 2018, Trautmann *et al* 2018). However, satellite gravimetry measures the vertically integrated total water storage (TWS), that includes water stored in ice, snow, canopy, soil moisture, groundwater, but also in wetlands, surface water bodies and river channels (Watkins *et al* 2015). GHMs, on the contrary, do not necessarily simulate all these storages, and also vary significantly in their complexity and the represented hydrologic processes (Schellekens *et al* 2017, Telteu *et al* 2021). Among others, their inability to correctly simulate GRACE TWS variations is often attributed to neglected processes, such as river and floodplain storage dynamics (Kim *et al* 2009). Getirana *et al* 2017 showed that river and surface water storages contribute to 8% of TWS variability globally, those storages are especially relevant at regional scale. Significant contribution of river water storages to TWS variations have been shown for the tropics, the monsoon-impacted sub-tropics, major river basins in arid regions, and in high

latitudes where the increase in river water due to snow melt is relevant (Felfelani *et al* 2017, Getirana *et al* 2017). While not all GHMs include a river water storage (Telteu *et al* 2021), a river routing scheme is the essential component to simulate lateral water flow and thus enable calibration of model parameters against river discharge observations, as it is traditionally done in hydrology.

However, the increasing availability and quality of global Earth observation data over the past decades, not only of GRACE TWS but also of other water storages and fluxes, allows to constrain model parameters not only against the single river discharge constraint, but against multiple, complementary observational data streams. In this context, multi-criteria calibration approaches evolve (Trautmann *et al* 2022, Dembélé *et al* 2020, Sirisena *et al* 2020, Mostafaie *et al* 2018), in which including river routing during model calibration might be dispensable to define parameter values. Furthermore, the variety of global observational data enables more data-driven approaches of hydrological modeling, that either use machine-learning techniques (Mosaffa *et al* 2022, Shen and Lawson 2021, Xu and Liang 2021) or combine these with process-based modeling knowledge in hybrid models (Kraft *et al* 2020, Reichstein *et al* 2022). Both such approaches require thousands of thousands of model runs during the calibration process and the parameter testing in the learning phase. Hence, it is computational not feasible, or even methodological possible, to perform a full global model run, including the lateral routing of discharge, in each iteration. While many efforts to reduce the computational costs of global routing schemes exist (Yamazaki *et al* 2013, Mizukami *et al* 2021), they remain a considerable factor for time and computational performance when compared to a model run without river routing. Additionally, the spatial context must be preserved to simulate the lateral water flow in routing schemes, what hinders the possibilities for spatial sub-sampling of grid cells for calibration and the parallelization of computational processes.

At the same time, the actual relevance of river routing for parameter calibration and validation against GRACE TWS at a global scale is rather unclear, also given the broad spatial and temporal resolution of GRACE TWS.

In the context of the development of new, largely data-driven modeling schemes and enhanced model calibration approaches against multiple observational data streams , it's essential to know whether computational resources need to be invested in river routing during parameter calibration, and what the consequences are if routing is only applied as a post-processing, i.e. after defining model parameters, to validate model simulations.

Therefore, we specifically investigate the need for consideration of river storage and its potential effect on model calibration and validation in global hydrological studies that apply GRACE TWS.

To do so, we use a parsimonious hydrological model that does not explicitly account for river dynamics. We constrain the model in a multi-criteria calibration approach either against original GRACE TWS estimates, or against TWS estimates from which river storage was removed, and compare the resulting simulations. In the second step, we apply a routing scheme on the calibrated model and validate the performance with and without additional consideration of river storage compared to the original GRACE TWS. Specifically, we focus on:

I. the sensitivity of model calibration and resulting hydrological simulations to the removal of river storage from GRACE TWS.

II. the need of river routing for validation of hydrological simulations with GRACE TWS observations at regional and global scales.

In the following, we provide an overview on the methodology of this study. In section 3, we present and discuss the results regarding the effect of river storage on model calibration, followed by its influence on validation against GRACE TWS. Finally, section 4 summarizes the implications for future global hydrological studies.

2. Data and methods

This study consists of 2 parts:

I) the effect of river storage on model calibration, and

II) the effect of river storage on model validation.

An overview on the methodologies and data for both parts is given in figure 1, and described in detail in sections 2.2 and 2.3. For all analysis, we apply the same hydrologic model, which is introduced in the following section.

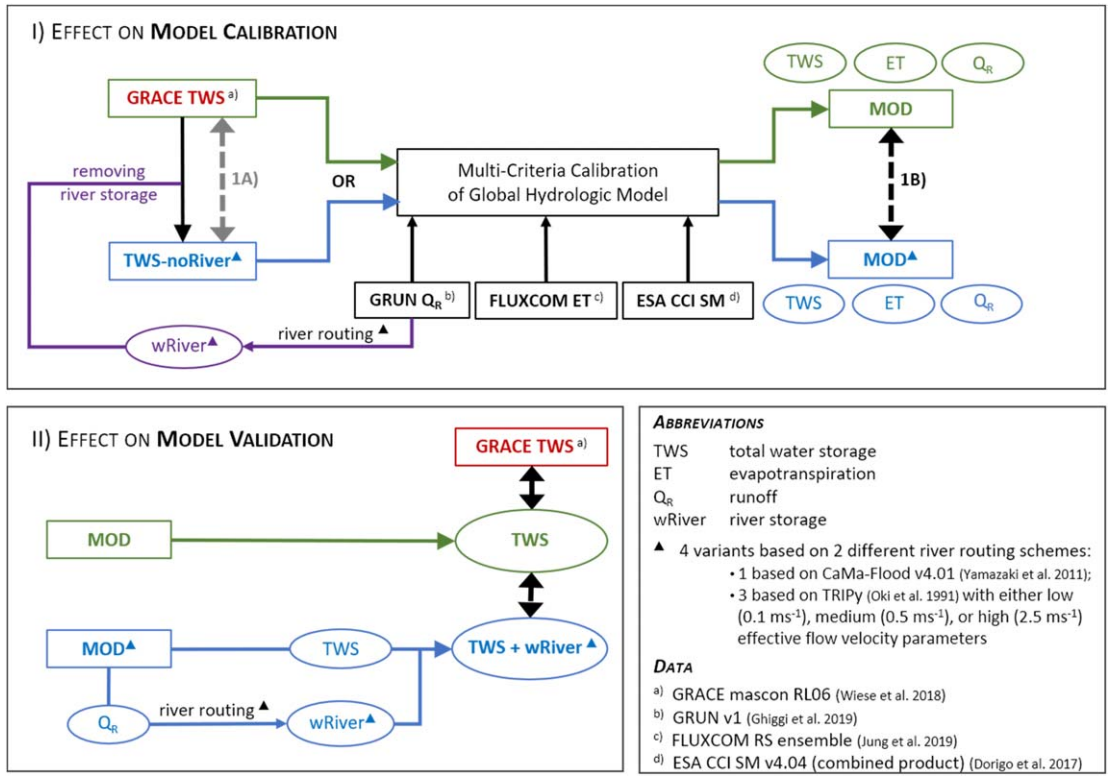

Figure 1. Schematic structure of the applied data and methodologies to assess the effect of river storage on model (I) calibration, and (II) validation. In (I) the TRIPy and the CaMa-Flood routing schemes are forced by GRUN Q_R to estimate wRiver, which is then removed from GRACE TWS. Either GRACE TWS or TWS-noRiver data is then used along with other observational data to calibrate a global hydrologic model. In the first step (1A) we compare GRACE TWS against TWS-noRiver to identify the wRiver signal in GRACE TWS data (section 3.1.1), and in the second step (1B) simulations from different calibrations are compared against each other to assess the effect of wRiver on model calibration (section 3.1.2). In II) we force the TRIPy and the CaMa-Flood routing scheme with modeled Q_R from I) and add it to modeled TWS to assess the effect of accounting for wRiver on model validation against the original GRACE TWS wRiver.

While the model is run for the time period 03/2000 to 12/2014, model calibration and validation are limited to the availability of observational data, such as GRACE TWS. Thus model calibration and validation are conducted for the period 2002 to 2014. Although all calibration and analysis considers this entire time series, the shown results focus on seasonal variations because of the rather short time period and the inability of the model to represent trends in TWS due to, e.g., groundwater depletion and melting of glaciers that hinder the analysis of trends.

2.1. Hydrological model

Exemplary for the variety of global hydrological models, we apply the conceptual hydrologic model introduced in Trautmann *et al* (2022). While being more parsimonious than its established counterparts, its structure reflects classical process representation of GHMs and the calibrated model achieves equally good and partially better performance as, e.g. models from the EartH2Observe model ensemble (Schellekens *et al* 2017), regarding different observational data (figure S5).

Forced by precipitation, air temperature and net radiation, the model simulates evapotranspiration (ET) and runoff (Q_R), and considers 4 water storages: a snow component, a 2-layer soil water storage, a delayed water storage component, and a deep soil water storage that interacts with the soil and delayed storage components. Simulated total water storage (TWS) is the sum of these 4 storages. While groundwater, surface water and river water are not implemented explicitly, they are assumed to be effectively included in the deep and slow storage components after calibration of associated model parameters against GRACE TWS.

We run the model on a $1° \times 1°$ latitude/longitude spatial resolution on daily time steps for the period 03/2000 to 12/2014, focusing on vegetated land area under near-natural conditions.

For regional analysis, we consider hydro-climatic regions obtained from cluster analysis of latitude, mean seasonal dynamics and amplitudes of TWS, ET and Q_R observational data (figure 2).

Further details are available in Trautmann *et al* (2022).

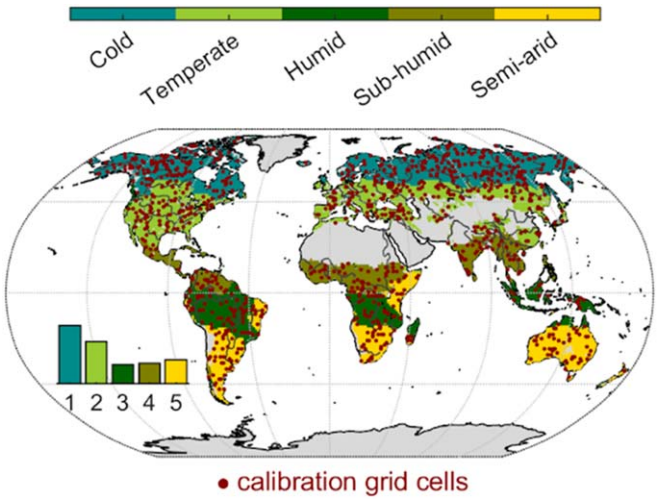

Figure 2. Hydro-climatic clusters of the study area with red dots indicating grid cells used in model calibration.

2.2. Effect of river storage on model calibration

To assess the effect of river water storage (wRiver) included in GRACE TWS estimates on model calibration, we constrain the model against monthly GRACE TWS variations of the JPL mascon solution (RLM06v2; Wiese *et al* 2018), from which estimates of wRiver were removed or not. To do so, we first estimate river storage variations, and then calibrate the hydrological model against GRACE data with or without river storage.

To estimate wRiver variations, we force spatially-explicit river routing schemes with observation-based runoff Q_R reconstructions from GRUN v1 (Ghiggi *et al* 2019). To do so, the monthly average gridded Q_R estimates are resampled to daily modeling time steps by replicating the monthly value.

Since the choice of the river routing scheme essentially affects simulated wRiver and discharge (Q_{Dis}) (Zhao *et al* 2017) we consider 2 different river routing schemes: 1) the simple routing scheme TRIPy (Oki *et al* 1999), and 2) the more sophisticated, widely used Catchment-based Macro-scale Floodplain (CaMa-Flood) river routing scheme (Yamazaki *et al* 2011).

TRIPy calculates Q_{Dis} from each grid cell along the river network based on Q_R and maps of flow direction and river sequence, using a linear reservoir algorithm. Thereby, the parameter effective flow velocity (*eff_vel*) [ms^{-1}] defines how fast Q_R is discharged from one grid cell to the next, i.e. how long water is stored in the grid cell's wRiver. While in reality, flow velocity varies spatially as it depends on land surface characteristics such as slope, TRIPy uses a globally uniform value for simplicity. To yet assess the sensitivity to *eff_vel* in TRIPy, we consider a range of global *eff_vel* values in different experiments, from low (0.1 ms^{-1}) to medium (0.5 ms^{-1}) to high (2.5 ms^{-1}) values, to derive corresponding estimates of wRiver that produce a range of river storage dynamics with large and fast variability for high *eff_vel*, and small and slow variability for low *eff_vel*.

Similar to TRIPy, CaMa-Flood v4.01 derives the time evolution of water storage from the water balance equation, considering the inflow from upstream grid cells, the input from runoff forcing generated at the respective grid cell, and the outflow to downstream grid cells. However, in contrast to TRIPy, it, next to Q_{Dis}, explicitly calculates flow velocity along a prescribed river network that is automatically generated with the Flexible Location of Waterways (FLOW) (Yamazaki *et al* 2009) method. Utilizing a parametrization based on the sub-grid topography obtained from HydroSHEDS, CaMa-Flood simulates water storage within the river channel, but also water storage in flood plains. Thus, CaMa-Flood allows a more dynamic simulation of Q_{Dis} and wRiver while considering the spatial variability of discharge-generating characteristics.

In the following, the experiments with river routing from CaMa-Flood and the 3 experiments from TRIPy are summarized with ▲, or denoted with *CaMa, 01*, *05*, or *25* if referred to explicitly.

Estimates of wRiver▲ are removed from GRACE TWS to obtain new estimates of TWS-noRiver▲. Before using either GRACE TWS or TWS-noRiver▲ for model calibration, we compare them to assess the contribution of wRiver to the TWS constraint regionally and globally (figure 1(1A); section 3.1.1). For this purpose, we calculate the Nash-Sutcliffe Efficiency (MEF, equation (1)) between GRACE TWS and each TWS-noRiver▲ to quantify their similarity for different spatial scales.

$$MEF = 1 - \frac{\sum_{i=1}^{n}(x_{obs,i} - x_{\mathrm{mod},i})^2}{\sum_{i=1}^{n}(x_{obs,i} - \bar{x}_{obs,i})^2} \quad (1)$$

where $x_{mod,i}$ corresponds to TWS-noRiver$^{\blacktriangle}$, $x_{obs,i}$ corresponds to GRACE TWS, and $\bar{x}_{obs}$ is the average of GRACE TWS at each data point i.

Either GRACE TWS or TWS-noRiver$^{\blacktriangle}$ estimates are then used along with other observational data including GRUN Q_R, FLUXCOM ET (Jung *et al* 2019), and ESA CCI soil moisture (Dorigo *et al* 2017) to constrain model parameters in a multi-criteria calibration approach. The approach (Text S1), described in Trautmann *et al* 2022, aims to derive the globally best performing parameter set regarding all constraints simultaneously, while considering each data stream's strengths and uncertainties. For each observational constraint we calculate a cost metric that is summed up to a total cost value which is optimized (minimized) using the CMAES algorithm (Hansen and Kern 2004) to derive the globally best performing parameter set. We perform calibration for a spatial subset of grid cells that is obtained by stratified random sampling among Koeppen-Geiger zones. The calibration subset mirrors the global and regional distribution of observed TWS, ET, Q_R, and wRiver, and therefore allows for efficient calibration of parameter values that are globally applicable. To appraise parameter equifinality and uncertainties of the optimization procedure, the calibration of each experiment is performed 10 times independently, with each of the 10 calibration runs comprising up to 10000 model runs, to derive 10 optimal parameter sets.

Finally, we analyze and compare the simulations that were calibrated against GRACE TWS (MOD) and those calibrated against the 4 different TWS-noRiver$^{\blacktriangle}$ estimates regarding TWS, ET and Q_R (figure 1(1B)) for global and regional mean seasonal dynamics. While taking into account the spread between 10 different calibration runs of each experiment, we focus on the best performing calibration run when calculating MEF between MOD (x_{obs} in equation (1)) and each MOD-R$^{\blacktriangle}$ (x_{mod} in equation (1)).

2.3. Effect on model validation

To assess the relevance of wRiver for validation of TWS simulations against GRACE TWS, we add wRiver to model simulations after model calibration. For this purpose, we apply the CaMa-Flood and the TRIPy routing scheme for each calibrated MOD-R$^{\blacktriangle}$, i.e., using the calibrated Q_R as forcing for the routing schemes. Thereby, we use the same routing scheme and parametrization as for the TWS-noRiver$^{\blacktriangle}$ constraint that was used for the respective calibration (e.g. TRIPy with *eff_vel* of 0.1 ms^{-1} for MOD-R01 that was calibrated against TWS-noRiver-01; and CaMa-Flood for MOD-CaMa that was calibtated against TWS-noRiver-CaMa).

We then add simulated wRiver to TWS of each MOD-R$^{\blacktriangle}$ and compare it against the TWS simulations from MOD without additional wRiver, as well as against the original GRACE TWS. For model validation, we calculate MEF between GRACE TWS and MOD resp. MOD-R$^{\blacktriangle}$ at the local grid-cell scale, as well as for global and regionally aggregated mean seasonal dynamics.

3. Results and discussion

3.1. Effect of river storage on model calibration

3.1.1. Effect of river storage on TWS patterns

Figure 3 compares the decrease in similarity between the original GRACE TWS and estimates of TWS from which wRiver is removed.

As expected, low *eff_vel* increases wRiver, and therefore, the lowest correspondence with the original GRACE TWS can be seen for TWS-noRiver-01, while the similarity increases with increasing *eff_vel*, so that TWS-noRiver-25 is nearly identical with the original GRACE TWS, because the high *eff_vel* causes the immediate depletion of wRiver to the next downstream grid cell and prevents the accumulation of wRiver except for grid cells in downstream areas of large catchments. Spatially, differences between GRACE TWS and TWS-noRiver-CaMa are similar to those from TRIPy with a medium *eff_flow* (TWS_noRiver-05).

Overall, spatial correspondence with GRACE TWS mainly decreases along larger river networks with a large water accumulation (figure 3(a)). The largest grid-wise changes are obtained in the *Cold* and *Humid* regions, where river networks are dense and streamflow and, thus, wRiver is large in absolute terms. On the contrary, removing wRiver from GRACE TWS has little effect in the *Semi-arid* region, where water accumulation is smaller than in humid regions.

Regarding seasonal dynamics (figure 3(b)), removing wRiver mainly changes the amplitude of TWS variations, which increases in snow affected regions and tends to decrease otherwise. In the *Cold* region, gradual snow melt, retarded infiltration due to shallow and frozen soils, slow discharge to downstream areas, as well as additional water input from upstream areas in large river networks dampen TWS variations. Removing wRiver

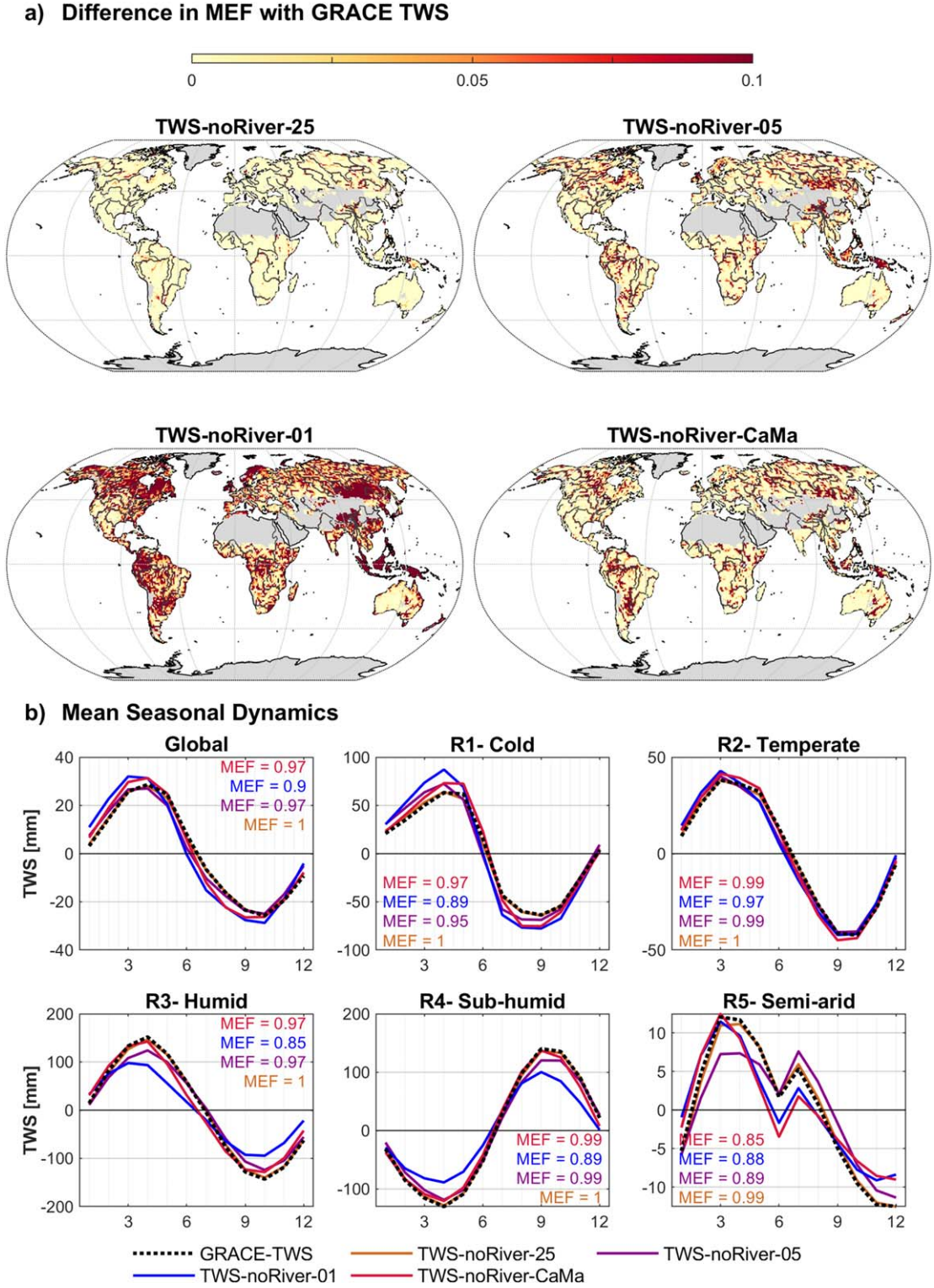

Figure 3. Comparison of the different TWS estimates (GRACE TWS and TWS-noRiver$^{\blacktriangle}$) that are used for model calibration in different experiments. (a) similarity between the original GRACE TWS and TWS-noRiver$^{\blacktriangle}$ in terms of gridwise Nash-Sutcliff efficiency (MEF); (b) global and regional mean seasonal dynamics of GRACE TWS and TWS-noRiver$^{\blacktriangle}$. For (a) and (b) MEF of TWS-noRiver$^{\blacktriangle}$ with the original GRACE TWS is calculated and subtracted from the optimal MEF of 1. Note the different ranges on the *y*-axis for TWS variations in panel (b).

attenuates these delaying effects, causing increased TWS variations and shifts the TWS peak to one month earlier. While in the *Cold* region seasonal TWS variations are mainly affected by snow accumulation and melt, TWS in other regions is dominated by liquid water storages (Trautmann *et al* 2018). In non-snow affected regions, Q_R increases with wetness, i.e. with TWS, which in turn reduces TWS. Due to this negative feedback, removing wRiver from TWS reduces the TWS amplitude. Due to the spatial variability of parametrization and

flow velocity in CaMa-Flood, TWS-noRiver-CaMa agrees with different TRIPy TWS-noRiver estimates in different regions, e.g. rather with high *eff_vel* parametrization in *Temperate*, *Humid* and *Sub-humid* regions, but with low *eff_vel* parametrization in the *Semi-arid* region.

3.1.2. Effect of river storage on model calibration

Figure 4 shows the mean seasonal dynamics of TWS, ET and Q_R simulated by MOD and MOD-R$^▲$ after model calibration. Respective observations are plotted for better evaluation of the calibration results, yet the following focuses on differences between MOD and MOD-R$^▲$. A detailed evaluation of performance of MOD against TWS, ET and Q_R observations can be found in Trautmann *et al* (2022).

Despite being calibrated against either GRACE TWS or TWS-noRiver$^▲$, overall little variance between calibrated MOD and MOD-R$^▲$ is evident. While we expected the model parameters to adapt to differences in the TWS constraint, the mean seasonal TWS simulations are nearly identical among experiments, globally and regionally. The same holds for ET. Some differences between calibration runs are obtained regarding Q_R, especially for *Cold*, *Temperate* and *Semi-arid* regions. However, while the spread is larger than for the other simulated variables, the differences of the best performing calibration runs are still negligible.

High agreement in simulated fluxes and TWS variations goes along with no systematic differences between calibrated parameter values of MOD and different MOD-R$^▲$. Hence, we do not find any evidence for biases between experiments that result from either considering wRiver or not in model calibration. While the calibrated parameter values don't vary significantly between different experiments, different calibration runs point to two parameter sets that achieve (nearly) equal good performance (figure S1). Affected are parameters that regulate the size of soil water storage and it's depletion by ET. The interplay of these parameters with the other outgoing flux Q_R also explains the spread of simulated Q_R. This parameter equifinality is not related to wRiver, but to the equifinality of baseflow and ET decay, especially under water limitation as discussed in Trautmann *et al* (2022).

The absence of a qualitative effect of using GRACE TWS or TWS-noRiver$^▲$ for model calibration is also evident when inspecting the total costs and cost components among the experiments (figure S1). They are fairly similar on average, while the spread of costs between different calibration runs of one experiment tends to be larger than the differences between experiments. This underlines that uncertainties arise mainly from other aspects of hydrological modeling than from the effect of wRiver included in the TWS constraint. However, we do see a tendency for elevated TWS costs for MOD-R01 and MOD-R05. Higher TWS costs and thus total costs suggest difficulties of the model to adapt to the TWS constraint, when the removed wRiver is large. This suggests that the comparatively large removal of wRiver is harder to reconcile with the other constraints of ET and Q_R, and thus may indicate that such large wRiver based on low *eff_vel* is not plausible. While it is notable that the spread in total, TWS, and Q_R costs of MOD-CaMa is comparatively smaller and it achieves the overall lowest total costs due to low soil moisture and ET costs, the general difference to the other calibrated models remains small, indicating no significant impact of the chosen routing scheme on the global multi-criteria calibration.

3.2. Effect of river storage on global model validation

While we did not find systematic differences between TWS simulations of MOD and MOD-R$^▲$ after calibration, explicitly accounting for wRiver and adding it to TWS of MOD-R$^▲$ causes systematic differences of model performance against the original GRACE TWS, locally as well as for regional and global mean seasonal dynamics (figure 5). The largest differences are notable for considering wRiver based on low *eff_vel*, while the largest improvement of model performance relative to the original GRACE TWS is achieved when adding wRiver based on medium *eff_vel* - globally and in most regions (figure 5(b)). While the choice of the routing scheme does not significantly affect model performance regarding seasonal dynamics of TWS for large spatial regions (figure 5(b)), it is relevant for simulating hydrological dynamics at smaller, e.g. catchment, scale. As such, Q_{Dis} from MOD-CaMa provides consistently good estimates of Q_{Dis} at various GRDC stations, while different MOD-R$^▲$ from TRIPy perform better for different stations, highlighting the benefit of spatially distributed river flow parameters as opposed to global average parameter values of *eff_vel* in TRIPy (figure S6). However, locally, the differences in model performance with GRACE TWS are less pronounced when wRiver from CaMa-Flood is added (MOD-CaMa). Overall, including wRiver improves TWS simulations at local scale especially in the tropics and Northern low- and wetlands where rivers accumulate water over large catchments (figure 5(a)). The importance of wRiver in the tropics has already been shown by previous studies (Getirana *et al* 2017). Similarly, the inability to reproduce observed TWS variations by models in the *Cold* region is among others attributed to missing representations of floodplain and river flow processes (Kim *et al* 2009). Indeed, in the *Cold* region, accounting for wRiver improves MEF for the majority of grid cells (figure 5(a)), highlighting the importance of additional inflow from upstream grid cells and the delay of water outflow in these regions. While MOD-R01 matches the timing of TWS variations slightly better, it underestimates the seasonal TWS amplitude

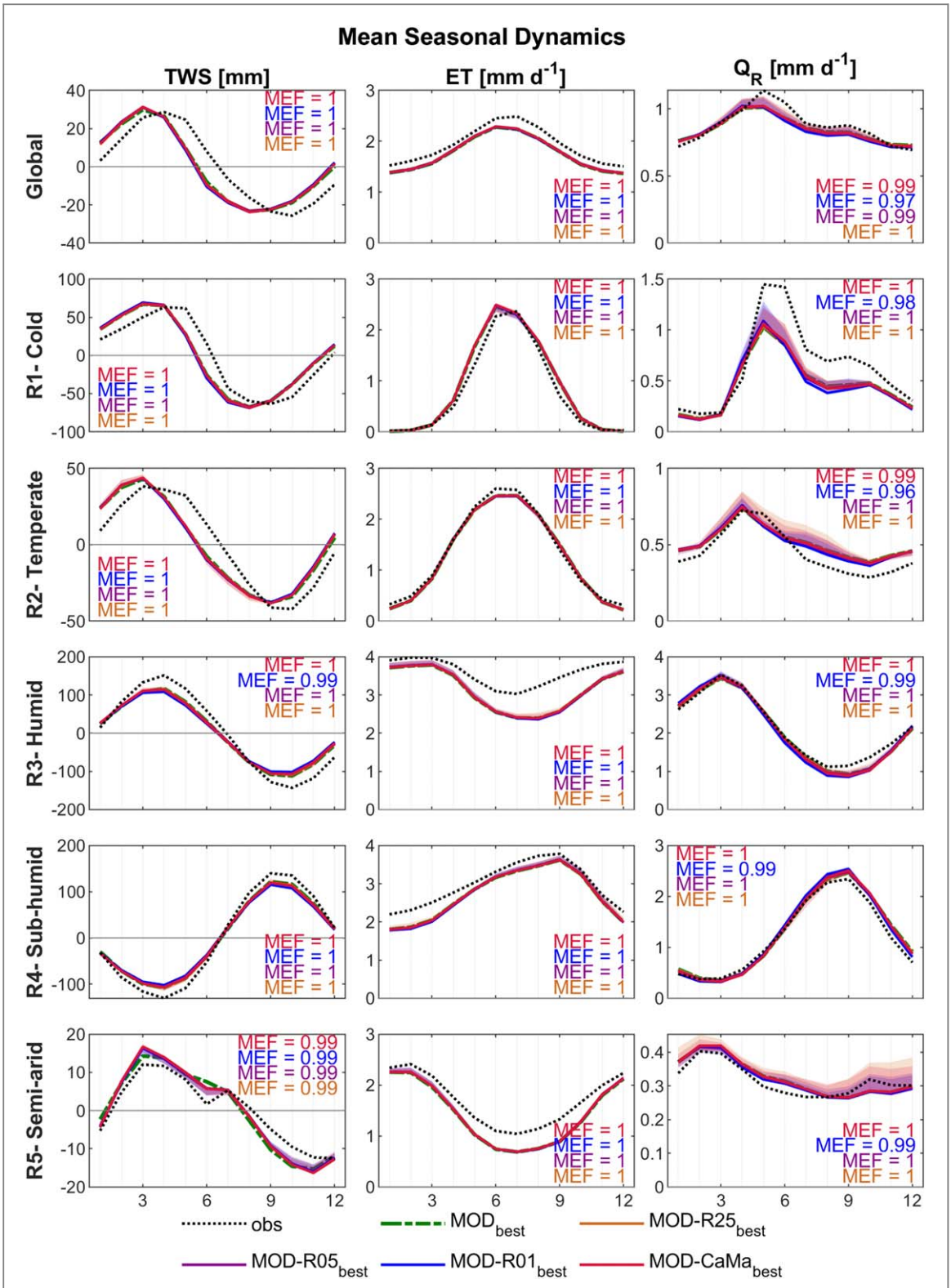

Figure 4. Mean seasonal dynamics of TWS, ET, and Q_R observations and calibrated simulations, summarized globally and for different regions. Uncertainty bands denote the spread between 10 different calibration runs of each experiment, while dashed/solid lines indicate the best performing calibration run of each experiment. Black dotted lines denote observational GRACE TWS, FLUXCOM ET, and GRUN Q_R. Listed MEF compares the best performing calibration run MOD▲ against the best performing calibration run of MOD (dashed line).

(figure 5(b)). On the contrary, MOD-R05 and MOD-CaMa reproduce TWS amplitude, yet still precede TWS variations, although not as much as MOD or MOD-R25. Hence, the phase-shift issue of simulated TWS in *Cold* regions, which is prevalent in many GHMs (Schellekens *et al* 2017), is unlikely to arise from unaccounted river storage variations only. The underestimation of TWS amplitude (and peak spring discharge, figure S6)) can also be affected by deficiencies in the precipitation forcing (Huffman *et al* 2000, Contractor *et al* 2020), but remaining

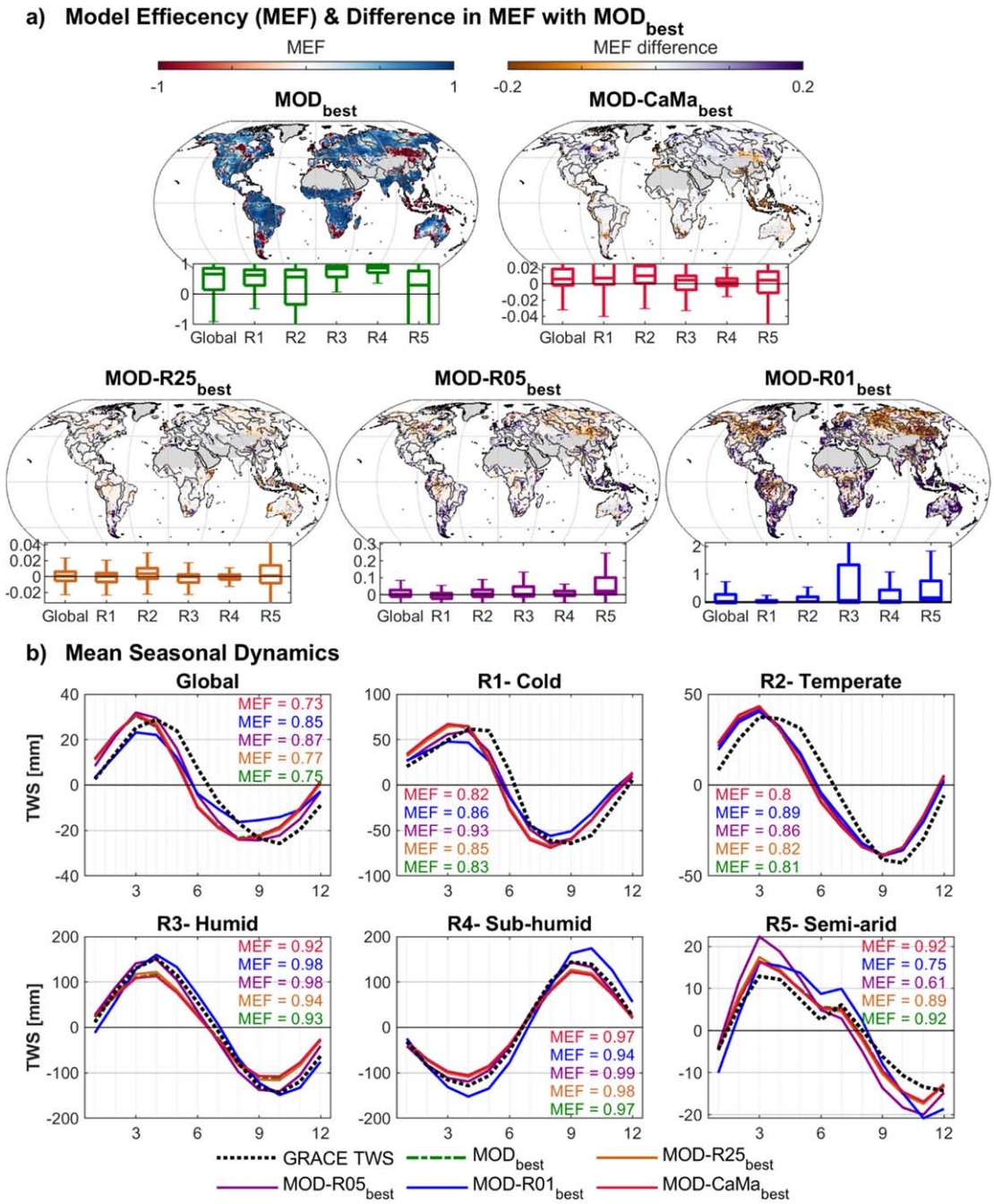

Figure 5. Comparison of the original GRACE TWS, TWS simulated by MOD (without river water storage) and TWS+wRiver simulated by MOD▲ (accounting for river water storage based on different routing schemes). (a) gridwise Nash Sutcliff efficiency (MEF) between GRACE TWS and MOD (upper left map), and differences of MEF obtained by MOD and MEF obtained by MOD-CaMa, MOD-R01, MOD-R05, MOD-R25, respectively; including their global and regional distributions (outlier are not plotted). (b) mean seasonal dynamics averaged globally and for different regions. Listed MEF compares MOD resp. MOD▲ with the original GRACE TWS. Note: In MODbest, individual storage components contribute with 0.22% (snow), 0.17% (soil), 0.4% (deep soil) resp. 0.20% (slow storage) to the global average seasonal TWS variations (Trautmann *et al* 2022).

difficulties in reproducing the timing of TWS dynamics indicate the potential importance of other missing processes such as freeze/thaw dynamics and permafrost (Yu *et al* 2020), and ice jam in river channels (Kim *et al* 2009).

The preceding of simulated TWS compared to GRACE TWS can also slightly be reduced at global scale, when wRiver is considered (figure 5(b)). However, a preceding of simulated TWS variations is still apparent for large areas, especially the *Temperate* region (figure 5(b)), indicating again the relevance of other processes than water delay in wRiver, such as irrigation, land cover changes and interactions between groundwater and surface water dynamics.

While including wRiver improves agreement with GRACE TWS over large areas, MEF decreases notably in the *Semi-arid* region when the TRIPy routing scheme is used (figure 5). The slightly better performance of

MOD-CaMa in semi-arid regions locally (figure 5(a)) as well as for the average seasonal dynamics in the *Semi-arid* region (figure 5(b)) indicates the benefits of the variable flow velocity and the accounting of flooding in CaMa-Flood, which is relevant in such regions that are characterized by rain- and dry-seasons. However, especially the regional average dynamics are not notably different from the other simulations by TRIPy and the local improvement remains low. Already MOD does not agree well with GRACE TWS in the *Semi-arid* regions (figure 5(a)), where TWS variability is sensitive to parameters that control ET under water limited conditions, which are poorly constrained (Trautmann *et al* 2022). In addition to the parameter equifinality issues, poor (initial) performance points to model structure uncertainties and deficiencies, such as missing processes of evaporation and percolation to groundwater from open water surfaces. Besides, the GRUN Q_R constraint is known for larger uncertainties in arid regions (Ghiggi *et al* 2019) and tends to inconsistencies with the other observation-based data in the *Semi-arid* region (Trautmann *et al* 2022). Therefore, comparatively good MEF of modeled and observed Q_R (figure 4) does not necessarily reflect good representation of Q_R, which is also underlined by poor representation of observed discharge at semi-arid GRDC stations (figure S6). As modeled Q_R is used to derive modeled wRiver, this further leads to poorer model performance in semi-arid regions when wRiver is added to TWS.

Besides, the general improvement of TWS when wRiver is added to modeled TWS, in all regions except for the semi-arid regions, highlights potential room for improvements in our approach of modeling the hydrology in semi-arid regions. Among others, the improved model performance by using CaMa-Flood in such regions indicates especially the importance of seasonal flooding and associated processes for TWS variability in semi-arid regions.

3.3. Transferability of results to other global hydrological modeling studies

While we cannot exclude that the findings shown in the previous sections are conditional on the specific model structure, calibration approach and data used in this study, we argue that our findings are of general relevance for global hydrological modeling studies across a spectrum of GHMs and data-driven approaches. The model used in this study is based on classic hydrologic process representation and despite its simple structure achieves good performance that is comparable to more complex state-of-the-art GHMs (figure S5). The identified key issues of model-data mismatches in cold and semi-arid regions are also shared among GHMs in general (Schellekens *et al* 2017) and appear unrelated to river storage variations. The identified problem of parameter and thus process identifiability is due to insufficient observational constraints, and this problem is expected to be even larger for more complex models with more parameters and for data-driven approaches that include machine learning methods and thus even more rely on the available data.

4. Implications

In this study, we showed that river storage has a relevant impact on seasonal TWS dynamics, in particular in cold and humid regions, and accounting for it when simulating TWS improves performance against GRACE TWS observations (section 3.2). However, we did not find a systematic impact of either including or excluding river storage in TWS for global-scale model calibration (section 3.1). Compared to river storage, restrictions from other observational constraints seem to be more relevant to define model parameters in such a model-data integration approach as presented here. For example, the main discrepancies between observed and modeled TWS may be related to missing processes other than the river water storage (especially in cold and semi-arid regions, see section 3.2), and issues of parameter identifiability due to insufficient and partly conflicting data constraints. Our findings hold across sensitivity experiments with different routing schemes and effective flow velocity parameters that produce a wide range of river storage dynamics. While we do not argue that river routing is of relevance at local and smaller regional scales, our findings show that the impact at larger regional up to global scales vanishes. Especially when using GRACE TWS for model calibration, the effect of small rivers and less dense river networks is likely smoothed out by its 250 km native resolution (Wiese *et al* 2018). As discussed, our findings are of general relevance for other global hydrological modeling approaches as well (section 3.3). Especially in the Era of Earth observations, they are of particular relevance for future global hydrological modeling approaches that follow a more data-driven perspective and such approaches in which model parameters are calibrated not only against a single river discharge constraint, but against multiple large-scale observational data streams of complementary water fluxes and storages. Based on the presented findings, it seems advisable to save the comparatively large computational resources needed by routing schemes when model parameters are calibrated against GRACE TWS and other Earth observation based data. Omitting the routing itself, but more significantly the resulting ability to subsample grid cells for calibration instead of demanding a full global simulation in each calibration iteration, as well as the easier implementation of parallel programming strategies can reduce the computational costs of global model calibration. The saved

computational resources can then be better used to address other issues, such as parameter and process identifiability, more efficiently.

Overall, we highly recommend considering river water storage when evaluating and analyzing modeled TWS, especially on regional scale. However, we suggest there is no need to apply river routing in global model calibration against GRACE TWS, especially if complementary observational data streams are considered or when a data-driven approach, with high computational demand is used.

Acknowledgments

Part of this research was financially supported by the Internship Support Program of UTokyo Institute of Industrial Science. This research was further supported by the National Research Foundation of Korea (NRF) grant Funded by the Korea Government (MSIT) (2021H1D3A2A03097768).

All experiments of this study, including the model implementation and calibration, were performed within the SINDBAD model data integration framework developed at the Max-Planck Institute for Biogeochemistry Jena.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.8085841. Data will be available from 20 June 2023 (Trautmann 2023).

Code and data availability statement

For this study we use monthly terrestrial water storage variations from the JPL GRACE Mascon RL06 Coastal Resolution Improvement Filtered version 1.0 (Watkins *et al* 2015, Wiese *et al* 2018), available at http://doi.org/10.5067/TEMSC-3MJC6. Average monthly runoff from the GRUN dataset (Ghiggi *et al* 2019) is publicly available and can be downloaded at http://doi.org/10.6084/m9.figshare.9228176. For model calibration we additionally used evapotranspiration from the FLUXCOM RS ensemble v1 (Jung *et al* 2019) and the combined ESA CCI soil moisture v4.04 (Dorigo *et al* 2017). The hydrologic model is forced by GPCP 1dd v1.2 precipitation (Huffman *et al* (2000)), CRU-NCEP v7 air temperature (Viovy *et al* 2018), and CERES Ed4 A net radiation (Loeb *et al* 2018).

The used CaMa-Flood river routing scheme v4.01 is freely available via http://hydro.iis.u-tokyo.ac.jp/~yamadai/cama-flood/.

The EartH2Observe model ensemble data (Schellekens *et al* 2017) is available via the Water Cycle Integrator portal (WCI, wci.earth2observe.eu) and http://doi.org/10.1016/10.5281/zenodo.167070.

Discharge measurements are provided by The Global Runoff Data Centre, 56068 Koblenz, Germany.

ORCID iDs

Tina Trautmann https://orcid.org/0000-0002-4692-5071
Hyungjun Kim https://orcid.org/0000-0003-1083-8416

References

Contractor S, Donat M G, Alexander L V, Ziese M, Meyer-Christoffer A, Schneider U, Rustemeier E, Becker A, Durre I and Vose R S 2020 Rainfall estimates on a gridded network (REGEN)—a global land-based gridded dataset of daily precipitation from 1950 to 2016 *Hydrol. Earth Syst. Sci.* **24** 919–43

Dembélé M, Hrachowitz M, Savenije H H G, Mariéthoz G and Schaefli B 2020 Improving the predictive skill of a distributed hydrological model by calibration on spatial patterns with multiple satellite data sets *Water Resour. Res.* **56** e2019WR026085

Döll P, Fritsche M, Eicker A and Müller Schmied H 2014 Seasonal water storage variations as impacted by water abstractions: comparing the output of a global hydrological model with GRACE and GPS observations *Surv. Geophys.* **35** 1311–31

Dorigo W *et al* 2017 ESA CCI soil moisture for improved earth system understanding: State-of-the art and future directions *Remote Sens. Environ.* **203** 185–215

Felfelani F, Wada Y, Longuevergne L and Pokhrel Y N 2017 Natural and human-induced terrestrial water storage change: a global analysis using hydrological models and GRACE *J. Hydrol.* **553** 105–18

Getirana A, Kumar S, Girotto M and Rodell M 2017 Rivers and floodplains as key components of global terrestrial water storage variability *Geophys. Res. Lett.* **44** 359–10

Ghiggi G, Humphrey V, Seneviratne S I and Gudmundsson L 2019 GRUN: an observation-based global gridded runoff dataset from 1902 to 2014, Earth System Science *Data* **11** 1655–167

Hansen N and Kern S 2004 Evaluating the CMA evolution strategy on multimodal test functions *Parallel Problem Solving from Nature - PPSN VIII* ed X Yao *et al* (Springer) vol 3242 (https://doi.org/10.1007/978-3-540-30217-9_29)
Huffman G J, Adler R, Morrissey M M, Bolvin D, Curtis S, Joyce R, McGavock B and Susskind J 2000 Global precipitation at one-degree resolution from multisatellite observations *Journal of Hydrometeorology* **2** 36–50
Jung M, Koirala S, Weber U, Ichii K, Gans F, Camps-Valls G, Papale D, Schwalm C, Tramontana G and Reichstein M 2019 The FLUXCOM ensemble of global land-atmosphere energy fluxes *Scientific data* **6** 1–14
Kim H, Yeh P J F, Oki T and Kanae S 2009 Role of rivers in the seasonal variations of terrestrial water storage over global basins *Geophys. Res. Lett.* 36
Kraft B, Jung M, Körner M and Reichstein M 2020 Hybrid modeling: fusion of a deep learning approach and a physics-based model for global hydrological modeling *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLIII-B2-2020** 1537–44
Kumar S V *et al* 2016 Assimilation of gridded GRACE terrestrial water storage estimates in the north american land data assimilation system *Journal of Hydrometeorology* **17** 1951–72
Loeb N G, Doelling D R, Wang H, Su W, Nguyen C, Corbett J G, Liang L, Mitrescu C, Rose F G and Kato S 2018 Clouds and the earth's radiant energy system (CERES) energy balanced and filled (EBAF) top-of atmosphere (TOA) edition-4.0 data product *J. Climate* **31** 895–918
Mizukami N, Clark M P, Gharari S, Kluzek E, Pan M, Lin P, Beck H E and Yamazaki D 2021 A vector-based river routing model for earth system models: parallelization and global applications *Journal of Advances in Modeling Earth Systems* **13** e2020MS002434
Mosaffa H, Sadeghi M, Mallakpour I, Naghdyzadegan Jahromi M and Pourghasemi H R 2022 Chapter 43 - Application of machine learning algorithms in hydrology *Computers in Earth and Environmental Sciences* 585–91
Mostafaie A, Forootan E, Safari A and Schumacher M 2018 Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using *in situ* runoff and daily GRACE data *Computational Geosciences* **22** 789–814
Oki T, Nishimura T and Dirmeyer P 1999 Assessment of annual runoff from land surface models using Total Runoff Integrating Pathways (TRIP) *Journal of the Meteorological Society of Japan. Ser. II* **77** 235–55
Reichstein M, Ahrens B, Kraft B, Camps-Valls G, Carvalhais N, Gans F, Gentine P and Winkler A J 2022 Combining system modeling and machine learning into hybrid ecosystem modeling *Knowledge Guided Machine Learning* (Chapman and Hall/CRC)
Scanlon B R, Zhang Z Z, Save H, Wiese D N, Landerer F W, Long D, Longuevergne L and Chen J l 2016 Global evaluation of new GRACE mascon products for hydrologic applications *Water Resour. Res.* **52** 9412–29
Schellekens J *et al* 2017 A global water resources ensemble of hydrological models: the eartH2Observe Tier-1 dataset *Earth System Science Data* **9** 389–413
Shen C and Lawson K 2021 Applications of deep learning in hydrology *Deep Learning for the Earth Sciences* (John Wiley & Sons, Ltd) pp 283–97
Sirisena T A J G, Maskey S and Ranasinghe R 2020 Hydrological model calibration with streamflow and remote sensing based evapotranspiration data in a data poor basin *Remote Sens.* **12** 3768
Telteu C E *et al* 2021 Understanding each other's models: an introduction and a standard representation of 16 global water models to support intercomparison, improvement, and communication, Geosci *Model Dev.* **14** 3843–78
Trautmann T, Koirala S, Carvalhais N, Eicker A, Fink M, Niemann C and Jung M 2018 Understanding terrestrial water storage variations in northern latitudes across scales *Hydrol. Earth Syst. Sci.* **22** 4061–82
Trautmann T, Koirala S, Carvalhais N, Güntner A and Jung M 2022 The importance of vegetation in understanding terrestrial water storage variations *Hydrol. Earth Syst. Sci.* **26** 1089–109
Trautmann T 2023 Scripts for Trautmann et al. 2023v1 (v1.1) *Zenodo* (https://doi.org/10.5281/zenodo.8085840)
Viovy N 2018 CRUNCEP Version 7 - atmospheric forcing data for the community land model *Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory.*
Watkins M M, Wiese D N, Yuan D-N, Boening C and Landerer F W 2015 Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons *Journal of Geophysical Research: Solid Earth* **120** 2648–71
Werth S, Güntner A, Petrovic S and Schmidt R 2009 Integration of GRACE mass variations into a global hydrological model *Earth Planet. Sci. Lett.* **277** 166–73
Wiese D N, Yuan D-N, Boening C, Landerer F W and Watkins M M 2018 JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height Release 06 Coastal Resolution Improvement (CRI) *Filtered Version 1.0, Ver. 1.0, PO.DAAC, CA, USA* (https://doi.org/10.5067/TEMSC-3MJC6)
Xu T and Liang F 2021 Machine learning for hydrologic sciences: an introductory overview *WIREs Water* **8** e1533
Yamazaki D, de Almeida G A M and Bates P D 2013 Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map *Water Resources Research* **49** 7221–35
Yamazaki D, Kanae S, Kim H and Oki T 2011 A physically based description of floodplain inundation dynamics in a global river routing model *Water Resour. Res.* 47
Yamazaki D, Oki T and Kanae S 2009 Deriving a global river network map and its sub-grid topographic characteristics from a fine-resolution flow direction map *Hydrol. Earth Syst. Sci.* **13** 2241–51
Yu L, Fatichi S, Zeng Y and Su Z 2020 The role of vadose zone physics in the ecohydrological response of a Tibetan meadow to freeze–thaw cycles *The Cryosphere* **14** 4653–73
Zhao F *et al* 2017 The critical role of the routing scheme in simulating peak river discharge in global hydrological models *Environ. Res. Lett.* **12** 075003