



## Understanding terrestrial water storage variations in northern latitudes across scales

Tina Trautmann<sup>1,2</sup>, Sujan Koirala<sup>1</sup>, Nuno Carvalhais<sup>1,3</sup>, Annette Eicker<sup>4</sup>, Manfred Fink<sup>5</sup>, Christoph Niemann<sup>1,5</sup>, and Martin Jung<sup>1</sup>

<sup>1</sup>Department of Biogeochemical Integration, Max-Planck-Institute for Biogeochemistry, 07745 Jena, Germany

<sup>2</sup>International Max Planck Research School for Global Biogeochemical Cycles, 07745 Jena, Germany

<sup>3</sup>CENSE, Departamento de Ciências e Engenharia do Ambiente, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, 2829-516, Portugal

<sup>4</sup>HafenCity University, 20457 Hamburg, Germany

<sup>5</sup>Department of Geography, Friedrich-Schiller University, 07743 Jena, Germany

**Correspondence:** Tina Trautmann (ttraut@bgc-jena.mpg.de)

Received: 27 November 2017 – Discussion started: 29 November 2017

Revised: 30 May 2018 – Accepted: 11 July 2018 – Published: 27 July 2018

**Abstract.** The GRACE satellites provide signals of total terrestrial water storage (TWS) variations over large spatial domains at seasonal to inter-annual timescales. While the GRACE data have been extensively and successfully used to assess spatio-temporal changes in TWS, little effort has been made to quantify the relative contributions of snowpacks, soil moisture, and other components to the integrated TWS signal across northern latitudes, which is essential to gain a better insight into the underlying hydrological processes. Therefore, this study aims to assess which storage component dominates the spatio-temporal patterns of TWS variations in the humid regions of northern mid- to high latitudes.

To do so, we constrained a rather parsimonious hydrological model with multiple state-of-the-art Earth observation products including GRACE TWS anomalies, estimates of snow water equivalent, evapotranspiration fluxes, and gridded runoff estimates. The optimized model demonstrates good agreement with observed hydrological spatio-temporal patterns and was used to assess the relative contributions of solid (snowpack) versus liquid (soil moisture, retained water) storage components to total TWS variations. In particular, we analysed whether the same storage component dominates TWS variations at seasonal and inter-annual temporal scales, and whether the dominating component is consistent across small to large spatial scales.

Consistent with previous studies, we show that snow dynamics control seasonal TWS variations across all spatial scales in the northern mid- to high latitudes. In contrast, we find that inter-annual variations of TWS are dominated by liquid water storages at all spatial scales. The relative contribution of snow to inter-annual TWS variations, though, increases when the spatial domain over which the storages are averaged becomes larger. This is due to a stronger spatial coherence of snow dynamics that are mainly driven by temperature, as opposed to spatially more heterogeneous liquid water anomalies, that cancel out when averaged over a larger spatial domain. The findings first highlight the effectiveness of our model–data fusion approach that jointly interprets multiple Earth observation data streams with a simple model. Secondly, they reveal that the determinants of TWS variations in snow-affected northern latitudes are scale-dependent. In particular, they seem to be not merely driven by snow variability, but rather are determined by liquid water storages on inter-annual timescales. We conclude that inferred driving mechanisms of TWS cannot simply be transferred from one scale to another, which is of particular relevance for understanding the short- and long-term variability of water resources.

## 1 Introduction

Since the start of the mission in 2002, measurements from the Gravity Recovery and Climate Experiment (GRACE) provide unprecedented estimates of changes in the terrestrial water storage (TWS) across large spatial domains (Tapley et al., 2004; Wahr et al., 2004). Due to their global coverage and independence from surface conditions, the data represent a unique opportunity to quantify spatio-temporal variations of the Earth's water resources (Alkama et al., 2010; Werth et al., 2009). Therefore, GRACE data have been widely used to diagnose patterns of hydrological variability (Seo et al., 2010; Rodell et al., 2009; Ramillien et al., 2006; Feng et al., 2013), to validate and improve model simulations (Döll et al., 2014; Güntner, 2008; Werth and Güntner, 2010; Chen et al., 2017; Eicker et al., 2014; Girotto et al., 2016; Schellekens et al., 2017), and to enhance our understanding of the water cycle on regional to global scales (Syed et al., 2009; Felfelani et al., 2017).

Despite the high potential of GRACE data for hydrological applications (Döll et al., 2015; Werth et al., 2009), the measured signal vertically integrates over all water storages on and within the land surface, which challenges the interpretation of the driving mechanism behind TWS variations. To facilitate insight into the underlying processes, hydrological models are frequently used to separate the measured TWS into its different components such as groundwater, soil moisture, and snowpacks (Felfelani et al., 2017). However, as a consequence of uncertain model structure, forcing, and parametrization, model-based partitioning is ambiguous (Güntner, 2008) and may lead to diverging conclusions, especially on regional scale (Long et al., 2015; Schellekens et al., 2017).

While the uncertainties of catchment-scale hydrological models are commonly reduced by calibrating the model parameters against discharge measurements, the majority of macro-scale models rely on a priori parametrization. So far, only a few models used to assess hydrological processes on continental to global scales are constrained by observations, and if so, they are mainly calibrated against the observed discharge of large river basins (Long et al., 2015; Döll et al., 2015). Recently, several studies showed the benefits of additionally including GRACE TWS data in model calibration (Werth and Güntner, 2010; Xie et al., 2012; Chen et al., 2017) or by means of data assimilation (Eicker et al., 2014; Forman et al., 2012; Kumar et al., 2016). However, although these approaches improve model simulations, they do not reduce the uncertainty in the partitioning of TWS due to the parameter equifinality problem (Güntner, 2008). Therefore, it is desirable to include multiple observations, ideally of several hydrological storages and fluxes, to constrain model results (Syed et al., 2009).

Nowadays, the increasing number and quality of Earth-observation-based products provides valuable information on a variety of hydrological variables over large scales, and thus

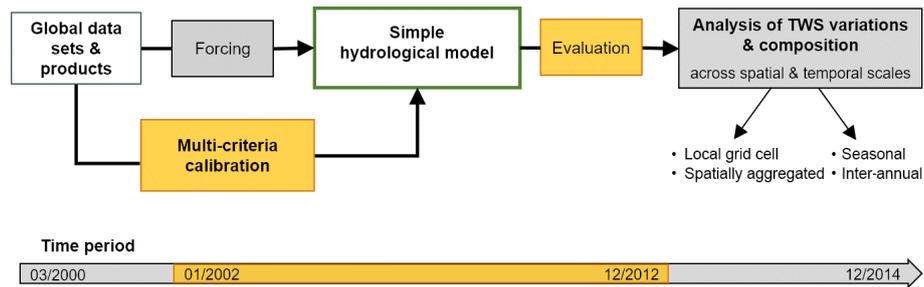
facilitates the constraint of model simulations with multiple data streams simultaneously. While this can provide a more robust understanding of how variations in water storages translate into the observed TWS (Werth and Güntner, 2010), it is very challenging in practice and has rarely been implemented.

On the one hand, this is due to the limitations and inherent uncertainties of each Earth-observation-based product that need to be considered when comparing simulations and observations. For example, satellite-based soil moisture retrievals only capture the upper 5 cm of soil under snow-free conditions and therefore are difficult to compare to modelled soil water (Lettenmaier et al., 2015), while large-scale observations of snow mass based on passive microwave sensors are known to suffer from uncertainties in deep and wet snow conditions (Niu et al., 2007), and multispectral sensors solely provide estimates of snow cover in the absence of clouds (Lettenmaier et al., 2015).

On the other hand, the application of multi-criteria calibration approaches is limited by the increasing complexity of most macro-scale hydrological models over time (Döll et al., 2015). This high model complexity is not only associated with conceptual issues related to over-parametrization (Jakeman and Hornberger, 1993) and large computational demand, but has also been shown to not necessarily improve model performance (Orth et al., 2015). Therefore, it is desirable to implement a rather parsimonious model structure (Sorooshian et al., 1993), especially in multi-criteria model–data fusion approaches.

Applying multiple observational constraints is particularly beneficial in regions where hydrological dynamics are poorly understood and thus their representation in models varies widely. This is the case for snow-dominated regions as the northern high latitudes (Schellekens et al., 2017), which are among the areas most prone to the impacts of climate change (Tallaksen et al., 2015). These regions have been experiencing the strongest surface warming over the last century globally (IPCC, 2014), a trend which is expected to be exacerbated in the future and to significantly change hydrological patterns (AMAP, 2017). Therefore, solid understanding of present hydrological processes and variations is crucial, yet the effect of complex snow dynamics on other storages and water resources is relatively unknown (van den Hurk et al., 2016; Kug et al., 2015). While it has been shown that snow mass is the primary component of seasonal variations of TWS in large northern basins (Niu et al., 2007; Rangelova et al., 2007), it is not known what drives the TWS variations on inter-annual or longer timescales in these regions. Moreover, most analysis has so far focused on individual river basins and do not provide a comprehensive picture over large spatial scales.

In this study, we therefore aim to investigate the contributions of snow compared to other (liquid) water reservoirs to spatio-temporal variations of TWS in the northern mid- to high latitudes. To do so, we establish a model–data fusion ap-



**Figure 1.** Experiment design and considered time periods for forcing and analysis (grey) as well as model calibration and evaluation (orange).

proach that integrates multiple Earth-observation-based data streams including GRACE TWS along with estimates of snow water equivalent (SWE), evapotranspiration, and runoff into a rather simple hydrological model. This model is designed as a combination of standard model formulations yet aims to maintain low complexity in order to facilitate multi-criteria calibration and to focus on variables that can be constrained by observations.

First, we explain the applied methods, including the implemented model, the data used, and the multi-criteria calibration approach. The following section presents and discusses the results obtained with the optimized model. In the results, we describe the calibrated model parameters and evaluate the model performance with respect to observed patterns of TWS and SWE. Subsequently, the relative contributions of snow and liquid water storages to TWS variations are assessed on seasonal and inter-annual scales. Thereby we first focus on spatially integrated values across the study domain, and secondly on the composition on local grid scale. Finally, we summarize our findings and draw the conclusions.

## 2 Data and methods

The following section provides an overview on the experimental set-up, followed by a more detailed description of the model, the input data, and the methods for model calibration and analysis.

### 2.1 Experiment design

To assess the composition of TWS variations in northern mid- to high latitudes, we optimized a simple hydrological model on daily time steps at a  $1^\circ \times 1^\circ$  latitude–longitude resolution. We defined the area of interest as humid land surface north of  $40^\circ$  N, excluding Greenland as well as grids with  $> 90\%$  permanent snow cover and  $> 50\%$  water fraction. Humid areas are derived based on an aridity index  $AI \geq 0.65$ , which was calculated as the ratio of precipitation and potential evapotranspiration (United Nations Environment, 1992). Therefore, we used the same precipitation and potential evapotranspiration data as for model forcing (see Sect. 2.3). To mask out grids with  $> 90\%$  permanent snow

cover and  $> 50\%$  water fraction, we applied the SYNMAP land cover classification (Jung et al., 2006). This dataset has an original resolution of 1 km and was used to determine the fraction of land cover classes within each  $1^\circ \times 1^\circ$  grid cell.

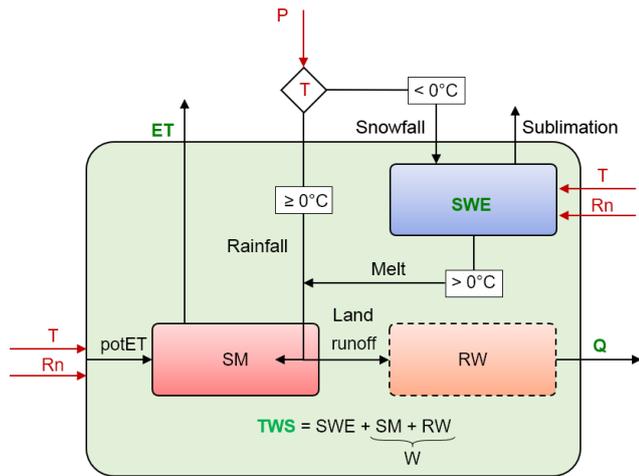
Forced with global observation-based climate data, the model parameters were constrained for a subset of the study domain by multiple Earth observation data products using a multi-criteria calibration approach. These products include terrestrial water storage anomalies as seen by the GRACE satellites (Watkins et al., 2015; Wiese, 2015), measurements of snow water equivalent obtained in the GlobSnow project (Luoju et al., 2014), evapotranspiration fluxes based on FLUXCOM (Tramontana et al., 2016), and runoff estimates for Europe from E-RUN based on E-OBS (Gudmundsson and Seneviratne, 2016). Once the model parameters were calibrated, we evaluated the model against the same data, taking into account the entire study domain. Finally we applied the calibrated model to quantify the contributions of snow and liquid water storages to the integrated TWS. Thereby we considered different spatial domains (local grid cell and spatially aggregated) and temporal scales (mean seasonal and inter-annual variations).

Due to the differences in the temporal coverage of the observational data streams, model calibration and evaluation were conducted for the period 2002–2012, while analysis of TWS components covers the whole period of 2000–2014.

An overview on the experiment design and the selected time periods is provided by Fig. 1, while the following sections give a detailed description of the individual steps.

### 2.2 Model description

We designed a conceptual hydrological model with low complexity and a total number of 10 adjustable parameters. The model considers major hydrological fluxes such as snowmelt, sublimation, infiltration, evapotranspiration, and (delayed) runoff and includes water storages in the snowpack, in the soil, and due to delay in runoff (Fig. 2). It is forced by precipitation ( $P$ ), air temperature ( $T$ ), and net radiation ( $R_n$ ) and calculates all hydrological processes on daily time steps for individual grid cells. A simple schematic diagram of the



**Figure 2.** Schematic structure of the model with calculation of TWS. Boxes denote the water storages (mm): snow water equivalent SWE, soil moisture SM, retained water RW, liquid water  $W$  and total terrestrial water storage TWS. Fluxes are represented by arrows. Red colour identifies forcing data: precipitation  $P$  ( $\text{mm day}^{-1}$ ), air temperature  $T$  ( $^{\circ}\text{C}$ ), and net radiation  $R_n$  ( $\text{MJ m}^{-2} \text{day}^{-1}$ ); green colour indicates variables constrained by observations: evapotranspiration  $ET$  ( $\text{mm day}^{-1}$ ), runoff  $Q$  ( $\text{mm day}^{-1}$ ), SWE (mm), and TWS (mm).

model is shown in Fig. 2, while a detailed description of modelled processes is provided in Sect. S1 in the Supplement.

In the first step, precipitation  $P$  is partitioned into liquid precipitation (rainfall) and snowfall based on a temperature threshold of  $0^{\circ}\text{C}$ . Accumulating snowfall increases the snowpack represented by the snow water equivalent (mm), which depletes by sublimation and melt if  $T$  exceeds  $0^{\circ}\text{C}$ . We calculate sublimation based on the GLEAM model (Miralles et al., 2011) and apply an extended degree-day approach to estimate snowmelt (Kustas et al., 1994). Since the presence of snow can be highly variable in one grid cell, we model the fractional snow cover (–) following Balsamo et al. (2009), which is used to scale snowmelt and sublimation.

Similar to the WaterGAP model (Döll et al., 2002), incoming water from rain and snowmelt is allocated to soil moisture (SM) and land runoff ( $Q_s$ ) depending on soil moisture conditions (Bergström, 1991). SM is represented by a one-layer bucket storage that depletes by evapotranspiration (ET). We calculate ET as the minimum of demand-limited potential ET following the Priestley–Taylor formula (Priestley and Taylor, 1972) and supply-limited ET following Teuling et al. (2006).

As land runoff results from an effective soil water recharge formulation, the calculated runoff is essentially all the water that cannot be stored in the soil. Thus, it implicitly contains both surface and subsurface runoff as well as the percolation to deeper water storages such as groundwater, as well as contributions from surface water bodies. To account for runoff contributions from slow-varying storages, we calcu-

late runoff from each grid cell ( $Q$ ) by applying an exponential delay function on  $Q_s$  (Orth et al., 2013). Based on mass balance, we derive the amount of retained land runoff (RW), which implicitly accounts for the effects of several water pools that are not explicitly represented in the model (groundwater, lakes, wetlands, and the river storage). The sum of RW and SM is then taken as the total liquid water storage ( $W$ ). Frozen soil water is not explicitly included in the model. Further, the model does not account for lateral flow of water among grid cells and does not consider river routing explicitly. While the effect of the routing can be significant in large river basins of humid regions (Kim et al., 2009), it is negligible on the spatial scale of a grid cell (as also shown by small influence of the delayed storage component), and at the temporal scale of monthly aggregated values. To ensure that the model calibration is not affected by river routing, we do not compare simulated runoff to measured river discharge of large basins in our model–data fusion approach.

Finally, the sum of liquid water storage and snow is taken as the modelled terrestrial water storage ( $TWS_{\text{mod}}$ ) of a grid cell for the given time step. Since the delayed runoff contribution is minor at the monthly timescale, we, for simplicity, only focus on the contributions of SWE and total  $W$  to TWS in this study.

### 2.3 Input data

As meteorological forcing we used globally available, daily cumulated gridded precipitation sums ( $\text{mm day}^{-1}$ ), average air temperature ( $^{\circ}\text{C}$ ), and net radiation ( $\text{MJ m}^{-2}$ ) from March 2000 to December 2014.

Precipitation values originate from the  $1^{\circ}$  daily precipitation product version 1.2 of the Global Precipitation Climatology Project (GPCP-1DD) (Huffman et al., 2000; Huffman and Bolvin, 2013), which combines remotely sensed precipitation and observations from gauges. Temperature was obtained from the CRUNCEP version 6.1 dataset (Viovy, 2015), which is a merged product of Climate Research Unit (CRU) TS.3.23 observation-based monthly climatology (1901–2013) (New et al., 2000) and the National Center for Environmental Prediction (NCEP) 6-hourly reanalysis data (1948–2014) (Kalnay et al., 1996). Net radiation is based on radiation fluxes of the SYN1deg Ed3A data product of the Clouds and the Earth’s Radiant Energy Systems (CERES) program of the US National Aeronautics and Space Administration (NASA) (Wielicki et al., 1996).

Rather than using a single data stream, e.g. discharge measurements at the outlet of large continental catchments as used in traditional large-scale hydrological studies, we calibrated the model against multiple observation-based data streams on the grid scale. The integrated datasets include terrestrial water storage anomalies ( $TWS_{\text{obs}}$ ) (mm), snow water equivalent ( $SWE_{\text{obs}}$ ) (mm), evapotranspiration ( $ET_{\text{obs}}$ ) ( $\text{mm day}^{-1}$ ), and gridded runoff estimates for Europe ( $Q_{\text{obs}}$ ) ( $\text{mm day}^{-1}$ ).

$TWS_{obs}$  is derived from the GRACE Tellus Mascon product version 2 based on the GRACE gravity fields Release 05, processed at NASA's Jet Propulsion Laboratory (JPL) (Watkins et al., 2015; Wiese, 2015). The GRACE solutions were corrected for geocentric motion coefficients, according to Swenson et al. (2008), and for variations in Earth's oblateness (C20 coefficient) obtained from satellite laser ranging (Cheng et al., 2013). The glacial isostatic adjustment has been accounted for using the model by A et al. (2013). The dataset provides monthly anomalies of equivalent water thickness relative to the January 2004–December 2009 time-mean baseline for the period 2002–2016. Unlike previous GRACE products based on spherical harmonic coefficients, the JPL RL05M dataset uses equal area  $3^\circ \times 3^\circ$  spherical cap mass concentration blocks (mascons) to solve for monthly gravity field variation. To ensure a clean separation along coastlines within land–ocean mascons, a Coastline Resolution Improvement (CRI) filter has been applied (Watkins et al., 2015). For each mascon, uncertainties were estimated by scaling the formal covariance matrix. To enable hydrological analysis at sub-mascon resolution, we used the provided gain factors to scale the original  $TWS_{obs}$  values.

To gain confidence in the partitioning of the integrated TWS, we additionally used SWE estimates from the European Space Agency's (ESA) GlobSnow SWE v2.0 product (Luoju et al., 2014). The dataset provides daily SWE values (mm) for the non-alpine Northern Hemisphere based on assimilating passive microwave satellite data and observed snow depth from weather stations by applying a semi-empirical snow emission model. Compared to data from stand-alone remote sensing approaches, GlobSnow SWE shows superior performance, even though validation against ground-based measurements still reveals a systematic underestimation of SWE under deep snow conditions due to a change in the microwave behaviour of the snowpack (Derksen et al., 2014; Takala et al., 2011; Luoju et al., 2014).

The ET product is based on FLUXCOM (<http://www.fluxcom.org>, last access: 8 April 2016), i.e. upscaled estimates of latent energy that were derived by integrating local eddy covariance measurements of FLUXNET sites, remote sensing, and meteorological data using the Random Forest (Breiman, 2001) machine learning algorithm (Tramontana et al., 2016). In this study, we apply the Random Forest (Breiman, 2001) realization of FLUXCOM-RS + METEO (see Tramontana et al., 2016 for details). While the product captures seasonality and spatial patterns of mean annual fluxes well, predictions of inter-annual variations remain highly uncertain (Tramontana et al., 2016). In addition, the performance of FLUXCOM ET was found to be lower in extreme environments that are not well represented by FLUXNET sites such as those in the Arctic. An underestimation of the order of 10%–20% of ET can be expected owing to missing energy balance correction prior to upscaling for this respective FLUXCOM ET realization. To calcu-

late  $ET_{obs}$  ( $\text{mm day}^{-1}$ ), we assume a constant latent heat of vaporization of  $2.45 \text{ MJ kg}^{-1}$ .

Similar to TWS that represents the vertically integrated water storage, observations of river discharge spatially integrate hydrological processes within a basin. Thus, they provide an invaluable tool for model validation at large scales. However, it is desirable to apply gridded products to evaluate model performance at local (grid) scale. Therefore, we used the observation-based gridded runoff product E-RUN version 1.1 (Gudmundsson and Seneviratne, 2016) as a constraint for runoff processes. This dataset is based on observed river flow from 2771 small European catchments that was spatially disaggregated to upstream grid cells using a machine learning approach. The data provide mean monthly runoff rates per unit area for each grid, so that river routing is not necessary to directly compare runoff estimates with modelled runoff. Similar to the ET data, gridded runoff estimates show high accuracy for the mean seasonal cycle across Europe, and poorer agreement regarding monthly time series and inter-annual variations (Gudmundsson and Seneviratne, 2016).

Table 1 summarizes the main features of the data used in this study. If required, the data streams were resampled from their original resolution to a consistent  $1^\circ \times 1^\circ$  latitude–longitude grid and common daily (meteorological forcing) and monthly (calibration data) time steps. Data preparation further included extraction of the relevant, overlapping time period and area under consideration.

## 2.4 Multi-criteria calibration

In this study, calibration is intended to identify the set of 10 model parameters (Table 2) that achieves the best fit between simulations and observations for all grid cells while regarding all observational data simultaneously. Thereby, we aimed to exploit the strength of each data stream, while considering known uncertainties and biases. For this purpose, we defined a cost function that takes into account the weakness of each observed variable and evaluates the overall model fit with one value of total cost (see subsequent section). To minimize total costs and thus find the optimal parameter values, we applied the covariance matrix evolution strategy (CMAES) (Hansen and Kern, 2004) search algorithm. The CMAES, as an evolutionary algorithm, is a stochastic, derivative-free method for non-linear, non-convex optimization problems. Compared to gradient-based approaches, it performs better on rough response surfaces with discontinuities, noise, local optima, and/or outliers and is a reliable tool even for global optimization (Hansen and Kern, 2004). Additionally, the CMAES guided search in the parameter space makes the algorithm less computationally demanding than other global optimization approaches, which enumerate a large number of possible solutions (e.g. Monte Carlo–Markov chain methods) (Bayer and Finkel, 2007).

**Table 1.** Overview on data applied for meteorological forcing and multi-criteria calibration and model evaluation (NH: Northern Hemisphere).

	Variable	Dataset	Coverage and resolution		Reference
			Spatial	Temporal	
Meteorological forcing					
$P$	precipitation	GPCP 1dd v1.2	$1^\circ \times 1^\circ$ global	daily 1996–present	Huffman et al. (2000), Huffman et al. (2016)
$T$	air temperature	CRUNCEP v6.1	$0.5^\circ \times 0.5^\circ$ global	daily 1901–2014	Viovy (2015)
$R_n$	net radiation	CERES SYN1deg Ed3A	$1^\circ \times 1^\circ$ global	3-hourly Mar 2000–May 2015	Wielicki et al. (1996)
Calibration and evaluation					
TWS	terrestrial water storage anomalies	GRACE Tellus JPL- RL05M v2	$0.5^\circ \times 0.5^\circ$ global	monthly 2002–2016	Watkins et al. (2015), Wiese et al. (2016b)
SWE	snow water equivalent	GlobSnow v2.0	$0.25^\circ \times 0.25^\circ$ non-alpine NH	daily 1979–2012	Luoju et al. (2014)
ET	evapotranspiration	FLUXCOM	$0.5^\circ \times 0.5^\circ$ global	daily 1982–2013	Tramontana et al. (2016)
$Q$	runoff	EU-RUN v1.1	$0.5^\circ \times 0.5^\circ$ Europe	monthly 1950–2015	Gudmundsson and Seneviratne (2016)

**Table 2.** Adjustable model parameters, their meaning, calibration range (theoretical range in brackets), optimized value including estimated uncertainty, and the corresponding equation in S1.

Parameter	Description	Unit	Range (theoretical)	Optimized			Eq.
				value	$\pm$ uncertainty (%)		
Snow							
$p_{sf}$	scaling factor for snowfall	–	0–3 ( $\infty$ )	0.67	$\pm 1 \times 10^{-3}$	(< 1 %)	(S2)
$sn_c$	minimum SWE that ensures complete snow cover of the grid	mm	0–500 ( $\infty$ )	80	$\pm 19$	(24 %)	(S3)
$m_t$	snowmelt factor for $T$	mm K <sup>-1</sup> day <sup>-1</sup>	0–10	2.63	$\pm 0.26$	(10 %)	(S4)
$m_r$	snowmelt factor for $R_n$	mm MJ <sup>-1</sup> day <sup>-1</sup>	0–3	0.90	$\pm 0.05$	(6 %)	(S4)
$sn_a$	sublimation resistance	–	0–3	0.44	$\pm 0.01$	(3 %)	(S5)
Soil							
$s_{exp}$	shape parameter of runoff–infiltration curve	–	0.1–5	1.46	$\pm 0.02$	(2 %)	(S12)
$s_{max}$	maximum soil water holding capacity	mm	10–1000 (0– $\infty$ )	515	$\pm 9$	(2 %)	(S12)
$\eta_a$	alpha coefficient in Priestley–Taylor formula	–	0–3	1.20	$\pm 0.01$	(1 %)	(S14)
$\epsilon_{sup}$	ET sensitivity and/or SM fraction available for ET	day <sup>-1</sup>	0–1	0.02	$\pm 6 \times 10^{-5}$	(< 1 %)	(S18)
Runoff							
$q_t$	recession timescale for land runoff	d	0.5 (0)–100	13	$\pm 4$	(31 %)	(S20)

In order to keep computational demands low and to avoid overfitting by a very small sample size, we perform calibration for a subset of 1000 randomly chosen grid cells. Within this iterative process, the model simulations are carried out on daily time steps, while costs are calculated based on monthly values. Further, each model run includes an initialization based on 10 random years that were selected a priori.

### Cost function

To objectively describe the goodness of fit, we defined a cost function based on model efficiency (Nash and Sutcliffe, 1970), but with explicit consideration of the uncertainty  $\sigma_i$  of the observed data stream as follows:

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

where  $x_{\text{obs},i}$  is the observed data,  $\bar{x}_{\text{obs}}$  is the average of  $x_{\text{obs}}$ , and  $x_{\text{mod},i}$  is the modelled data of each space–time point  $i$ . Similar to model efficiency, the criterion reflects the overall fit in terms of variances and biases, yet with an optimal value of 0 and a range from 0 to  $\infty$ . Costs are calculated for each variable separately, considering only grid cells and time steps with available observations, which vary for the different data streams. Additionally, to overcome the sensitivity to outliers arising from data uncertainties or inconsistencies, we adopted a 5th-percentile outlier removal criterion (Trischenko, 2002), i.e. the data points with the highest 5 % residuals  $x_{\text{obs}} - x_{\text{mod}}$  were excluded in the cost function.

The costs of each observed variable and its modelled counterpart are then added equally to derive a single value of total cost (Eq. 2). Since a perfect simulation would yield a total cost of 0, calibration aims to find the global minimum of  $\text{cost}_{\text{total}}$ .

$$\text{cost}_{\text{total}} = \text{cost}_{\text{TWS}} + \text{cost}_{\text{SWE}} + \text{cost}_{\text{ET}} + \text{cost}_Q \quad (2)$$

As the uncertainty  $\sigma$  of observational data in Eq. (1) is adapted to best reflect the strength of the individual data stream, we preselected the strongest aspect of the data to be included in the cost function. Owing to the larger uncertainties of  $\text{ET}_{\text{obs}}$  and  $Q_{\text{obs}}$  on inter-annual scales, we only employed the grid's mean seasonal cycles, while the full monthly time series of gridded  $\text{TWS}_{\text{obs}}$  and  $\text{SWE}_{\text{obs}}$  were taken into account.

As  $\text{ET}_{\text{obs}}$  and  $Q_{\text{obs}}$  do not explicitly provide uncertainty estimates, we assume an uncertainty of 10 % and minimum of 0.1 mm, respectively. In order to define  $\sigma$  of  $\text{TWS}_{\text{obs}}$  we utilized the spatially and temporally varying uncertainty information provided with the GRACE data. Additionally, the monthly values of observed and modelled TWS datasets were translated as anomalies to a common time-mean baseline of their overlapping period 1 January 2002–31 December 2012 before calculating the cost for TWS.

For SWE, we applied an absolute uncertainty of 35 mm based on reported differences to ground measurements (Liu et al., 2014; Luoju et al., 2014). Since GlobSnow SWE saturates above approx. 100 mm (Luoju et al., 2014), we do not penalize model simulations when both  $\text{SWE}_{\text{obs}}$  and  $\text{SWE}_{\text{mod}}$  are larger than 100 mm in order to prevent the propagation of data biases to calibrated model parameters.

For maps of the temporal average uncertainties see Sect. S2.

## 2.5 Evaluation of model performance

Once the parameters were optimized, we applied the model for the entire study domain and evaluated its performance

regarding all grid cells (6050) in terms of Pearson correlation coefficient  $r$  and root mean square error RMSE for each variable with observational data. On the one hand, the overall performance at local scale was assessed by calculating  $r$  and RMSE for the monthly time series of each grid individually. On the other hand, the model performance over the entire study domain was evaluated by comparing the seasonal and inter-annual dynamics of the regional average. Therefore, we defined inter-annual variation (IAV) as the deviation of the monthly values from the mean seasonal cycle (MSC). As with the calibration, we focused on the common time period 2002–2012 and considered only the grid cells and time steps with available observations.

In order to benchmark our model against current state-of-the-art hydrological models, we compared its simulations with the multi-model ensemble of the global hydrological and land surface models of the earth2Observe dataset (Schellekens et al., 2017). This ensemble includes HTESSEL-CaMa (Balsamo et al., 2009), JULES (Best et al., 2011; Clark et al., 2011), LISFLOOD (van der Knijff et al., 2010), ORCHIDEE (Krinner et al., 2005; Ngo-Duc et al., 2007; d'Orgeval et al., 2008), SURFEX-TRIP (Alkama et al., 2010; Decharme et al., 2013), W3RA (van Dijk and Warren, 2010; van Dijk et al., 2014), WaterGAP3 (Flörke et al., 2013; Döll et al., 2009), PCR-GLOBWB (van Beek et al., 2011; Wada et al., 2014), and SWBM (Orth et al., 2013). For consistency, we processed the model estimates in the same manner as our model simulations to directly compare modelled SWE and TWS to observations from GlobSnow and GRACE, respectively. While each model provides simulated SWE, they vary in the representation of other storage components. We calculated modelled TWS for each model by summing up the available water storage components. Thus, the variables contributing to modelled TWS vary between the models, which impedes detailed comparison. Additionally, we calculated the multi-model mean of SWE and TWS simulations.

## 2.6 Analysis of TWS variations and composition

Finally, the contribution of snow and liquid water to seasonal and inter-annual TWS variability was quantified across spatial scales. For this, we ran the model with optimized parameters for the entire study domain from 2000 to 2014 and translated simulated storages as anomalies to the time-mean baseline. As in the model evaluation, the MSC and IAV of  $\text{SWE}_{\text{mod}}$ ,  $W$ , and  $\text{TWS}_{\text{mod}}$  anomalies were calculated at local scale for each grid individually and as spatial average over all grid cells. To assess storage variability, the variance in the MSC and the IAV of each storage component was computed. Assuming negligible covariance of snow and liquid water (see Sect. S8), their relative contribution to TWS variance was calculated as the contribution ratio CR:

$$CR = \frac{\text{var}(W)}{\text{var}(TWS_{\text{mod}})} - \frac{\text{var}(SWE_{\text{mod}})}{\text{var}(TWS_{\text{mod}})}. \quad (3)$$

While  $CR = 0$  indicates equal contribution of snow and liquid water to TWS variability, positive (negative) values of  $CR$  imply that variations of  $TWS_{\text{mod}}$  mainly result from variations in liquid water (snowpack), with  $CR = +1$  meaning that all variation is explained by liquid water and  $CR = -1$  suggests determination solely by snow.

From Eq. (3) and the assumption that  $\text{var}(W) + \text{var}(SWE) = \text{var}(TWS)$ , the percentage contribution of liquid water storages to the variability of TWS can be inferred as  $CW$ :

$$CW = \frac{\text{var}(W)}{\text{var}(TWS_{\text{mod}})} = \frac{CR + 1}{2}. \quad (4)$$

As this study intends to analyse the effects of storage components on TWS at different spatial scales (local grid scale and large (regional) spatial averages), the difference in spatial heterogeneities of these components is considered. Some storage components, e.g. soil moisture anomalies, have much larger spatial variability than others. Due to this high small-scale heterogeneity, the effect on larger regional scale might be smaller than expected, as different local scale heterogeneities compensate for each other when the regional averages are calculated (Jung et al., 2017). Thus, we assessed the spatial coherence of simulated patterns of SWE and  $W$  by calculating the proportion of total positive and total negative covariances among grid cells (Eqs. 4 and 5 in Jung et al., 2017). If the sum of positive covariances outweighs the sum of negative covariances, it implies some degree of spatial coherence of the anomalies. Spatial coherence of anomalies then causes a larger variance in the averaged anomalies compared to the sum of the variances of individual grid cells. This assessment of spatial coherence of SWE and  $W$  anomalies allows for understanding different contributions of SWE and  $W$  to TWS variability at local scale compared to the regional scale.

### 3 Results and discussion

The following sections present and discuss the results obtained with the calibrated model. First, we review the calibration approach and the optimized parameter values. Then the model is validated with respect to its overall performance at grid scale, as well as the reproduction of average seasonal (MSC) and inter-annual (IAV) dynamics. Subsequently, we assess the driving component of spatially integrated TWS variations and the relative contributions of snow and liquid water to TWS variability on local scale. Finally, we summarize the results across spatio-temporal scales.

#### 3.1 Model optimization

Optimization of the model identifies the parameter values listed in Table 2 as being most suitable regarding all data constraints simultaneously. The CMAES search algorithm converged after 3272 function evaluations as no further improvement of  $\text{costs}_{\text{total}}$  could be achieved, which suggests a reliable estimate of the global optimal parameter set. The individual cost terms obtained with default and optimized parameter values are contrasted in Table S1 in the Supplement. Overall, this parameter set obtained for a subset of 1000 random grids is reasonable with respect to reported “plausible” parameter ranges, with none of them reaching their physically and/or technically defined upper and lower calibration bounds.

In detail, snowfall is reduced by  $p_{\text{sf}}$  to 67 % of precipitation occurring at  $T < 0^\circ\text{C}$ . This reduction agrees with Behrangi et al. (2016), who found that GPCP overestimated snowfall over Eurasian high latitudes by about 20 % compared to other precipitation products. Similar, overestimation of precipitation undercatch correction in GPCP has been reported by Swenson (2010). Taking into account the mismatch in temporal and spatial domains, as well as the experimental definitions, reducing GPCP snowfall in our study by 33 % is roughly consistent with both studies. Therefore,  $p_{\text{sf}}$  allows the reduction of inconsistencies between the precipitation forcing and the water storages as given by GlobSnow SWE and GRACE TWS.

Further, each grid is assumed to be completely covered by snow if  $SWE \geq 80 \text{ mm}$  ( $sn_c$ ). On the one hand, the snowpack can be reduced by sublimation, with  $sn_a = 0.44$  indicating relatively high sublimation resistance, compared to a default of  $sn_a = 0.95$  proposed by Miralles et al. (2011). The divergence probably results from interaction with snowmelt, as net radiation also contributes to melt with  $0.9 \text{ mm MJ}^{-1}$  ( $m_r$ ) if  $T$  exceeds  $0^\circ\text{C}$ . On the other hand, melt is mainly induced by temperature, as the estimated degree-day factor ( $m_t$ ) is  $2.63 \text{ mm K}^{-1}$ , which is close to typical values of  $3 \text{ mm K}^{-1}$  (Müller Schmied et al., 2014; Stacke, 2011). These parameter interactions underline an equifinality issue between modelled snowmelt and sublimation due to missing data constraints, resulting in larger parameter uncertainties for  $sn_a$ ,  $m_r$ , and  $m_t$ . However, for the objective of this study it is not primarily relevant whether sublimation- or radiation-induced melt decreases the snowpack, as the total snow loss amount remains relatively unchanged for different parameter combinations.

The maximum soil water holding capacity is set to 515 mm after calibration, a comparatively high value that is likely to include storages in surface water bodies such as lakes and wetlands within our study domain. The optimized value of  $s_{\text{exp}}$  is 1.46, which suggests a non-linear relationship between soil moisture storage and runoff generation. For the same amount of incoming water (rainfall and snowmelt), the non-linear relationship produces a smaller runoff and larger infiltration than a linear relationship ( $s_{\text{exp}} = 1$ ).

Regarding evapotranspiration, the alpha coefficient ( $e_a$ ) in the Priestley–Taylor formula is generally taken as 1.26 for well-watered crops based on experimental observations (Priestley and Taylor, 1972; Eichinger et al., 1996). Thus, the optimized value of 1.20 for  $e_a$  reflects a plausible value. Further,  $e_{\text{sup}}$  indicates that 2 % of the available soil moisture can evaporate per day (including transpiration), which lies within the range of site-specific ET sensitivities from 0.001 to 0.5 day<sup>-1</sup> and is close to the median value (5 %) (Teuling et al., 2006).

Finally, the calibrated recession timescale that delays land runoff is 13 days ( $q_t$ ). Compared to much smaller alpine catchments for which Orth et al. (2013) reported  $q_t$  of 2 days, the longer delay coefficients are reasonable at a spatial resolution of 1° × 1° grids, because the elevation gradients are much smaller within a large spatial area. At first glance, 13 days appear to be quite a short effective time period, as the delay is supposed to comprise contributions from much slower depleting reservoirs, such as lakes and deep groundwater. However, implementing and calibrating a simple groundwater storage, which is recharged with some proportion of land runoff and linearly depletes over time, led to similar retardation times.

The uncertainty in the optimized parameter vector was estimated by quantifying each parameter's standard error as the square root of the product between the diagonal elements of the parameters' covariance matrix (calculated from the Jacobian matrix) and the sum of residual squares according to Omlin and Reichert (1999) and Draper and Smith (1981). The resulting relative parameter uncertainty is particularly instructive for comparing how well individual parameters could be constrained.

Most parameters were well constrained (Table 2), suggesting that our model–data fusion method, fed by multiple observation streams, succeeded in reducing the initial theoretical parameter ranges (up to 500 %) to much narrower ranges. Nonetheless, some parameters have a larger uncertainty range than others (e.g.  $q_t$ ,  $sn_c$ ,  $m_t$ ), which may highlight a limitation in suitable observations to constrain them, as well as a lower sensitivity of the model results and the cost function used. Further, given that the model only considers the spatial variability of climate, the uncertainty in global parameters obtained from inversion may reflect the natural variations in these parameters that arise from differences in local land surface characteristics such as topography or land cover.

We adopted the calibrated parameter values as global constants for model simulations over the entire study domain. Even though the globally uniform parameters may not provide perfect simulation for all grids over a large study domain, this approach represents a compromise between a priori parametrization of the model and its calibration at local or regional (e.g. basin) scale. While local and regional model calibration enables good adaption to geographic characteristics, it easily leads to overfitting of the model and thus propa-

gates the constraints' inherent errors and uncertainties in the modelling result. As these uncertainties often vary in space, globally uniform parameter values diminish overfitting uncertainties. In addition, calibration for several independent grids is computationally demanding and subsequently requires a parameter regionalization approach (He et al., 2011). Since such approaches are not commonly accepted (Sood and Smakhtin, 2015; Bierkens et al., 2015), macro-scale models mostly apply a priori parameters based on empirical values or on expert knowledge, which may yet lead to suboptimal simulations (Beck et al., 2016; Sood and Smakhtin, 2015).

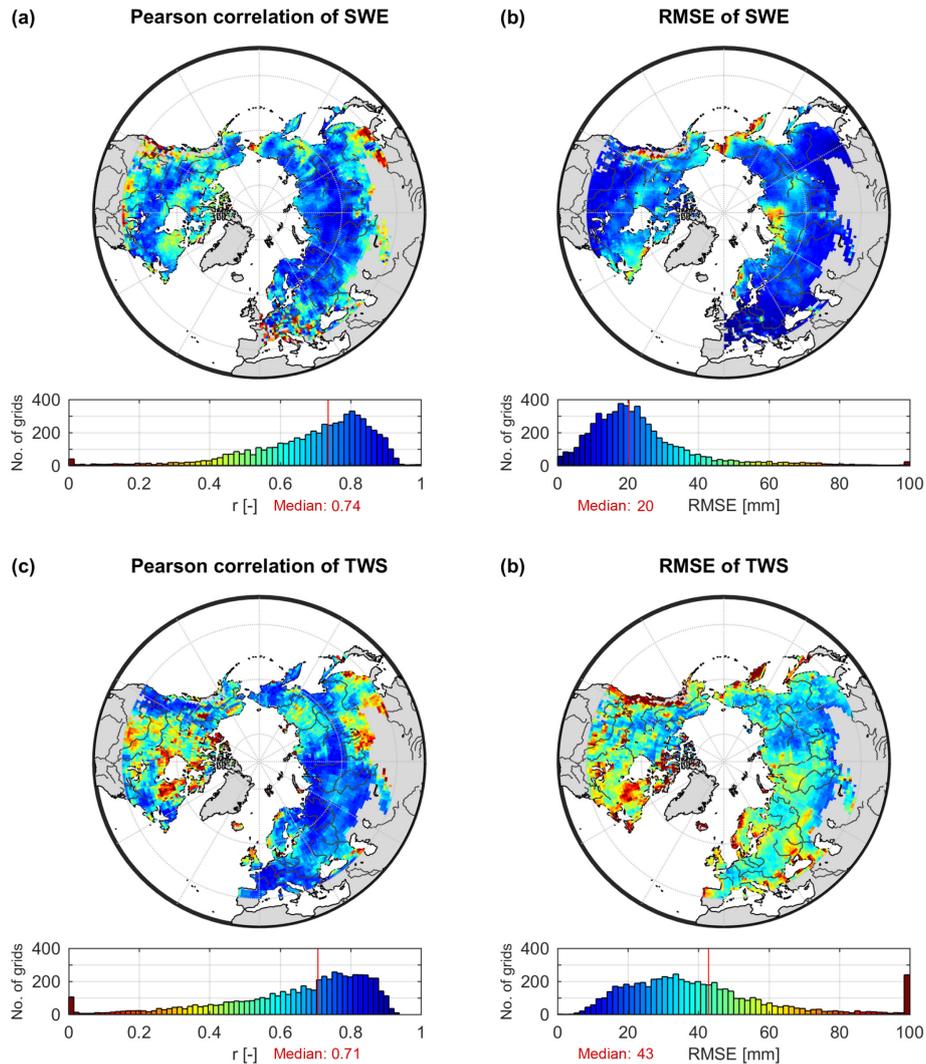
### 3.2 Model performance

For model validation, we used the optimized parameter values to simulate hydrological fluxes and states of the 2002–2012 period over the entire study domain and evaluated the model results against the observation-based data of TWS, SWE, ET, and  $Q$ .

In general, all observed patterns are reproduced very well, taking into account the specific data weaknesses. We achieve a “near-perfect” correlation of 0.99 and 0.94 for mean seasonal variations of ET and  $Q$ , respectively. The median RMSE of mean seasonal ET is 11 and 9.5 mm month<sup>-1</sup> for  $Q$ , which represent 15 % resp. 17 % of the average observed annual amplitude. At the inter-annual scale, though, larger discrepancies exist, which at least partly arise from larger uncertainties in  $ET_{\text{obs}}$  and  $Q_{\text{obs}}$  (Sect. S4). Thus, we assume high confidence in modelled ET and  $Q$  fluxes and subsequently focus on evaluation of the water storages TWS and SWE.

#### 3.2.1 Performance on local grid scale

Overall, the model performs well compared to the observations of monthly time series of SWE and TWS (Fig. 3). More than half of the grid cells obtain correlation values higher than 0.74 between  $SWE_{\text{obs}}$  and  $SWE_{\text{mod}}$ . In general, the median RMSE is 20 mm, which is smaller than the average uncertainty of 35 mm in  $SWE_{\text{obs}}$ . The correlation reduces in lower latitudes where seasonal snow accumulation and thus variability is small. Further, the correlation is also relatively weaker in arctic North America and the Rocky Mountains, while larger deviations between observed and modelled snow quantities centre around mountainous and coastal regions (e.g. Rocky Mountains, Kamchatka), and regions with the largest seasonal snow accumulation (Labrador Peninsula, North Siberian Lowland and northern West Siberian Plain). There are several reasons for this relatively poorer performance. First, the GlobSnow measurements do not cover mountainous areas due to the sub-grid variability of snow depth and high uncertainties in the microwave measurements in complex alpine terrains (Takala et al., 2011). As the resampling and the coarse resolution of each grid in this study compound a distinct alpine or non-alpine classification, these

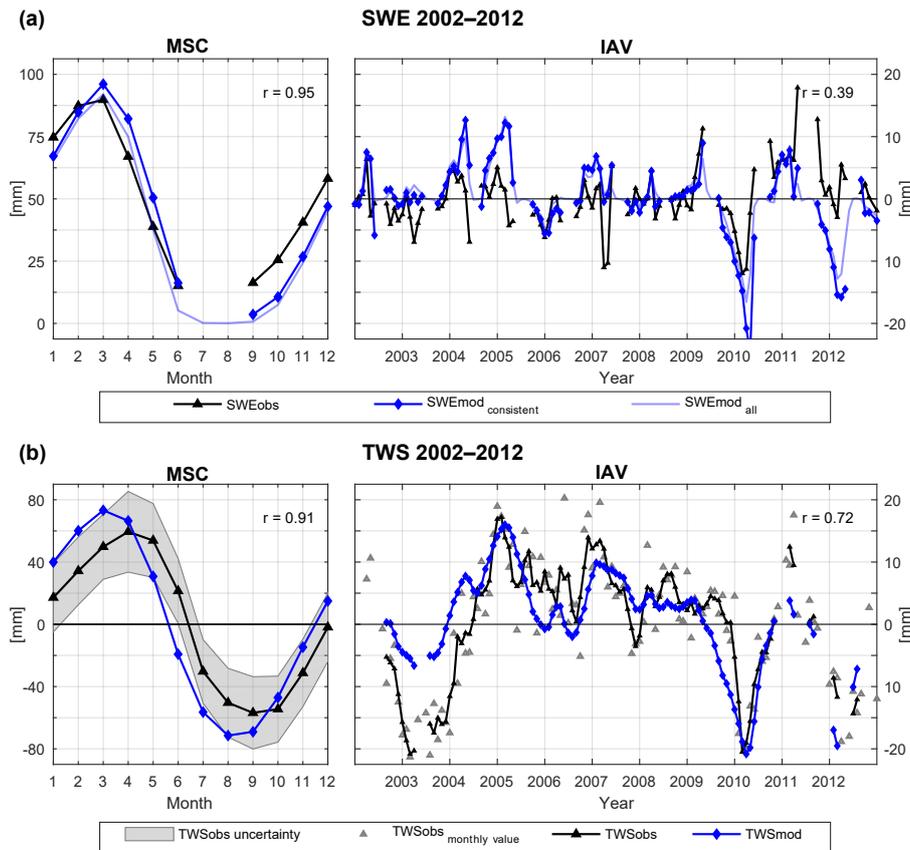


**Figure 3.** Pearson correlation coefficient  $r$  (a, c) and root mean square error RMSE (b, d) between monthly values of modelled SWE and GlobSnow SWE (a, b), as well as modelled TWS and GRACE TWS (c, d) for the period 2002–2012 and for each  $1^\circ \times 1^\circ$  grid cell of the study domain. Values of  $r$  are truncated to the range 0–1 (a, c), and values of RMSE to the range 0–100 mm (b, d).

uncertainties leak to the surrounding areas. Second, neither the input forcing data nor our model include the sub-grid scale heterogeneity of climate (e.g. precipitation and temperature) and hydrological processes, which may be significant in near-mountain or coastal regions. Additionally, the accuracy of observed large snow accumulation is limited as the radar-retrieval methods tend to saturate at large  $SWE_{\text{obs}}$  values, which then leads to large RMSE of the model simulation.

Similar to SWE, more than half of the grid cells show a strong correlation of 0.71 between  $TWS_{\text{obs}}$  and  $TWS_{\text{mod}}$ , which reflects a realistic temporal variation in the model simulation. Compared to SWE, the RMSE of TWS is somewhat higher, yet the median of 43 mm still reflects the range of  $\pm 22$  mm average uncertainty in GRACE  $TWS_{\text{obs}}$  of the study domain (Wiese, 2015). However, when comparing GRACE

TWS with model simulations, several aspects have to be considered. First,  $TWS_{\text{obs}}$  as an integrated signal comprises all water storages, not all of which are (sufficiently) represented in the model structure. Second, although GRACE TWS passed through various pre-processing steps, the models that account for postglacial rebound or leakage between neighbouring grid cells, for example, introduce their own uncertainties and do not remove the effects completely. Further, with a native resolution of  $3^\circ$ , uncertainties remain for grids that comprise large variability at sub-grid scale and depend on the model used to estimate GRACE scaling factors (Wiese et al., 2016a). Altogether this is reflected in higher RMSE in arctic regions (e.g. surrounding the Hudson Bay), as well as in heterogeneous coastal and mountainous regions. Additionally, our model shows a weaker performance in subarctic and arctic wetlands, and in central North America and east-



**Figure 4.** Spatially averaged mean seasonal cycle (MSC) of the period 2002–2012 as well as inter-annual variability (IAV, difference between monthly values and the MSC) for (a) SWE and (b) TWS. In (a),  $SWE_{\text{mod consistent}}$  refers to modelled SWE considering only data points with available  $SWE_{\text{obs}}$ , while  $SWE_{\text{mod all}}$  incorporates all time steps for all grids of the study domain. Correlation  $r$  is calculated only for consistent data points. In (b) IAV,  $TWS_{\text{obs monthly value}}$  shows the original IAV of individual  $TWS_{\text{obs}}$  months, while  $TWS_{\text{obs}}$  and  $TWS_{\text{mod}}$  are smoothed using a 3-month average moving window filter. Correlation  $r$  refers to the smoothed values. For the MSC in (b) no smoothing is applied.

ern Eurasia. The latter are both relatively dry regions that are rather dominated by inter-annual TWS variations (Humphrey et al., 2016). Discrepancies between  $TWS_{\text{obs}}$  and  $TWS_{\text{mod}}$  thus relate to a low signal-to-noise ratio in  $TWS_{\text{obs}}$  due to small seasonal TWS variations. However, the anthropogenic influence for irrigational withdrawal is very large in these regions, yet such processes are not considered in our model. We also lack explicit surface water storages (including wetland dynamics), which may be the reason for poorer performance, especially in North American wetland regions.

### 3.2.2 Performance of the spatially integrated simulations

Since the aim of this study is to analyse the composition of TWS across temporal scales, we additionally evaluated average (spatially integrated) MSC and IAV of SWE and TWS (Fig. 4). While the mean seasonal variations of both observational data streams are relatively robust and have been used for model evaluation before (Alkama et al., 2010; Döll et

al., 2014; Schellekens et al., 2017; Zhang et al., 2017), their inter-annual variations are more uncertain and contain considerable noise. This clearly reduces the information content in the observational data, so that we evaluate the IAV in more qualitative terms.

As with the comparison at grid scale, the spatially averaged  $SWE_{\text{mod}}$  compares well to  $SWE_{\text{obs}}$ , with a correlation of 0.95 suggesting a good reproduction of seasonal snow accumulation and ablation processes (Fig. 4a). Owing to the high uncertainty of  $SWE_{\text{obs}}$  peaks due to signal saturation, the higher amplitude of  $SWE_{\text{mod}}$  seems reasonable. Although inter-annual variations are not as well represented as the MSC, general tendencies, e.g. increasing/decreasing positive/negative anomalies, coincide.

Similar to SWE, the spatial average of TWS shows high correlation of 0.91 for seasonal variations, with positive anomalies from December to May–June and negative anomalies during summer and autumn months (Fig. 4b). Even though the modelled amplitude is slightly larger than the observed one, it stays within the uncertainty range of  $TWS_{\text{obs}}$

for most months, suggesting reliable simulations. However,  $TWS_{\text{mod}}$  precedes  $TWS_{\text{obs}}$  on average by 1 month, reaching the maximum in March instead of April, and the minimum in August instead of September. A similar phase shift of 1 month between GRACE TWS and modelled TWS has been reported by several state-of-the-art global models (Döll et al., 2014; Schellekens et al., 2017). It should be noted that some areas such as eastern North America, Kamchatka, Scandinavia, and western Europe do not show phase differences, while the lag in south-eastern Eurasia is even larger, as already suggested by lower overall correlation (Fig. S5). In general, the disagreement in timing is attributed to the lack of sufficient water storages and the delay mechanism within the model, so that the modelled system reacts too fast (Schellekens et al., 2017; Döll et al., 2014; Schmidt et al., 2008). Thus, we implemented model variants with an explicit groundwater storage to delay depletion of TWS, with spatially varying soil properties to better represent heterogeneous infiltration and runoff rates, as well as a variant that applied a more sophisticated approach to calculate snow dynamics based on energy balance. Despite the efforts, we achieved no improvement in terms of reducing the phase shift. Therefore, the question arose as to whether it is not primarily the model formulation that prevents correction of the temporal delay, but rather the combination of forcing data and observational constraints. To further preclude possible errors due to such data inconsistencies, e.g. between GRACE TWS and GlobSnow SWE, we excluded GlobSnow SWE data from calibration. Although this could slightly improve the agreement of TWS MSC, it led to unrealistic behaviour of snow dynamics, and thus did not offer any advantages. Besides, we found no major differences in the magnitude or spatial distribution of the phase shift resulting from the precipitation forcing (GPCP vs. WFDEI) or compared to other GRACE solutions (Sect. S6). Further, the lag in TWS simulation can occur due to several mechanisms and processes that are not yet considered in the current model structure, such as lateral flow and surface storages (wetland and lakes), vegetation processes, glacier melt, and human influence with dams and reservoirs. However, we do not observe a general or systematic relationship with either elevation, land cover type, soil properties, or the occurrence of lakes and wetlands. There is a tendency that larger negative lags occur more frequently in regions with sporadic permafrost, but the ranges of permafrost fractions are large for both short and long lags in TWS, suggesting a complex interaction between permafrost extent and its effect on lag in seasonal TWS dynamics. Finally, potential biases in the timing of ET due to snow cover and/or vegetation processes may also affect the timing of the depletion of SM and TWS. Additionally, high uncertainties of the precipitation forcing and GlobSnow SWE product in (near-)mountain regions, as well as leakage errors in the GRACE signal influence the accuracy of both  $TWS_{\text{obs}}$  and  $TWS_{\text{mod}}$ . Although these shortcomings should be kept in mind, we assumed that they do not significantly affect our

results regarding to the relative contributions of snow and liquid water to TWS.

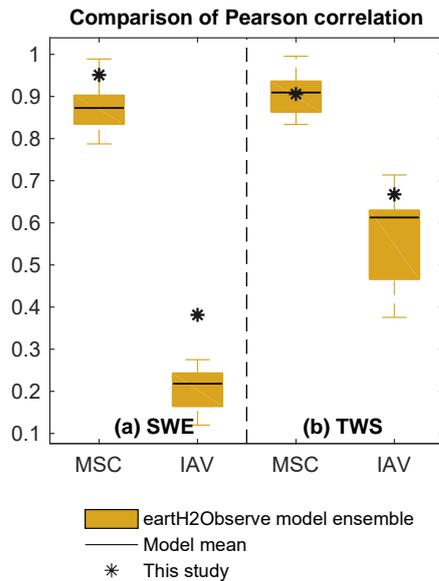
In terms of inter-annual variations, the variance in monthly  $TWS_{\text{obs}}$  values is highly underestimated by modelled TWS, which on the one hand relates to noise within the GRACE signal, but on the other hand may again reflect missing process representation in the model. To reduce the noise, we applied a 3-month moving-average filter on the monthly time series. The smoothed time series then shows fairly good agreement of inter-annual dynamics, with correlation  $r = 0.68$  (Fig. 4b, solid lines). Only the amplitude of the large negative anomaly in 2003 is not captured by the model. While the spatial pattern of this negative TWS anomaly can be simulated, the forcing data do not show large anomalies in 2003, so that the model fails to reproduce the magnitude of observed TWS, especially in North America. Issues with the precipitation forcing are further suggested by a negative SWE anomaly of on average 5 mm (see Fig. 4a), indicated in the GlobSnow data, that is not captured by the model, either. The reason why this snow anomaly is not captured by the forcing remains unclear at this point – it persists when using the WFDEI forcing dataset. Besides, the agreement between GRACE and modelled TWS IAV gets substantially better when isolating inter-annual variations by removing the trends in both TWS time series (increase in correlation  $r$  from to 0.77). This suggests that the trend in GRACE TWS is to some extent either subject to observational issues or represents a process that is not captured by the model.

### 3.2.3 Comparison with the earth2Observe model ensemble

Compared to the model ensemble of the earth2Observe dataset, we achieve an equally good or better performance for the spatially integrated SWE and TWS on both MSC and IAV scales (Figs. 5 and S6). Besides, the majority of the model ensemble obtains similar spatial patterns of performance criteria for SWE as well as for TWS (not shown).

The average observed MSC of SWE is captured with a correlation in the range of 0.79 (PCR-GLOBWB) to 0.99 (ORCHIDEE), whereby only ORCHIDEE shows a better correlation than our model ( $r = 0.95$ ). However, modelled snow accumulation exceeds that of  $SWE_{\text{obs}}$  for the majority of the models, which is also reflected in higher RMSE (Figs. S6–S8). On IAV scales, the correlation is lower in general, yet again we obtain a better fit ( $r = 0.39$ ) than the model ensemble ( $r = 0.12$  – ORCHIDEE – to  $0.28$  – LISFLOOD). However, it remains uncertain whether the discrepancies between  $SWE_{\text{obs}}$  and  $SWE_{\text{mod}}$  represent model deficiencies or evolve from issues related to the GlobSnow SWE retrieval (Schellekens et al., 2017).

Regarding average seasonal TWS variations, our model performs as well as the model ensemble ( $r = 0.91$ ), with the range of the earth2Observe ensemble spanning from  $r = 0.83$  (ORCHIDEE) to  $r = 1.00$  (PCR-GLOBWB). The



**Figure 5.** Pearson correlation for the spatially integrated SWE (a) and TWS (b) achieved by this study compared to the model ensemble of the earth2Observe dataset across temporal scales. In each box, the edges represent the 25 % and 75 % percentiles of the model ensemble, while the solid black line marks the performance of the ensemble mean.

amplitudes in the MSC of  $TWS_{mod}$  (95 to 156 mm) are comparable to the observed amplitude of 118 mm, except for SWBM, whose amplitude is twice as large as that of  $TWS_{obs}$ . This discrepancy is reflected in relatively high RMSE values for SWBM (Fig. S8). The model ensemble precedes observed seasonal TWS variations by 1 to 1.4 months, similar to our estimates of  $TWS_{mod}$  (−1.1 month). Only PCR-GLOBWB, with a higher correlation than other models, shows a smaller average lag of less than 1 month (−0.3 months). This minor difference results from balancing out the preceding and succeeding of  $TWS_{mod}$  in different regions over the study domain. Additionally, Schellekens et al. (2017) found that PCR-GLOBWB shows unrealistic snow accumulation over time in Europe and boreal North America. Except for PCR-GLOBWB, the majority of the models obtain comparable spatial pattern of preceding TWS, with small differences at regional scales. Even though the difference in the MSC is commonly attributed to the lack or inadequate size and number of water storages (Kim et al., 2009), a relationship between model performance and model complexity is not obvious. Relatively complex models, such as HTESSSEL, SURFEX, and JULES, show similar phase differences to simpler models, such as SWBM and our model (−1.0 and −1.1 months, respectively). Since Schellekens et al. (2017) found the largest phase differences in cold regions, they postulate the implementation of processes associated with snow as an important factor for this phase lag. In this context, constraining the model with snow observations as

done in our study should increase confidence in the representation of snow processes. Nevertheless, we obtain a similar phase difference, which points to the importance of other hydrological processes and storages even in snow-affected regions.

Although our modelling framework assimilates information from more data streams compared to the model simulations in the Earth2Observe ensemble, e.g. GRACE and GlobSnow data, we only used a subset of 1000 random grid cells to constrain the model parameters. Despite this, our model performs better than the Earth2Observe ensemble over the whole domain (6050 grids). This improvement in model performance is also consistent among several modelled variables and not limited to storage components only. This suggests that remote sensing data, with larger spatial coverage than site measurements, have large potential to improve hydrological simulations over a large domain. In addition, remote sensing data also hold potential beyond their use as an observational constraint and can provide information on identifying and formulating functional relationships across several spatial and temporal scales, which should be addressed in future efforts.

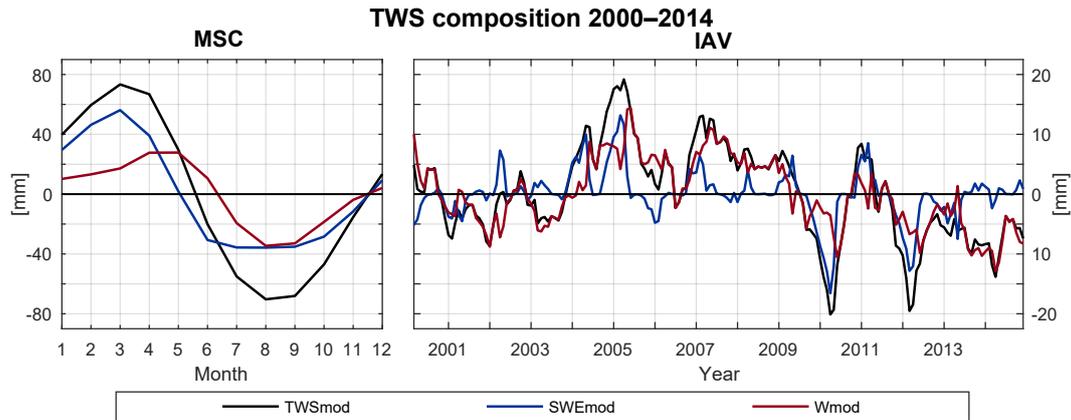
All in all, we conclude that our simple model with a global uniform parameter set achieves considerably good performance regarding observed patterns, especially compared to well-established, more complex models, and with respect to its simplicity and given uncertainties of forcing and calibration data. Thus, we found the model results to be suitable to analyse the composition of TWS across spatial and temporal scales.

### 3.3 TWS variation and composition

#### 3.3.1 Spatially integrated

To assess the average composition of seasonal and inter-annual TWS variations, we spatially integrated modelled TWS anomalies as well as modelled anomalies of snow (SWE) and liquid water storages ( $W$ ) across all grids of the study domain (Fig. 6).

Regarding the MSC, all water storages show a clear seasonal pattern. The maximum  $TWS_{mod}$  in March coincides with the maximum in  $SWE_{mod}$ . On the contrary,  $W$  remains nearly constant throughout the winter, related to the lack of evapotranspiration losses and missing infiltration due to prevailing solid precipitation. Starting from March, snowmelt decreases  $SWE_{mod}$ , and thus  $TWS_{mod}$ , progressively. Thereby  $TWS_{mod}$  declines with some delay, as positive  $W$  anomalies in April and May suggest buffering of melt water in the soil and on the surface before being transferred to runoff or evapotranspired. During the summer months, snow is absent, while  $W$  decreases due to higher summertime evapotranspiration, and preceding runoff of temporarily stored water. With  $W$  and  $SWE_{mod}$  being at their minimum in August–September,  $TWS_{mod}$  reaches its minimum,



**Figure 6.** Spatially averaged mean seasonal cycle (MSC) of the period 2000–2014 as well as inter-annual variability (IAV, difference between monthly values and the MSC) for modelled TWS, SWE, and  $W$  anomalies to the time-mean of 2000–2014.

too, before starting to increase again in September–October. This rise relates to dropping evapotranspiration rates in combination with more precipitation input (increasing  $W$ ) and beginning snow accumulation (increasing  $SWE_{\text{mod}}$ ). Despite the interplay of  $SWE_{\text{mod}}$  and  $W$  on seasonal variations of the integrated  $TWS_{\text{mod}}$ , the amplitude of  $W$  (62 mm) is considerably lower than the one of  $SWE_{\text{mod}}$  (92 mm) and  $TWS_{\text{mod}}$  (144 mm). Thus, the seasonal accumulation of snow largely determines the magnitude of  $TWS_{\text{mod}}$ . Additionally,  $W$  anomalies at least partly result from snowmelt, whereas liquid water does not influence the snowpack. Thus, we conclude that average seasonal TWS variations in northern mid- to high latitudes are mainly driven by annual snow accumulation and ablation processes. The CR (Eq. 3) based on the spatially averaged MSC underlines this, as  $CR = -0.26$  indicates that variations in  $SWE_{\text{mod}}$  explain 63 % of seasonal  $TWS_{\text{mod}}$  variability ( $CW = 37\%$ ). This agrees with previous findings of Güntner et al. (2007), who found that SWE contributes to 68 % of seasonal TWS variations in cold regions using the WaterGAP model.

On IAV scales, the pattern seems less homogeneous (Fig. 6). In contrast to the MSC,  $CR = 0.25$  suggests a larger influence from liquid water anomalies ( $CW = 62.5\%$ ) than snow anomalies on inter-annual TWS variations. Thereby, we found the main contributor to  $TWS_{\text{mod}}$  anomalies being dependent on the phase of previous precipitation anomalies, in that they define whether snowfall anomalies lead to anomalies in the  $SWE_{\text{mod}}$ , or whether rain anomalies directly influence  $W$ . Additionally, precipitation input shows larger inter-annual variability than evapotranspiration or runoff losses, and thus dominates the change in water storages on IAV scales (not shown). Large  $TWS_{\text{mod}}$  anomalies, such as in 2005, 2010, and 2012, follow anomalies in wintertime precipitation and go along with anomalies in  $SWE_{\text{mod}}$  (Fig. 6). On the contrary, summertime anomalies related to  $W$  are usually less pronounced in their magnitude (e.g. 2003, 2006). We attribute this to the accumulating effects of snow stor-

age anomalies over the cold period, as they integrate all anomalies of previous cold months while the impact of evapotranspiration and runoff is reduced. Accordingly, the largest  $TWS_{\text{mod}}$  anomalies are obtained in early spring before snowmelt starts and when snow accumulation is highest. Nevertheless, since  $W$  is influenced by the quantity of snowmelt, anomalies in  $SWE_{\text{mod}}$  implicate subsequent changes in  $W$ . Additionally, snowpack anomalies are eliminated each summer, while  $W$  anomalies dominate the summer. As a result,  $W$  anomalies in any case affect  $TWS_{\text{mod}}$  variability on IAV scales when analysing the spatial average composition.

Besides, Güntner et al. (2007) demonstrated a shift from short-term storages with high seasonality such as SWE on MSC scales towards storages with longer delay times on IAV scales. Although modelled  $W$  mainly represents soil moisture, it implicitly includes surface water and groundwater storages. Thus, its contribution of  $CW = 62.5\%$  to inter-annual TWS variations is roughly comparable to 55 % contribution from soil moisture (27 %) and surface water (28 %) in cold regions, found by Güntner et al. (2007). Despite this, the relatively large influence of surface water bodies shown by Güntner et al. (2007) suggests that the lack of explicit surface water storages in this study may contribute to the remaining discrepancies with GRACE and the lower magnitude of modelled inter-annual TWS variability compared to GRACE estimates.

### 3.3.2 Local grid scale

Based on CR (Eq. 3), Fig. 7 shows the relative contribution of  $SWE_{\text{mod}}$  and  $W$  variances to total  $TWS_{\text{mod}}$  variability on MSC and IAV timescales for each grid. Thereby, blue colours represent prevailing  $SWE_{\text{mod}}$  variations as indicated by  $CR < 0$ , while red colours show the dominance of variations in  $W$  ( $CR > 0$ ).

Accordingly, variations in the MSC of  $TWS_{\text{mod}}$  are mainly influenced by snow in northern regions, with the mean  $CR = -0.30$  indicating that on average 65 % of seasonal  $TWS_{\text{mod}}$  variability can be explained by  $SWE_{\text{mod}}$  ( $CW = 35\%$ ) (Fig. 7a). The contribution of the variation in liquid water in general increases southwards and prevails seasonal  $TWS_{\text{mod}}$  variability south of approximately  $50^\circ$  latitude. An exception is Europe, where the influence of  $W$  reaches up to  $60^\circ$  latitude, and where the transition to snow-dominated regions is more gradual. Since the calculated variations in  $RW$  are low, the majority of modelled  $W$  represents variability in  $SM$ .

This obtained pattern confirms earlier studies that showed the dominance of snow in northern latitudes in North America (Rangelova et al., 2007), and prevailing soil moisture dynamics further south, e.g. in the Mississippi basin (Ngo-Duc et al., 2007; Güntner et al., 2007). Besides, the north extent of dominating  $W$  reflects the temperature gradient in North America and Eurasia. Comparison with average annual temperature suggests that for  $T > 10^\circ\text{C}$  variability of  $W$  dominates, while for  $T < 0^\circ\text{C}$   $SWE_{\text{mod}}$  dynamics prevail. This is plausible, as temperature determines annual snow accumulation, and the relative contribution of liquid water increases in the absence of snow. Yet, it further highlights the dependency on the temperature dataset used, especially in a model that calculates snowfall and snowmelt based on temperature thresholds as ours does.

Opposed to the MSC, the variability of  $W$  dominates  $TWS_{\text{mod}}$  variations on IAV scales in the entire study domain, as is clearly indicated by average  $CR = 0.63$  (Fig. 7b). Inter-annual variations of  $SWE_{\text{mod}}$  seem to be relevant only in regions that receive the highest annual snow amounts, such as the Canadian Arctic Archipelago, the northern west coast of North America, north-eastern Siberia, and the northern West Siberian Plain. Due to a prolonged cold period there, the time span with rainfall and evapotranspiration is short, decreasing the occurrence of potential variability in  $W$ . However, even in these regions the influence of  $SWE_{\text{mod}}$  is low compared to the MSC. This reduced importance of snow on inter-annual scales again agrees with the findings of Güntner et al. (2007).

Apart from that, and since we already showed a link between average  $TWS_{\text{mod}}$  IAV and previous precipitation anomalies, and as precipitation represents the main model forcing data, we investigated the relative contribution of rainfall and snowfall to inter-annual variability of total precipitation (Fig. 8). Similar to the composition of  $TWS_{\text{mod}}$  on local scale, rain anomalies prevail for most of the grid cells (mean  $CR = 0.68$ ). This is consistent when snowfall is not scaled by  $p_{\text{sf}}$  and thus suggests that the greater contribution of  $W$  to inter-annual variations of  $TWS_{\text{mod}}$  on local scale relates to highly variable (liquid) summertime precipitation events. On the contrary, monthly snowfall seems less variable between years, resulting in less pronounced variations in  $SWE_{\text{mod}}$  compared to  $W$ . Exceptions are regions of high maximum  $SWE_{\text{mod}}$ , which accordingly show a considerable

relative contribution of snow to the inter-annual  $TWS_{\text{mod}}$  variability.

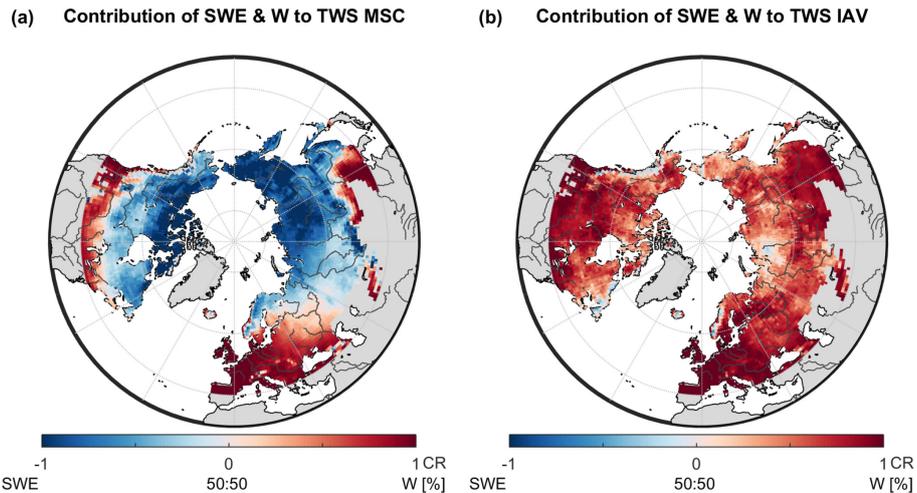
### 3.3.3 Comparison of different scales

Figure 9 summarizes the above-presented contributions to  $TWS_{\text{mod}}$  variability across spatial and temporal scales. As explained in the previous sections, we obtained two scale-dependent differences in the relative contribution to  $TWS_{\text{mod}}$  variability: (1) in general between temporal scales, and (2) for inter-annual variability between spatial scales.

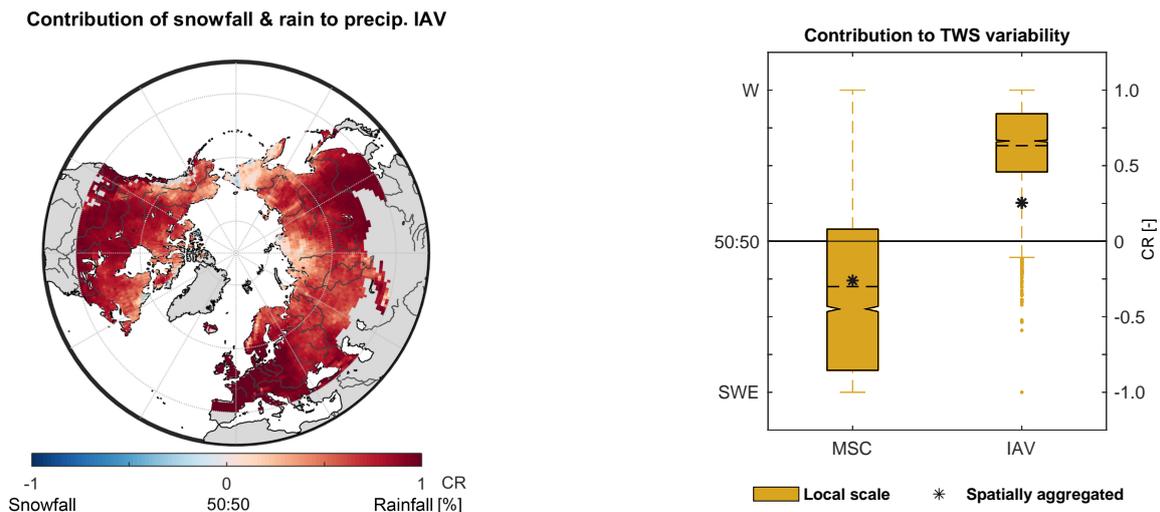
Regarding (1), Fig. 9 emphasizes again that seasonal variations of  $TWS_{\text{mod}}$  are mostly determined by seasonal snow dynamics, while on inter-annual scales  $TWS_{\text{mod}}$  variability mainly originates from variations in liquid water. As previously stated, determination by  $SWE_{\text{mod}}$  dynamics on MSC scales relates to the pronounced magnitude of seasonal snow variations in northern mid- to high latitudes. In comparison, average monthly changes in  $W$  were found to be minor and additionally influenced by snow ablation. Thereby, the spatially integrated  $CR$  (black star) roughly agrees with the average of local contributions (dashed line).

Concerning IAV scales, we found that the determination of  $TWS_{\text{mod}}$  variability by  $W$  relates to larger inter-annual variations in liquid precipitation compared to snowfall. However, considerable differences between spatial scales exist (Fig. 9). Opposed to the MSC, the spatially integrated  $CR$  (black star) for IAV is not within the interquartile range of local  $CR$ . This indicates a relatively larger effect of  $SWE_{\text{mod}}$  variations when looking on the spatially integrated values compared to local values. Liquid water storages are highly heterogeneous in space, mainly due to heterogeneous rainfall anomalies. On the contrary, snow variability is affected by fewer factors, and mainly regulated by a range of temperature values that control freezing and melting. Temperature anomalies in turn show sizeable spatial coherence across large areas. To assess the spatial coherence of  $W$  compared to  $SWE_{\text{mod}}$ , we calculated the proportion of total positive and total negative covariances among grid cells (Fig. 10).

For inter-annual variations of  $SWE_{\text{mod}}$ , the sum of positive covariances prevails (Fig. 10a), whereas positive and negative values are more in balance for  $W$  (Fig. 10b). This suggests that  $SWE_{\text{mod}}$  anomalies are more spatially coherent than anomalies of  $W$ . Thus, when spatially averaging, the more homogeneous snow anomaly patterns gain importance. On the contrary, opposed anomalies of  $W$  compensate for each other more strongly. This leads to a relatively larger influence of  $SWE_{\text{mod}}$  to the spatially integrated inter-annual  $TWS_{\text{mod}}$  variability compared to when analysing the local grid scale. Since positive covariation clearly dominates for temperature anomalies, the spatial coherence of  $SWE_{\text{mod}}$  relates to their homogeneous patterns (Fig. 10c). Similar to  $W$ , positive covariances only slightly outweigh year-to-year variations in rainfall (Fig. 10d). The same is true for snowfall (not shown). Therefore, the spatial coherence of  $SWE_{\text{mod}}$



**Figure 7.** Relative contribution based on CR (Eq. 3) of modelled snow (SWE) and liquid water ( $W$ ) storage anomalies to (a) mean seasonal variations from 2000 to 2014 of modelled TWS anomalies, and (b) inter-annual variations of modelled TWS anomalies for each grid cell of the study domain.



**Figure 8.** Relative contribution based on CR (Eq. 3) of modelled snowfall and rainfall to total precipitation ( $P$ ) anomalies on inter-annual (IAV) scales for each grid cell of the study domain.

**Figure 9.** Relative contribution of snow (SWE) and liquid water ( $W$ ) to TWS variability on different spatial (local grid scale, spatially integrated) and temporal (mean seasonal MSC, inter-annual IAV) scales based on CR (Eq. 3). The box plots represent the distribution of grid cell CR, with the dashed line marking the corresponding average. The star represents the CR calculated for the spatially integrated values.

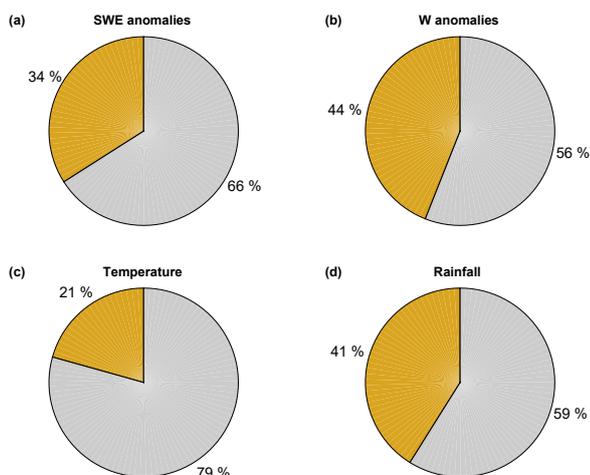
anomalies is less pronounced than for temperature, as snow is additionally influenced by snowfall anomalies. Regarding  $W$  anomalies, this indicates that the spatial heterogeneity in our model, which misses explicit information on soils, topography, etc., mainly results from inhomogeneous patterns in rainfall anomalies. Thereby, the slightly more balanced positive and negative covariations for  $W$  compared to rainfall can be ascribed to the additional impact of primarily radiation-driven evapotranspiration to SM.

In order to ensure that these results are not artificially caused by the forcing data, we did the same analysis running the model with rain and snowfall estimates of the WFDEI

product (Weedon et al., 2014). Since we observed the same patterns, we assume our findings to be robust (Sect. S7.1).

### 3.4 Limitations of the approach

Although the model of this study reproduces observed hydrological patterns well and achieves comparable results to state-of-the-art models, its low complexity and the applied



**Figure 10.** Proportion of total positive (grey) and negative (orange) covariances among grid cells for inter-annual variations of (a) snow (SWE), (b) liquid water storages ( $W$ ), (c) temperature, and (d) rainfall.

calibration approach are associated with limitations in terms of process understanding and predictive power.

First of all, the simple structure only allows inferences on represented processes, which likely include effects of fluxes and storages not considered explicitly. For example, the model does not resolve individual liquid water storages such as deep groundwater and surface water explicitly. As discussed previously, our delayed land runoff comprises various (intermediate) storages and delay times, and thus cannot be associated with one distinct storage component. Even though soil moisture is distinguished from these slowly varying reservoirs, its quantity and pattern have not been directly validated. Future research is required to increase confidence by including remote-sensing-based data of soil moisture in calibration and/or validation. However, these satellite data still have limited value as the microwave signals can only capture moisture in the upper 5 cm of soil and do not provide estimates under snow cover and dense vegetation (Döll et al., 2015; Lettenmaier et al., 2015). Therefore, a multi-layer soil scheme is needed to compare model outputs to satellite-derived soil moisture estimates, as was successfully demonstrated by Albergel et al. (2017) for example.

Further, the model does not include any human-induced changes in water storages, which yet contribute to observed TWS variability in many regions (Döll et al., 2015; Rodell et al., 2015). Other simplified or ignored hydrological processes include the coincident occurrence of rain and snowfall, liquid water capacity of snow, interception, freeze–thaw dynamics within the soil, capillary rise, and other surface–groundwater interactions, the effect of vegetation or other surface properties, and lateral flow from one grid cell to another. In the downstream areas of large basins especially, the latter repre-

sents a potential input that may significantly affect total TWS (Kim et al., 2009) and thus may contribute to the discrepancy between  $TWS_{obs}$  and  $TWS_{mod}$  in some regions. Besides, the model does not account for spatial variability of topography and land surface characteristics.

With regards to model parameters, we apply a global uniform parameter set and do not regionalize the parameters according to spatially distributed physio-geographical characteristics. In contrast, most macro-scale hydrological models include spatially distributed soil properties to define parameters related to infiltration, soil water holding capacity, and percolation, as well as vegetation types to assess the effects of different plant functional types on evapotranspiration and canopy storage (Sood and Smakhtin, 2015). Our model only implicitly considers the effects of vegetation, for example on ET, but not its spatial variability, as the associated impacts are included in the observational constraint. Spatial variability of model parameters might affect the relative contributions of different storage components to TWS variability at different spatial scales. However, the comparison with earthH2Observe models, which generally involve spatial heterogeneity in model parameters, suggests that the main conclusions remain unchanged. Additionally, we want to highlight that the spatial distribution of model parameters depends on assumptions and some degree of simplification as well and thus does not necessarily improve model performance compared to a global uniform parameter set obtained from multiple observational data. Further, as we encountered issues with parameter equifinality, especially between modelled snowmelt and sublimation, future efforts should include a stronger utilization of runoff data in the calibration and validation process. This would help to better constrain water fluxes to the atmosphere and liquid water fluxes, which can contribute to the runoff.

Finally, though the implemented cost function explicitly accounts for the uncertainty of the calibration data and additional uncertainties of other input data, their processing and characteristics remain partly unaddressed.

## 4 Conclusion

In this study, we assessed the relative contributions of snowpack versus soil and retained water variations to the variability of total terrestrial water storage (TWS) for northern mid- to high latitudes. To do so, we constrained a parsimonious hydrological model with multi-criteria calibration against multiple Earth observation data streams, including TWS from GRACE satellites and snowpack estimates from GlobSnow. The optimized model showed considerably good agreement with observed patterns of hydrological fluxes and states, and was found to perform comparably to or better than simulations from state-of-the-art macro-scale hydrological models. This underlines the potential of simple hydrological models tied to observational data streams as powerful

tools to diagnose and understand large-scale water cycle patterns. Further, it highlights the benefits of considering multiple complementary data constraints to overcome their individual shortcomings.

Consistent with previous studies, we found that seasonal TWS variations are dominated by the development of snowpacks during winter months in most places of the mid- to high northern latitudes. In contrast to this seasonal pattern, our study reveals that not snow but anomalies in liquid water storages, mainly comprising soil moisture, drive inter-annual TWS variations in almost the entire spatial domain. This counter-intuitive pattern was found to relate to larger rainfall anomalies compared to snowfall anomalies.

Apart from the timescale-dependent dominant controls on TWS variations, we additionally observed different behaviour across spatial scales. In terms of seasonal variations, the spatially integrated contribution reflects the average of the spatial domain. However, and more interestingly, the relative contribution of snowpack variations to total TWS inter-annual anomalies appears to be larger when spatially integrated than at local scale. We found this pattern results from stronger spatial coherence of snowpack anomalies compared to anomalies in other storages, such that the latter cancel out more strongly than the former when calculating an average across large spatial domains. The stronger spatial coherence of snowpack anomalies seems to be driven by the nature of spatially coherent temperature anomalies that determine snow accumulation and melt. These findings imply that patterns from large-scale integrated signals should not be associated with locally operating processes, since spatial covariations of climatic variables can confound the picture.

Overall, our study underlines the benefits of GRACE TWS as a model constraint, and moreover, stresses the importance of temporal and spatial scale when assessing the determinants of TWS variability. Clearly, insights obtained at one scale cannot be transferred to another, as is often (unintentionally) done. Hence, TWS variations in northern latitudes seem to be not merely subject to snow variability, but rather are driven by soil moisture on inter-annual scales – which may be of considerable importance when assessing long-term water availability in the context of climate changes.

**Data availability.** The hydrological model and all simulation results from this study can be obtained from the corresponding author on request. Due to the large size and quantity of the data, along with potentially few users of the “same” simulation results in the foreseeable future, it is not hosted on any public platform. The forcing and calibration data is publicly available as specified with the corresponding citation. For  $ET_{obs}$ , the subset of FLUXCOM (<http://www.fluxcom.org>, Jung and FLUXCOM team, 2016) data used in this study is available from the data portal of the Max Planck Institute for Biogeochemistry (<https://www.bgc-jena.mpg.de/geodb/projects/Home.php>) after registration.

**Supplement.** The supplement related to this article is available online at: <https://doi.org/10.5194/hess-22-4061-2018-supplement>.

**Author contributions.** TT, SK, and MJ designed the research in extensive collaboration with NC, AE, and MF. CN processed and integrated runoff estimates from E-RUN in model calibration. NC contributed to parameter estimation and uncertainty analysis. TT conducted the analysis. All co-authors contributed to the preparation of the paper.

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

**Special issue statement.** This article is part of the special issue “Integration of Earth observations and models for global water resource assessment”. It does not belong to a conference.

**Acknowledgements.** This research was carried out within the initiative for the development of the SINDBAD (Strategies to Integrate Data and Biogeochemical models) framework.

The article processing charges for this open-access publication were covered by the Max Planck Society.

Edited by: Jean-Christophe Calvet

Reviewed by: Vincent Humphrey and one anonymous referee

## References

- A, G., Wahr, J., and Zhong, S.: Computations of the viscoelastic response of a 3-D compressible Earth to surface loading: an application to Glacial Isostatic Adjustment in Antarctica and Canada, *Geophys. J. Int.*, 192, 557–572, <https://doi.org/10.1093/gji/ggs030>, 2013.
- Albergel, C., Munier, S., Leroux, D. J., Dewaele, H., Fairbairn, D., Barbu, A. L., Gelati, E., Dorigo, W., Faroux, S., Meurey, C., Le Moigne, P., Decharme, B., Mahfouf, J.-F., and Calvet, J.-C.: Sequential assimilation of satellite-derived vegetation and soil moisture products using SURFEX\_v8.0: LDAS-Monde assessment over the Euro-Mediterranean area, *Geosci. Model Dev.*, 10, 3889–3912, <https://doi.org/10.5194/gmd-10-3889-2017>, 2017.
- Alkama, R., Decharme, B., Douville, H., Becker, M., Cazenave, A., Sheffield, J., Voldoire, A., Tyteca, S., and Le Moigne, P.: Global Evaluation of the ISBA-TRIP Continental Hydrological System. Part I: Comparison to GRACE Terrestrial Water Storage Estimates and In Situ River Discharges, *J. Hydrometeorol.*, 11, 583–600, <https://doi.org/10.1175/2010jhm1211.1>, 2010.
- AMAP: Snow, Water, Ice and Permafrost in the Arctic (SWIPA) 2017, Arctic Monitoring and Assessment Programme (AMAP), Oslo, Norway, xiv+269, 2017.
- Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., and Betts, A. K.: A Revised Hydrology for the ECMWF Model: Verification from Field

- Site to Terrestrial Water Storage and Impact in the Integrated Forecast System, *J. Hydrometeorol.*, 10, 623–643, <https://doi.org/10.1175/2008jhm1068.1>, 2009.
- Bayer, P. and Finkel, M.: Optimization of concentration control by evolution strategies: Formulation, application, and assessment of remedial solutions, *Water Resour. Res.*, 43, W02410, <https://doi.org/10.1029/2005WR004753>, 2007.
- Beck, H. E., v. Dijk, A. I. J. M., d. Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A.: Global-scale regionalization of hydrologic model parameters, *Water Resour. Res.*, 52, 3599–3622, <https://doi.org/10.1002/2015WR018247>, 2016.
- Behrangi, A., Christensen, M., Richardson, M., Lebsack, M., Stephens, G., Huffman, G. J., Bolvin, D., Adler, R. F., Gardner, A., Lambriksen, B., and Fetzer, E.: Status of high-latitude precipitation estimates from observations and reanalyses, *J. Geophys. Res.-Atmos.*, 121, 4468–4486, <https://doi.org/10.1002/2015jd024546>, 2016.
- Bergström, S.: Principles and Confidence in Hydrological Modelling, *Nord. Hydrol.*, 22, 123–136, 1991.
- Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, *Geosci. Model Dev.*, 4, 677–699, <https://doi.org/10.5194/gmd-4-677-2011>, 2011.
- Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., Sudicky, E., Sutanudjaja, E. H., van de Giesen, N., Winsemius, H., and Wood, E. F.: Hyper-resolution global hydrological modelling: what is next?, *Hydrol. Process.*, 29, 310–320, <https://doi.org/10.1002/hyp.10391>, 2015.
- Breiman, L.: Random Forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/a:1010933404324>, 2001.
- Chen, X., Long, D., Hong, Y., Zeng, C., and Yan, D.: Improved modeling of snow and glacier melting by a progressive two-stage calibration strategy with GRACE and multisource data: How snow and glacier meltwater contributes to the runoff of the Upper Brahmaputra River basin?, *Water Resour. Res.*, 53, 2431–2466, <https://doi.org/10.1002/2016wr019656>, 2017.
- Cheng, M., Tapley, B. D., and Ries, J. C.: Deceleration in the Earth's oblateness, *J. Geophys. Res.-Solid Ea.*, 118, 740–747, <https://doi.org/10.1002/jgrb.50058>, 2013.
- Clark, D. B., Mercado, L. M., Sitch, S., Jones, C. D., Gedney, N., Best, M. J., Pryor, M., Rooney, G. G., Essery, R. L. H., Blyth, E., Boucher, O., Harding, R. J., Huntingford, C., and Cox, P. M.: The Joint UK Land Environment Simulator (JULES), model description – Part 2: Carbon fluxes and vegetation dynamics, *Geosci. Model Dev.*, 4, 701–722, <https://doi.org/10.5194/gmd-4-701-2011>, 2011.
- Decharme, B., Martin, E., and Faroux, S.: Reconciling soil thermal and hydrological lower boundary conditions in land surface models, *J. Geophys. Res.-Atmos.*, 118, 7819–7834, <https://doi.org/10.1002/jgrd.50631>, 2013.
- Derksen, C., Lemmetyinen, J., Toose, P., Silis, A., Pulliainen, J., and Sturm, M.: Physical properties of Arctic versus subarctic snow: Implications for high latitude passive microwave snow water equivalent retrievals, *J. Geophys. Res.-Atmos.*, 119, 7254–7270, <https://doi.org/10.1002/2013jd021264>, 2014.
- Döll, P., Mueller Schmied, H., Schuh, C., Portmann, F. T., and Eicker, A.: Global-scale assessment of groundwater depletion and related groundwater abstractions: Combining hydrological modeling with information from well observations and GRACE satellites, *Water Resour. Res.*, 50, 5698–5720, <https://doi.org/10.1002/2014wr015595>, 2014.
- Döll, P., Kaspar, F., and Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation, *J. Hydrol.*, 270, 105–134, 2002.
- Döll, P., Fiedler, K., and Zhang, J.: Global-scale analysis of river flow alterations due to water withdrawals and reservoirs, *Hydrol. Earth Syst. Sci.*, 13, 2413–2432, <https://doi.org/10.5194/hess-13-2413-2009>, 2009.
- Döll, P., Fritsche, M., Eicker, A., and Müller Schmied, H.: 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, 1311–1331, <https://doi.org/10.1007/s10712-014-9282-2>, 2014.
- Döll, P., Douville, H., Güntner, A., Müller Schmied, H., and Wada, Y.: Modelling Freshwater Resources at the Global Scale: Challenges and Prospects, *Surv. Geophys.*, 37, 195–221, <https://doi.org/10.1007/s10712-015-9343-1>, 2015.
- d’Orgeval, T., Polcher, J., and de Rosnay, P.: Sensitivity of the West African hydrological cycle in ORCHIDEE to infiltration processes, *Hydrol. Earth Syst. Sci.*, 12, 1387–1401, <https://doi.org/10.5194/hess-12-1387-2008>, 2008.
- Draper, N. and Smith, H.: Applied Regression Analysis, John Wiley, New York, 1981.
- Eichinger, W. E., Parlange, M. B., and Stricker, H.: On the concept of equilibrium evaporation and the value of the Priestly–Taylor coefficient, *Water Resour. Res.*, 32, 161–164, 1996.
- Eicker, A., Schumacher, M., Kusche, J., Döll, P., and Müller-Schmied, H.: Calibration/data assimilation approach for integrating GRACE data into the WaterGAP global hydrology model (WGHM) using an ensemble Kalman filter: First results, *Surv. Geophys.*, 35, 1285–1309, <https://doi.org/10.1007/s10712-014-9309-8>, 2014.
- Felfelani, F., Wada, Y., Longuevergne, L., and Pokhrel, Y. N.: Natural and human-induced terrestrial water storage change: A global analysis using hydrological models and GRACE, *J. Hydrol.*, 553, 105–118, <https://doi.org/10.1016/j.jhydrol.2017.07.048>, 2017.
- Feng, W., Zhong, M., Lemoine, J.-M., Biancale, R., Hsu, H.-T., and Xia, J.: Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements, *Water Resour. Res.*, 49, 2110–2118, <https://doi.org/10.1002/wrcr.20192>, 2013.
- Flörke, M., Kynast, E., Bärlund, I., Eisner, S., Wimmer, F., and Alcamo, J.: Domestic and industrial water uses of the past 60 years as a mirror of socio-economic development: A global simulation study, *Global Environ. Change*, 23, 144–156, <https://doi.org/10.1016/j.gloenvcha.2012.10.018>, 2013.
- Forman, B. A., Reichle, R. H., and Rodell, M.: Assimilation of terrestrial water storage from GRACE in a snow-dominated basin, *Water Resour. Res.*, 48, W01507, <https://doi.org/10.1029/2011wr011239>, 2012.

- Giroto, M., De Lannoy, G. J. M., Reichle, R. H., and Rodell, M.: Assimilation of gridded terrestrial water storage observations from GRACE into a land surface model, *Water Resour. Res.*, 52, 4164–4183, <https://doi.org/10.1002/2015wr018417>, 2016.
- Gudmundsson, L. and Seneviratne, S. I.: Observation-based gridded runoff estimates for Europe (E-RUN version 1.1), *Earth Syst. Sci. Data*, 8, 279–295, <https://doi.org/10.5194/essd-8-279-2016>, 2016.
- 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., Stuck, J., Werth, S., Döll, P., Verzano, K., and Merz, B.: A global analysis of temporal and spatial variations in continental water storage, *Water Resour. Res.*, 43, W05416, <https://doi.org/10.1029/2006WR005247>, 2007.
- Hansen, N. and Kern, S.: Evaluating the CMA Evolution Strategy on Multimodal Test Functions, in: *Parallel Problem Solving from Nature – PPSN VIII*, edited by: Yao, X., Burke, E., Lozano, J. A., Smith, J., Merelo-Guervós, J. J., Bullinaria, J. A., Rowe, J., Tino, P., Kabán, A., and Schwefel, H.-P., Springer, Berlin, 2004.
- He, Y., Bárdossy, A., and Zehe, E.: A review of regionalisation for continuous streamflow simulation, *Hydrol. Earth Syst. Sci.*, 15, 3539–3553, <https://doi.org/10.5194/hess-15-3539-2011>, 2011.
- Huffman, G. J. and Bolvin, D.: Version 1.2 GPCP One-Degree Daily Precipitation Data Set Documentation, NASA, Goddard Space Flight Center, Greenbelt, MD, USA <https://rda.ucar.edu/datasets/ds728.3/#!docs> (last access: July 2018), 2013.
- Huffman, G. J., Adler, R., Morrissey, M. M., Bolvin, D., Curtis, S., Joyce, R., McGavock, B., and Susskind, J.: Global Precipitation at One-Degree Resolution from Multisatellite Observations, *J. Hydrometeorol.*, 2, 36–50, 2000.
- Huffman, G. J., Bolvin, D. T., and Adler, R. F.: GPCP Version 1.2 One-Degree Daily Precipitation Data Set, Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, <https://doi.org/10.5065/D6D50K46>, 2016.
- Humphrey, V., Gudmundsson, L., and Seneviratne, S. I.: Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes, *Surv. Geophys.*, 37, 357–395, <https://doi.org/10.1007/s10712-016-9367-1>, 2016.
- IPCC: Climate Change 2014: Synthesis Report, in: *Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Geneva, Switzerland, 3–87, 2014.
- Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, 29, 2637–2649, 1993.
- Jung, M. and FLUXCOM team: FLUXCOM (RS+METEO) Global Land Carbon Fluxes using CRUNCEP climate data, FLUXCOM Data Portal, <https://www.bgc-jena.mpg.de/geodb/projects/Home.php>, 2016.
- Jung, M., Henkel, K., Herold, M., and Churkina, G.: Exploiting synergies of global land cover products for carbon cycle modeling, *Remote Sens. Environ.*, 101, 534–553, <https://doi.org/10.1016/j.rse.2006.01.020>, 2006.
- Jung, M., Reichstein, M., Schwalm, C. R., Huntingford, C., Sitch, S., Ahlstrom, A., Arneeth, A., Camps-Valls, G., Ciais, P., Friedlingstein, P., Gans, F., Ichii, K., Jain, A. K., Kato, E., Papale, D., Poulter, B., Raduly, B., Rodenbeck, C., Tramontana, G., Viovy, N., Wang, Y. P., Weber, U., Zaehle, S., and Zeng, N.: Compensatory water effects link yearly global land CO<sub>2</sub> sink changes to temperature, *Nature*, 541, 516–520, <https://doi.org/10.1038/nature20780>, 2017.
- Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., and Woollen, J.: The NCEP/NCAR 40-year reanalysis project, *B. Am. Meteorol. Soc.*, 77, 437–471, 1996.
- Kim, H., Yeh, P. J. F., Oki, T., and Kanae, S.: Role of rivers in the seasonal variations of terrestrial water storage over global basins, *Geophys. Res. Lett.*, 36, L17402, <https://doi.org/10.1029/2009GL039006>, 2009.
- Krinner, G., Viovy, N., de Noblet-Ducoudré, N., Ogé, J., Polcher, J., Friedlingstein, P., Ciais, P., Sitch, S., and Prentice, I. C.: A dynamic global vegetation model for studies of the coupled atmosphere-biosphere system, *Global Biogeochem. Cy.*, 19, GB1015, <https://doi.org/10.1029/2003GB002199>, 2005.
- Kug, J.-S., Jeong, J.-H., Jang, Y.-S., Kim, B.-M., Folland, C. K., Min, S.-K., and Son, S.-W.: Two distinct influences of Arctic warming on cold winters over North America and East Asia, *Nat. Geosci.*, 8, 759–763, <https://doi.org/10.1038/ngeo2517>, 2015.
- Kumar, S. V., Zaitchik, B. F., Peters-Lidard, C. D., Rodell, M., Reichle, R., Li, B., Jasinski, M., Mocko, D., Getirana, A., Lannoy, G. D., Cosh, M. H., Hain, C. R., Anderson, M., Arsenault, K. R., Xia, Y., and Ek, M.: Assimilation of Gridded GRACE Terrestrial Water Storage Estimates in the North American Land Data Assimilation System, *J. Hydrometeorol.*, 17, 1951–1972, <https://doi.org/10.1175/jhm-d-15-0157.1>, 2016.
- Kustas, W. P., Rango, A., and Uijlenhoet, R.: A simple energy budget algorithm for the snowmelt runoff model, *Water Resour. Res.*, 30, 1515–1527, 1994.
- Lettenmaier, D. P., Alsdorf, D., Dozier, J., Huffman, G. J., Pan, M., and Wood, E. F.: Inroads of remote sensing into hydrologic science during the WRR era, *Water Resour. Res.*, 51, 7309–7342, <https://doi.org/10.1002/2015wr017616>, 2015.
- Liu, J., Li, Z., Huang, L., and Tian, B.: Hemispheric-scale comparison of monthly passive microwave snow water equivalent products, *J. Appl. Remote Sens.*, 8, 084688, <https://doi.org/10.1117/1.JRS.8.084688>, 2014.
- Long, D., Longuevergne, L., and Scanlon, B. R.: Global analysis of approaches for deriving total water storage changes from GRACE satellites, *Water Resour. Res.*, 51, 2574–2594, <https://doi.org/10.1002/2014wr016853>, 2015.
- Luo, J., Pulliainen, J., Takala, M., Lemmetyinen, J., Kangwa, M., Eskelinen, M., Metsämäki, S., Solberg, R., Salberg, A.-B., Bippus, G., Ripper, E., Nagler, T., Derksen, C., Wiesmann, A., Wunderle, S., Hüsler, F., Fontana, F., and Foppa, N.: GlobSnow2 – Final Report, European Space Agency study contract report, Finnish Meteorological Institute, Helsinki, [http://www.globsnow.info/docs/GlobSnow\\_2\\_Final\\_Report\\_release.pdf](http://www.globsnow.info/docs/GlobSnow_2_Final_Report_release.pdf) (last access: July 2018), 2014.
- Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, *Hydrol. Earth Syst. Sci.*, 15, 453–469, <https://doi.org/10.5194/hess-15-453-2011>, 2011.
- Müller Schmied, H., Eisner, S., Franz, D., Wattenbach, M., Portmann, F. T., Flörke, M., and Döll, P.: Sensitivity of

- simulated global-scale freshwater fluxes and storages to input data, hydrological model structure, human water use and calibration, *Hydrol. Earth Syst. Sci.*, 18, 3511–3538, <https://doi.org/10.5194/hess-18-3511-2014>, 2014.
- Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models Part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, 1970.
- New, M., Hulme, M., and Jones, P.: Representing twentieth-century space-time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate, *J. Climate*, 13, 2217–2238, 2000.
- Ngo-Duc, T., Laval, K., Ramillien, G., Polcher, J., and Cazenave, A.: Validation of the land water storage simulated by Organising Carbon and Hydrology in Dynamic Ecosystems (ORCHIDEE) with Gravity Recovery and Climate Experiment (GRACE) data, *Water Resour. Res.*, 43, W04427, <https://doi.org/10.1029/2006WR004941>, 2007.
- Niu, G.-Y., Seo, K.-W., Yang, Z.-L., Wilson, C., Su, H., Chen, J., and Rodell, M.: Retrieving snow mass from GRACE terrestrial water storage change with a land surface model, *Geophys. Res. Lett.*, 34, L15704, <https://doi.org/10.1029/2007gl030413>, 2007.
- Omlin, M. and Reichert, P.: A comparison of techniques for the estimation of model prediction uncertainty, *Ecol. Model.*, 115, 45–59, [https://doi.org/10.1016/S0304-3800\(98\)00174-4](https://doi.org/10.1016/S0304-3800(98)00174-4), 1999.
- Orth, R., Koster, R. D., and Seneviratne, S. I.: Inferring Soil Moisture Memory from Streamflow Observations Using a Simple Water Balance Model, *J. Hydrometeorol.*, 14, 1773–1790, <https://doi.org/10.1175/jhm-d-12-099.1>, 2013.
- Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does model performance improve with complexity? A case study with three hydrological models, *J. Hydrol.*, 523, 147–159, <https://doi.org/10.1016/j.jhydrol.2015.01.044>, 2015.
- Priestley, C. H. B. and Taylor, R. J.: On the assessment of surface heat flux and evaporation using large-scale parameters, *Mon. Weather Rev.*, 100, 81–92, 1972.
- Ramillien, G., Lombard, A., Cazenave, A., Ivins, E. R., Llubes, M., Remy, F., and Biancale, R.: Interannual variations of the mass balance of the Antarctica and Greenland ice sheets from GRACE, *Global Planet. Change*, 53, 198–208, <https://doi.org/10.1016/j.gloplacha.2006.06.003>, 2006.
- Rangelova, E., van der Wal, W., Braun, A., Sideris, M. G., and Wu, P.: Analysis of Gravity Recovery and Climate Experiment time-variable mass redistribution signals over North America by means of principal component analysis, *J. Geophys. Res.-Ea. Surf.*, 112, F03002, <https://doi.org/10.1029/2006JF000615>, 2007.
- Rodell, M., Velicogna, I., and Famiglietti, J. S.: Satellite-based estimates of groundwater depletion in India, *Nature*, 460, 999–1002, <https://doi.org/10.1038/nature08238>, 2009.
- Rodell, M., Beaudoin, H. K., L'Ecuyer, T. S., Olson, W. S., Famiglietti, J. S., Houser, P. R., Adler, R., Bosilovich, M. G., Clayson, C. A., Chambers, D., Clark, E., Fetzer, E. J., Gao, X., Gu, G., Hilburn, K., Huffman, G. J., Lettenmaier, D. P., Liu, W. T., Robertson, F. R., Schlosser, C. A., Sheffield, J., and Wood, E. F.: The Observed State of the Water Cycle in the Early Twenty-First Century, *J. Climate*, 28, 8289–8318, <https://doi.org/10.1175/jcli-d-14-00555.1>, 2015.
- 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.
- Schmidt, R., Petrovic, S., Güntner, A., Barthelmes, F., Wunsch, J., and Kusche, J.: Periodic components of water storage changes from GRACE and global hydrology models, *J. Geophys. Res.-Solid Ea.*, 113, B08419, <https://doi.org/10.1029/2007JB005363>, 2008.
- Seo, K.-W., Ryu, D., Kim, B.-M., Waliser, D. E., Tian, B., and Eom, J.: GRACE and AMSR-E-based estimates of winter season solid precipitation accumulation in the Arctic drainage region, *J. Geophys. Res.*, 115, D20117, <https://doi.org/10.1029/2009jd013504>, 2010.
- Sood, A. and Smakhtin, V.: Global hydrological models: a review, *Hydrolog. Sci. J.*, 60, 549–565, <https://doi.org/10.1080/02626667.2014.950580>, 2015.
- Sorooshian, S., Duan, Q., and Gupta, V. K.: Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model, *Water Resour. Res.*, 29, 1185–1194, 1993.
- Stacke, T.: Development of a dynamical wetlands hydrology scheme and its application under different climate conditions, Max-Planck-Institute for Meteorology, Hamburg, 2011.
- Swenson, S.: Assessing High-Latitude Winter Precipitation from Global Precipitation Analyses Using GRACE, *J. Hydrometeorol.*, 11, 405–420, <https://doi.org/10.1175/2009jhm1194.1>, 2010.
- Swenson, S., Chambers, D., and Wahr, J.: Estimating geocenter variations from a combination of GRACE and ocean model output, *J. Geophys. Res.*, 113, B08410, <https://doi.org/10.1029/2007jb005338>, 2008.
- Syed, T. H., Famiglietti, J. S., and Chambers, D. P.: GRACE-Based Estimates of Terrestrial Freshwater Discharge from Basin to Continental Scales, *J. Hydrometeorol.*, 10, 22–40, <https://doi.org/10.1175/2008jhm993.1>, 2009.
- Takala, M., Luojus, K., Pulliainen, J., Derksen, C., Lemmetyinen, J., Kärnä, J.-P., Koskinen, J., and Bojkov, B.: Estimating northern hemisphere snow water equivalent for climate research through assimilation of space-borne radiometer data and ground-based measurements, *Remote Sens. Environ.*, 115, 3517–3529, <https://doi.org/10.1016/j.rse.2011.08.014>, 2011.
- Tallaksen, L. M., Burkhart, J. F., Stordal, F., Berntsen, T. K., Bryn, A., Eitzelmüller, B., Hagen, J. O. M., Hamran, S.-E., Halvorsen, R., Kääh, A., Kristjánsson, J. E., Krüger, K., Lande, T. S., Schuler, T. V., Westermann, S., Wisland, D., and Xu, C.-Y.: Land Atmosphere Interactions in Cold Environments (LATICE): The role of Atmosphere–Biosphere–Cryosphere–Hydrosphere interactions in a changing climate, EGU General Assembly, Vienna, Austria, 2015.
- Tapley, B. D., Bettadpur, S., Watkins, M., and Reigber, C.: The gravity recovery and climate experiment: Mission overview and early results, *Geophys. Res. Lett.*, 31, L09607, <https://doi.org/10.1029/2004gl019920>, 2004.
- Teuling, A. J., Seneviratne, S. I., Williams, C., and Troch, P. A.: Observed timescales of evapotranspiration response to soil moisture, *Geophys. Res. Lett.*, 33, L23403, <https://doi.org/10.1029/2006gl028178>, 2006.

- Tramontana, G., Jung, M., Schwalm, C. R., Ichii, K., Camps-Valls, G., Ráduly, B., Reichstein, M., Arain, M. A., Cescatti, A., Kiely, G., Merbold, L., Serrano-Ortiz, P., Sickert, S., Wolf, S., and Papale, D.: Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms, *Biogeosciences*, 13, 4291–4313, <https://doi.org/10.5194/bg-13-4291-2016>, 2016.
- Trischenko, A. P.: Removing Unwanted Fluctuations in the AVHRR Thermal Calibration Data Using Robust Techniques, *J. Atmos. Ocean. Tech.*, 19, 1939–1953, 2002.
- United Nations Environment: World atlas of desertification/UNEP, United Nations Environment Programme, Edward Arnold, London, Baltimore, 1992.
- van Beek, L. P. H., Wada, Y., and Bierkens, M. F. P.: Global monthly water stress: 1. Water balance and water availability, *Water Resour. Res.*, 47, W07517, <https://doi.org/10.1029/2010WR009791>, 2011.
- van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J., Ducharne, A., Cheruy, F., Viovy, N., Puma, M. J., Wada, Y., Li, W., Jia, B., Alessandri, A., Lawrence, D. M., Weedon, G. P., Ellis, R., Hagemann, S., Mao, J., Flanner, M. G., Zampieri, M., Matera, S., Law, R. M., and Sheffield, J.: LS3MIP (v1.0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project – aims, setup and expected outcome, *Geosci. Model Dev.*, 9, 2809–2832, <https://doi.org/10.5194/gmd-9-2809-2016>, 2016.
- 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, *Int. J. Geogr. Inform. Sci.*, 24, 189–212, <https://doi.org/10.1080/13658810802549154>, 2010.
- van Dijk, A. I. J. M. and Warren, G.: The Australian Water Resources Assessment System, Technical Report 4, Landscape Model (version 0.5) Evaluation against observations, CSIRO, Water for a Healthy Country National Research Flagship, Clayton, Australia, 2010.
- van Dijk, A. I. J. M., Renzullo, L. J., Wada, Y., and Tregoning, P.: A global water cycle reanalysis (2003–2012) merging satellite gravimetry and altimetry observations with a hydrological multi-model ensemble, *Hydrol. Earth Syst. Sci.*, 18, 2955–2973, <https://doi.org/10.5194/hess-18-2955-2014>, 2014.
- Viovy, N.: CRU-NCEPv6.1 Dataset, <http://dods.extra.cea.fr/data/p529viov/cruncep/>, last access: 15 September 2015.
- Wada, Y., Wisser, D., and Bierkens, M. F. P.: Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources, *Earth Syst. Dynam.*, 5, 15–40, <https://doi.org/10.5194/esd-5-15-2014>, 2014.
- Wahr, J., Swenson, S., Zlotnicki, V., and Velicogna, I.: Time-variable gravity from GRACE: First results, *Geophys. Res. Lett.*, 31, L11501, <https://doi.org/10.1029/2004gl019779>, 2004.
- Watkins, M. M., Wiese, D. N., Yuan, D. N., Boening, C., and Landerer, F. W.: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons, *J. Geophys. Res.*, 120, 2648–2671, 2015.
- 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.
- Werth, S., Güntner, A., Petrovic, S., and Schmidt, R.: Integration of GRACE mass variations into a global hydrological model, *Earth Planet. Sc. Lett.*, 277, 166–173, <https://doi.org/10.1016/j.epsl.2008.10.021>, 2009.
- Wielicki, B. A., Barkstrom, B. R., Harrison, E. F., Lee, R. B. I., Smith, L. G., and Cooper, J. E.: Clouds and the Earth's Radiant Energy System (CERES): An Earth Observing System Experiment, *B. Am. Meteorol. Soc.*, 77, 853–868, 1996.
- Wiese, D. N.: GRACE monthly global water mass grids NETCDF RELEASE 5.0 Ver. 5.0 Mascon Ver. 2, PO.DAAC, CA, USA, 2015.
- Wiese, D. N., Landerer, F. W., and Watkins, M. M.: Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution, *Water Resour. Res.*, 52, 7490–7502, <https://doi.org/10.1002/2016wr019344>, 2016a.
- Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W., and Watkins, M. M.: JPL GRACE Mascon Ocean, Ice, and Hydrology Equivalent Water Height RL05M.1 CRI Filtered Version 2, Ver. 2, PO.DAAC, CA, USA, <http://dx.doi.org/10.5067/TEMSC-2LCR5>, 2016b.
- Xie, H., Longuevergne, L., Ringler, C., and Scanlon, B. R.: Calibration and evaluation of a semi-distributed watershed model of Sub-Saharan Africa using GRACE data, *Hydrol. Earth Syst. Sci.*, 16, 3083–3099, <https://doi.org/10.5194/hess-16-3083-2012>, 2012.
- Zhang, L., Dobslaw, H., Stacked, 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.