https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

A physically guided deep learning reconstruction of terrestrial water storage anomalies at 0.1° across China

Xueying Li [1*], Yan Sun [2*], Xihui Gu [3], Niko Wanders [4], Bridget R. Scanlon [5], and Louise J. Slater [1]

1. School of Geography and the Environment, University of Oxford, Oxford, UK

2. School of Computer Science, University of Sydney, Sydney, Australia

3. School of Geography and Information Engineering, China University of Geosciences, Wuhan, China

4. Department of Physical Geography, Utrecht University, Utrecht, The Netherlands

5. Bureau of Economic Geology, Jackson School of Geosciences, The University of Texas at Austin, Austin, USA

Correspondence: Xueying Li (xueying.li@ouce.ox.ac.uk); Yan Sun (ysun9899@uni.sydney.edu.au)

Abstract

Terrestrial water storage (TWS), comprising all surface and subsurface water components, is a key indicator of water availability. The Gravity Recovery and Climate Experiment (GRACE) satellite mission provides large-scale estimates of TWS anomalies (TWSA), but its coarse spatial resolution (3°, approximately 300 km) limits the analysis of hydrologic processes at sub-regional scales. Using a physically-guided deep learning framework, we downscale TWSA from the original 3° GRACE mascons to 0.1° (approximately 10 km) across China, generating a standard version (2002–2019) with comprehensive observations used for model constraints and independent evaluation and an extended version (2020–2023) to support more recent hydrologic analyses. The downscaled TWSA preserves large-scale GRACE signals at the 3° grid scale (median correlation coefficient (*CC*): 0.95; root-mean-square error (*RMSE*): 1.38 cm) and basin scale (median *CC*: 0.94; *RMSE*: 1.72 cm), with a low median uncertainty (0.88 cm) across China. Its reliability is supported by high consistency with physically informed TWSA spatial patterns at the 0.1° resolution (median *CC*: 0.91) and internally consistent water balance closure beyond the native GRACE resolution (median *CC*: 0.80; *RMSE*: 1.44 cm). Evaluation against independent observations demonstrates that the downscaled TWSA agrees well with groundwater variations in intensively irrigated regions (*CC*: 0.65 for irrigation intensity > 50%) and annual glacier elevation change in cryospheric areas (*CC*: 0.97). The datasets improve fine-scale characterization of TWS variability and associated hydrologic processes in China, and can be used as a reference for evaluating performance of high-resolution hydrologic models. The two versions of the dataset are available at https://doi.org/10.5281/zenodo.19502906.

1 Introduction

Terrestrial water storage (TWS), the sum of all surface and subsurface storage components, is a crucial determinant of the global water budget (Immerzeel et al., 2020; Scanlon et al., 2023). TWS change plays a key role in determining water availability (Rodell et al., 2018) and modulating water flux interactions (Tapley et al., 2019), and is closely linked to hydro-climatic extremes such as droughts (Long et al., 2013; Pokhrel et al., 2021) and floods (Reager et al., 2014; Sun et al., 2024). Over the past two decades, TWS has changed substantially in response to climate change/variability (Guan et al., 2023; Li et al., 2022; Rodell et al., 2018) and accelerating human pressures (Döll et al., 2014; Gu et al., 2019; Guan et al., 2025). Monitoring TWS variations across different spatial scales is essential to understand the response of the terrestrial water system to a changing environment, providing valuable guidance for water resources management and climate change adaptation.

Despite its importance, TWS remains understudied relative to hydrologic fluxes, such as precipitation, evapotranspiration (ET), and streamflow, primarily because TWS represents an integration of different storage components and is difficult to measure directly in situ, particularly over large spatial domains. Although hydrologic and land surface models simulate some water storage components (e.g., soil moisture), they face challenges in explicitly resolving all TWS components and their interactions, particularly for groundwater and glacier processes (Scanlon et al., 2018; Tiwari et al., 2025). Recent advances in satellite gravimetry have provided an unprecedented opportunity to quantify TWS changes

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

at large spatial scales, with availability of 20+ years of TWS anomaly (TWSA) observations from the Gravity Recovery and Climate Experiment (GRACE) and GRACE Follow-On (GRACE-FO) missions (Tapley et al., 2019). Using GRACE and GRACE-FO retrievals, changes in TWS have been quantified from regional to global scales (Li et al., 2022; Rodell et al., 2018; Scanlon et al., 2023), constraining the estimation of individual storage components (Li et al., 2019a; Zhao et al., 2022), closure of the water balance (Li et al., 2019b), and calibration of model parameters (Bai et al., 2018; Chen et al., 2017). However, an inherent disadvantage of the GRACE data lies in its coarse spatial resolution (3°; approximately 110,000 km^2 at the Equator), linked to the design of the satellite orbit and accuracy of the instruments (Rowlands et al., 2005; Wahr et al., 2006). Although post-processing (e.g., regularization algorithms) can be used to generate GRACE Level-3 products at nominal resolutions of 0.25°–1°, the adjacent grid cells in these Level-3 products are not statistically independent and thus do not represent an effective improvement in spatial resolution (Loomis et al., 2019; Save et al., 2016; Wiese et al., 2016). The coarse resolution severely limits its application of satellite-based TWSA for investigating hydrologic processes at sub-regional scales (Chen et al., 2022; Li et al., 2025), which are often more sensitive to human intervention and climate variability than regional to continental scales.

Frameworks of downscaling GRACE/-FO observations are mainly categorized into dynamic (model-based) and statistical (data-based) approaches (Pascal et al., 2022). Dynamic downscaling, such as data assimilation methods, integrates GRACE observations into hydrologic or land surface models (Eicker et al., 2014; Li et al., 2019a). Despite its physical interpretability, this approach entails substantial computational demands. More critically, dynamic downscaling largely relies on the structure, parameters, and process description of physical models, potentially distorting the original signals of GRACE measurements (Gerdener et al., 2023). Statistical downscaling frameworks are more widely adopted, using data-driven approaches such as machine learning (ML) to establish relationships between GRACE/-FO TWSA observations and selected predictors. To date, downscaled TWSA datasets have primarily been generated by regression-based ML algorithms, including multiple linear regression (Pascal et al., 2022), partial least squares regression (Vishwakarma et al., 2021), random forest (Arshad et al., 2024; Wang et al., 2024), support vector machine (Kalu et al., 2024; Yazdian et al., 2023), and long-short term memory (Li and Kusche, 2026). These approaches link coarse-resolution TWSA with aggregated predictors and subsequently infer fine-scale variability, assuming a consistent relationship between TWSA and predictors across different spatial scales. One limitation of this assumption is that spatially heterogeneous interactions among variables across scales are not explicitly represented, which limits model generalization in regions with complex hydrologic processes. A recent application of convolutional neural network (CNN) to TWSA downscaling (Gou and Soja, 2024) captures spatial structural dependencies and multi-scale feature interactions through end-to-end learning, incorporating both GRACE signals and physically informed TWSA patterns to constrain training processes.

Despite these efforts, current TWSA downscaling remains insufficient for sub-regional investigations, particularly in regions subject to strong climate variability, intensive human interventions, and heterogeneity in the land surface, such as China. Available downscaled TWSA products in China, derived from both global and regional ML models, are provided at resolutions of 0.25° (Li and Kusche, 2026) and 0.5° (Gou and Soja, 2024; Vishwakarma et al., 2021; Xiong et al., 2025). The spatial resolution of these products remains coarse for fine-scale analysis, particularly when compared with commonly used hydrologic models and reanalysis datasets, such as PCR-GLOBWB 2 (Sutanudjaja et al., 2018) and ERA5-Land (Muñoz-Sabater et al., 2021), both available at a finer resolution of 0.1° (approximately 10 km). In addition, glacier dynamics are not considered in most datasets, although glacier mass balance is a sensitive indicator of climate warming and a major contributor to TWS changes in high mountain regions of Southwest and Northwest China (Hugonnet et al., 2021; Li et al., 2022). Moreover, current downscaled TWSAs are primarily produced using climatic and hydrologic predictors, which may limit their representation in regions with intensive human-water interactions, such as highly irrigated areas relying on ground and surface water abstractions. Therefore, the relatively coarse spatial resolution, omission of glacier dynamics, and lack of anthropogenic predictors call for further improvements of TWSA downscaling in China, especially for sub-regional analysis.

Here we use a deep learning framework to downscale TWSA from the original 3° GRACE/-FO mascons ($TWSA_{GRACE}$) to a much higher spatial resolution of 0.1° ($TWSA_{DS}$), incorporating physically informed

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

TWSA patterns ($TWSA_{PI}$) jointly derived from hydrologic simulations and satellite altimetry. A CNN-based U-Net model was selected for multi-scale feature fusion, via its encoder-decoder architecture that enhances generalization by isolating high-resolution details. Twelve predictors were included to comprehensively characterize TWS variations, associated with the hydrologic cycle, land processes, and human-water interactions. We provide a standard dataset of monthly 0.1° TWSA across China for 2002–2019, a period with comprehensive observations available for robust quality assessment. The gap period between GRACE and GRACE-FO (July 2017–May 2018) is reconstructed using the U-Net deep learning model. An extended dataset (2020–2023) is also provided to support more recent hydrologic analyses, but due to limited data availability, glacier mass balance is not constrained by independent satellite observations during 2020–2023 in the current version (Li and Sun, 2026a). The downscaled 0.1° TWSA was evaluated through: (1) large-scale comparisons with original GRACE/-FO observations; (2) assessments against physically informed spatial patterns; (3) water balance closure across small to medium catchments beyond the native GRACE resolution; (4) validation across irrigated and glacierized regions using independent observations; and (5) comparisons with representative downscaling studies (Gou and Soja, 2024; Xiong et al., 2025). The new datasets are valuable not only for investigating sub-regional TWS changes and related hydrologic analyses in China, but also for improving performance of high-resolution hydrologic models and providing insights into fine-scale water resources management.

2 Study area and data processing

2.1 Study area

We generate high-resolution TWSA across China and surrounding regions (15°–57°N, 69°–138°E; Fig. 1). Covering approximately 9.6 million km², China encompasses strong gradients in climate, topography, and human activities (Liu et al., 2017; Piao et al., 2010). Nine major river basins are delineated, including the Songhua-Liaohe, Haihe, Yellow, Huaihe, Yangtze, Pearl, Southeast, Southwest, and Continental basins. Northern basins such as the Haihe and Yellow River basins are characterized by limited water resources and intensive groundwater abstraction (Feng et al., 2013), whereas southern basins including the Yangtze and Pearl are dominated by monsoon-driven precipitation and frequent flooding (Zhang et al., 2006). In addition, the Tibetan Plateau and surrounding high mountain regions act as major water sources for downstream basins, with glacier and snowpack contributing significantly to regional water storage variability (Li et al., 2022; Li et al., 2025; Yao et al., 2022).

Recent climate change/variability and accelerating human pressures have jointly altered TWS across China. Severe threats to water security include warming-driven glacier mass losses from high mountains (Li et al., 2022; Li et al., 2025; Yao et al., 2022) and large-scale groundwater depletion from intensive irrigation (Feng et al., 2013; Long et al., 2025; Long et al., 2020). These processes pose two major challenges for hydrologic analysis: cryospheric dynamics and intensive human-water interactions, which remain inadequately resolved. In this study, we selected the Southeastern Tibetan Plateau (SETP) and North China Plain (NCP) as two representative subregions for assessing downscaled TWSA based on independent observations. The SETP (28°–32°N, 92°–98°E) covers the Eastern Nyainqentanglha Ranges, Eastern Himalayas, and Western Hengduan Mountains, and features the largest maritime glaciers on the Tibetan Plateau with a mean elevation exceeding 4 km (Fig. 1). Glaciers over the SETP are mainly distributed in the Parlung Tsangpo River basin and adjacent areas, where glacier meltwater accounts for more than 50% of the total runoff in this region (Zhao et al., 2022). Under a warming climate, glaciers of the SETP have been shrinking for decades, leading to significant losses in TWS over this region.

In contrast to climate-driven TWS declines in the SETP, the NCP has emerged as a global hotspot of groundwater depletion, largely driven by extensive irrigation supported by groundwater abstraction (Scanlon et al., 2023). Here the NCP refers to the region bordered by the Yanshan Mountains in the north, the Taihang Mountains in the west, the Yellow River in the south, and the Bohai Gulf in the northeast (35°–40°N, 113°–120°E) (Yang et al., 2022). Reported groundwater levels in the NCP declined at a rate of 1–2 m/year in the late 20th century (Yang et al., 2021), leading to an estimated cumulative depletion of about 60 km^3 from the 1960s to 2008 (Cao et al., 2013). China's South-to-North Water Diversion Project, operating since 2014, has reduced the rate of groundwater depletion in the

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

NCP in recent years (Long et al., 2025; Long et al., 2020; Yang et al., 2022), but this region remains a critical hotspot for analysing TWS variability due to its long-term groundwater stress.

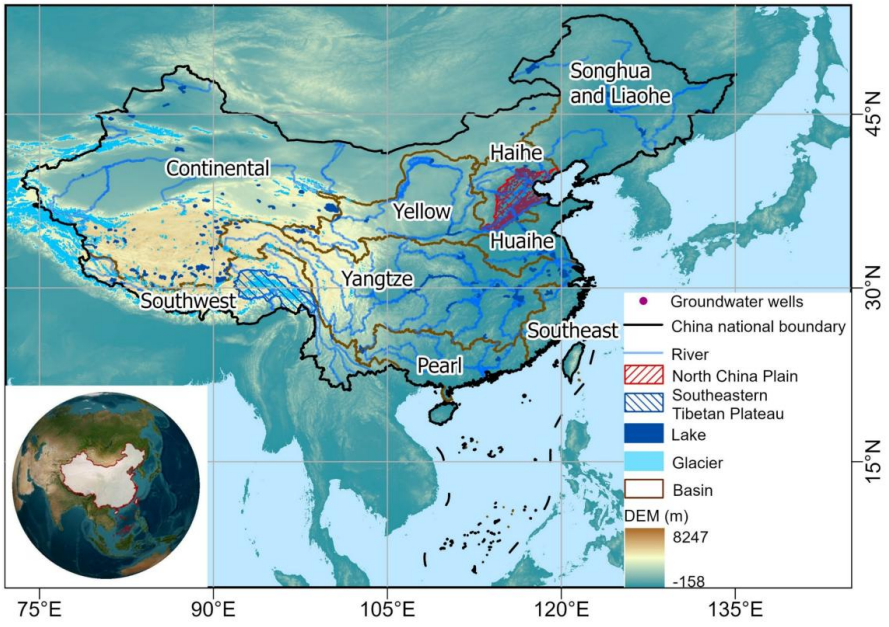

Fig. 1 China's national boundary, and the locations and elevations of major rivers, lakes, glaciers, and nine large river basins. Two representative subregions, the North China Plain and the Southeastern Tibetan Plateau, are marked, and locations of in-situ groundwater wells in the North China Plain are also shown.

2.2 Data processing

Datasets used in this study include labels and predictors of the U-Net model and ancillary data for model validation (Table 1). We combine the state-of-the-art datasets from multiple sources, including satellite remote sensing, hydrologic simulations, and reanalysis data to better represent the climatic and anthropogenic drivers underlying TWS variability.

Table 1 Information of data sets used in this study.

Variable	Product name / data accessibility	Original spatial resolution	Temporal resolution	Temporal coverage
	Coarse-resolution label of TWSA observation			
$TWSA_{GRACE}$	JPL-M	3°	Monthly	Apr 2002−Jun 2017; Jun 2018−present
	High-resolution label of TWSA spatial pattern			
$TWSA_{PI}$	PCR GLOBWB 2	5 arcmin	Monthly	1981−2024
	Hugonnet et al. (2021)	100 m	Monthly	2000−2019
	Dynamic predictors			
Precipitation	ERA5L	0.1°	Monthly	1950−present
Evapotranspiration	ERA5L	0.1°	Monthly	1950−present
Runoff	ERA5L	0.1°	Monthly	1950−present
Temperature	ERA5L	0.1°	Monthly	1950−present
NDVI	MOD13 A3	1 km	Monthly	2000−present
	Static predictors			
Irrigation intensity	FAO	5 arcmin	–	2005

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

Groundwater irrigation area	FAO	5 arcmin	–	2005
Population	GPWv4	5 arcmin	–	2000, 2005, 2010, 2015, 2020
Elevation	SRTM	1 arc-second	–	2000
Ancillary data sets				
Groundwater level	National Earth System Science Data Center	–	Monthly	2005−2018
Glacier mass balance	Zhao et al. (2022)	Region average	Annual	2003−2019

Note: Glacier mass balance estimates with broad spatial coverage are available up to 2019 and are therefore included only in the standard version of downscaled TWSA (2002–2019). In the extended period (2020–2023), the reconstructed TWSA primarily reflects variations in other water storage components (e.g., lakes, reservoirs, snow, soil moisture, and groundwater) simulated by PCR-GLOBWB 2 (Li and Sun, 2026a).

2.2.1 TWSA from GRACE and GRACE-FO mascon solution

We used the GRACE/-FO mascon solutions provided by NASA Jet Propulsion Laboratory (JPL) (Wiese et al., 2016) based on the latest Release Number 06 (RL06) version. Compared to previous spherical harmonic solutions, mascon solutions eliminate the need for post-processing and suffer less from leakage errors, and thus better preserve GRACE and GRACE-FO signals (Landerer and Swenson, 2012; Wiese et al., 2022). The JPL-M (mascon) provides monthly TWSA as anomalies relative to the 2004–2009 time-mean baseline. Missing monthly TWSA during the GRACE period (Jan 2003–Jun 2017) caused by instrument failure was linearly interpolated using the nearest two monthly estimates, whereas the missing data in the gap period between the GRACE and GRACE-FO missions (Jul 2017–May 2018) were reconstructed through the U-Net model. Several post-processing algorithms have been applied to the JPL-M to reprocess the original signals to provide user-friendly products; that is, applying the Coastal Resolution Improvement (CRI) filter to reduce land/ocean leakage errors and using gridded gain factors provided by the Community Land Model (CLM) to redistribute mass at a resolution of 0.5° within each 3° mascon element (Wiese et al., 2016). In this study, we use the JPL-M data sets without the land-grid scaling factors so that the satellite retrievals are entirely independent of hydrologic models. In addition, we used the block mean method to upgrade the JPL-M data from the nominal resolution of 0.5° into its native resolution 3°, aiming to reveal the effective resolution of GRACE/-FO satellites.

2.2.2 High-resolution spatial pattern of physically informed TWSA

Physically informed TWSA ($TWSA_{PI}$) was the sum of individual water storage components, including simulated storage anomalies derived from PCR-GLOBWB 2 (Sutanudjaja et al., 2018) and glacier anomalies derived from satellite altimetry (Hugonnet et al., 2021). Note that $TWSA_{PI}$ is used as a reference for the U-Net model to learn the pixel-scale structure (spatial distribution of high and low values), instead of constraining model outputs as ground truth.

PCR-GLOBWB 2 (Sutanudjaja et al., 2018) is a global hydrologic and water resources model that accounts for anthropogenic water use and management (e.g., irrigation, groundwater pumping, and reservoir regulation), and simulates surface (rivers, lakes, reservoirs, canopy, and snow) and subsurface (soil and groundwater) water storage. As one of the model outputs, monthly TWS simulated by PCR-GLOBWB 2 is the sum of these storage components, with high spatial resolution of 5 arcmin (approximately 0.0833°). We interpolated the 5-arcmin data to 0.1° resolution, and the model-simulated TWS was calculated as TWSA by subtracting the 2004–2009 time-mean baseline, to be consistent with GRACE data processing. Importantly, because PCR-GLOBWB 2 does not include data assimilation of GRACE measurements, the modelled TWSAs and GRACE TWSAs are entirely independent.

Due to the lack of glacier processes in PCR-GLOBWB 2 (Janzing et al., 2025), we incorporate multisource satellite altimetry retrievals to add glacier changes, obtained from the latest estimation of global glacier mass balance by Hugonnet et al. (2021). This glacier database provides monthly changes in glacier volume relative to the year 2000 at a spatial resolution of 100 m during 2000–2019. It was generated based on glacier inventories from Randolph Glacier Inventory 6.0 (RGI 6.0) (Pfeffer et al., 2014) and elevation changes from a large number of satellite archives, such as ASTER stereo images and ArcticDEM (Hugonnet et al., 2021). We first obtained monthly time series of cumulative changes in glacier volume relative to the year 2000, and divided them by glacier areas determined by RGI 6.0

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

to calculate cumulative changes in glacier thickness (Eq.1). Second, changes in glacier thickness were converted into liquid water equivalent (LWE) by multiplying the ratio of glacier density to water density (Eq.2). Finally, the water-equivalent contribution from glacier changes is area-weighted based on the area of each 0.1° grid cell, to obtain changes in glacier storage (expressed in water equivalent) at the mean of 0.1° grid cell (Eqs.3–4). For consistency with GRACE/-FO data, remote sensing-based glacier changes were calculated as anomalies by subtracting the 2004–2009 time-mean baseline.

$$h_g = \frac{V_g}{A_g} \tag{1}$$

$$LWE = h_g \cdot \frac{\rho_{glacier}}{\rho_{water}} \tag{2}$$

$$LWE' = LWE \cdot \frac{A_g}{A_{grid}} \tag{3}$$

$$A_{grid} = R^2 \cdot (\omega_2 - \omega_1)(\sin\varphi_2 - \sin\varphi_1) \tag{4}$$

where hg, Vg, and Ag represent thickness, volume, and area of glaciers, respectively; LWE indicates liquid water equivalent corresponding to ice thickness, while LWE' denotes area-weighted water equivalent of each 0.1° grid cell; ρ is average density, where average density of glaciers ($\rho_{glacier}$) is 850 kg/m^3 (Huss, 2013) and average water density (ρ_{water}) is 1000 kg/m^3; A_{grid} denotes area of each grid cell, computed based on a spherical Earth approximation; R is mean radius of the Earth (6,371 km); ω_1 and ω_2 are the western and eastern longitudinal boundaries of the grid cell (in radians); and φ_1 and φ_2 denote the southern and northern latitudinal boundaries (in radians).

By summing storage anomalies derived from PCR-GLOBWB 2 and glacier anomalies derived from satellite altimetry, our physically informed TWSA (TWSA$_{PI}$) explicitly accounts for multiple components of TWS change to ensure a comprehensive representation. The normalization relative to the same baseline period with GRACE processing (2004–2009) reduces potential inconsistency associated with different data sources and processing schemes. Given that physical models may introduce substantial noise (Gou and Soja, 2024), we applied an outlier filtering approach to TWSA$_{PI}$ to improve the training robustness. As training was conducted at a monthly time step (Section 3.2), the lower and upper outlier bounds were determined as the 1st and 99th percentiles of the data distribution within each monthly map, and excluded from the computation of loss functions.

2.2.3 Input predictors

We considered a total of twelve predictors from four categories to capture the combined effects of climatic and anthropogenic factors on TWS variations. First, precipitation, ET, and runoff are three critical hydrologic fluxes in the water balance and highly related to TWS variations. We obtained monthly estimates of these variables at a spatial resolution of 0.1° from the latest land component of the fifth generation of European ReAnalysis (ERA5) data, ERA5-Land (simplified as ERA5L) (Muñoz-Sabater et al., 2021), provided by the European Centre for Medium Range Forecasts (ECMWF).

Second, we considered cumulative anomalies of precipitation, ET, and runoff to characterize the memory effect of hydrologic fluxes on TWS variations. On the basis of the water balance principle (input fluxes – output fluxes = TWS changes), TWS is affected by accumulated flux variables in a given period (Eq.5). Cumulative anomalies of hydrologic fluxes represent long-term water surplus or deficit and have been shown to improve the performance of deep learning models in reconstructing TWS in previous studies (Li et al., 2022; Wang et al., 2024). In this study, flux variables (precipitation, ET, and runoff) were calculated as monthly anomalies relative to the 2004–2009 baseline, consistent with GRACE data processing, and then cumulatively integrated from a fixed starting point of April 2002.

$$\begin{aligned} TWS_n &= TWS_0 + \Delta TWS_1 + \Delta TWS_2 + \cdots + \Delta TWS_n \\ &= TWS_0 + \left(flux_{In,1} - flux_{Out,1}\right) + \left(flux_{In,2} - flux_{Out,2}\right) + \cdots + \left(flux_{In,n} - flux_{Out,n}\right) \\ &= TWS_0 + \left(flux_{In,1} + flux_{In,2} + \cdots + flux_{In,n}\right) - \left(flux_{Out,1} + flux_{Out,2} + \cdots + flux_{Out,n}\right) \end{aligned} \tag{5}$$

Third, temperature, vegetation dynamics, and topography were included to describe land surface processes. Temperature largely affects the energy budget and is a dominant driver of TWS changes over cryospheric regions (Li et al., 2022). We obtained 2 m air temperature from ERA5L datasets with 0.1°

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

resolution. Vegetation dynamics were represented by Normalized Difference Vegetation Index (NDVI), which was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS) vegetation index product (MOD13 A3) at the monthly scale and 1 km resolution. NDVI reflects the vegetation greenness and canopy density, closely linked to transpiration and soil moisture uptake. Topography was represented by the Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM) 1 Arc-Second Global data set (approximately 30 m resolution). The elevation plays a key role in regulating runoff pathways and groundwater recharge, further influencing the spatial variability of TWS.

Fourth, we incorporated predictors related to human interventions in the water cycle, including irrigation intensity (fraction of area equipped for irrigation), fraction of irrigated area supplied by groundwater, and population density. Irrigation information was derived from the latest version (v5.0) of the Global Map of Irrigation Areas from the Food and Agriculture Organization (FAO) at 5 arcmin (0.0833°). Population density serves as a proxy for domestic and municipal water demand, provided by the Gridded Population of the World dataset, Version 4 (GPWv4), with a resolution of 5 arcmin (0.0833°). This dataset is primarily based on national censuses and population registers, without extensive post-processing or ancillary data-driven refinement.

The native spatial resolutions of these twelve predictors are equal to or finer than the target downscaling resolution of 0.1°, thereby providing sufficient spatial details to learn high-resolution TWS patterns. Eight predictors are dynamic, with monthly values related to the water balance (precipitation, evapotranspiration, runoff, and their accumulations), energy budget (temperature), and vegetation (NDVI). In contrast, four predictors describing irrigation, population, and topography are treated as quasi-static factors due to their relatively slow temporal variability. Subject to data availability, we input the average state (annual average population during 2000–2020) or representative state (the irrigation map obtained in 2005 and DEM map in 2000) of these static predictors into the U-Net model as time-invariant input channels. All features were linearly normalized on the basis of their 1st percentiles and 99th percentiles to reduce the impacts of outliers.

2.2.4 Ancillary datasets

Ancillary datasets include in-situ observations of groundwater levels across the NCP and independent estimation of cumulative glacier surface elevation over the SETP. We collected observed groundwater level during 2005–2018 of 559 groundwater wells across the NCP, provided by the National Earth System Science Data Center, National Science and Technology Infrastructure of China. These groundwater wells provide monthly groundwater level variations with at least 12 months of records, which are compared with TWSA at the corresponding grid cells. Because we aim to evaluate the temporal variability of TWSA rather than absolute magnitude (Section 4.3), observed groundwater level variations are considered representative without the need for conversion to groundwater storage. Such conversion requires the estimation of specific yield, which is highly uncertain over the NCP and may introduce additional errors.

In addition, glacier estimates over the SETP were obtained from Zhao et al. (2022), based on satellite altimetry observations from ICESat and CryoSat-2, which are independent of those included in $TWSA_{PI}$. This dataset provides regionally averaged cumulative surface elevation over the SETP at seasonal and annual scales, and is used to evaluate the temporal variability of TWSA in this glacierized area.

3 Methodology

3.1 U-Net model structure

The input to the U-Net is a four-dimensional tensor of shape $B \times C \times 30 \times 30$, where B denotes the global batch size and C represents the number of features channels derived from input predictors. The global batch size was set to 512 ($B = 512$), with a total of 12 feature channels, including 8 dynamic and 4 static features ($C = 12$). All predictors were processed to 0.1° and grouped into 30×30 spatial patches, corresponding to the spatial extent of a single 3° GRACE grid cell (i.e., 3° = 30×0.1°). The network maps multi-channel input patches to a single-channel output at 0.1° resolution, representing the predicted TWSA at each grid cell. The U-Net model adopts an encoder-decoder architecture with symmetric skip connections to capture both large-scale contextual information and fine-scale spatial details (Fig. 2). The encoder progressively extracts hierarchical features through convolutional blocks

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

with downsampling, while the decoder restores spatial resolution via upsampling. The number of feature channels increases in the encoder and decreases in the decoder. Skip connections are applied at each stage to preserve spatial information. Detailed layer configurations, including channel dimensions and feature map sizes, are illustrated in Fig. 2. Finally, a 1 × 1 convolution layer maps the decoded features to a single-channel output, representing the predicted TWSA at 0.1° resolution.

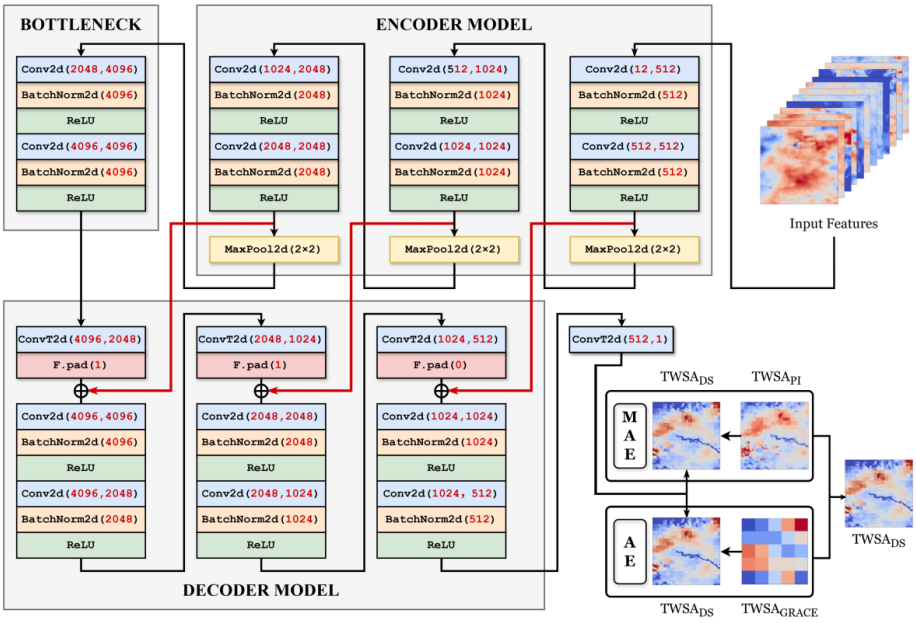

Fig. 2 The structure of U-Net model for TWSA downscaling. Multi-channel predictors at 0.1° resolution are mapped to high-resolution TWSA through an encoder-decoder framework with skip connections. The encoder and decoder models are both built upon a unified modular structure consisting of two convolutional layers. The red lines denote skip connections, which are used to fuse low-level features with high-level features. The symbol ⊕ denotes the concatenation of two tensors with the same shape along the channel dimension. Model training is constrained by a dual-loss design, enforcing agreement with GRACE TWSA at the patch scale and with physically informed patterns at the pixel scale.

3.2 Training strategy and uncertainty estimation

For model training, the study domain was represented as a 420 × 690 matrix of 0.1° pixels, corresponding to a 14 × 23 matrix of 3° grid cells (i.e., 30 × 30 pixels per grid cell), covering China and its surrounding regions (15°–57°N, 69°–138°E). Coastal GRACE grid cells with a land fraction larger than 20% were retained to ensure adequate terrestrial signal. The spatial domain was partitioned into 7 × 11 non-overlapping blocks. Each block consists of 60 × 60 pixels and corresponds to four effective GRACE grid cells, improving the stability of the patch-scale constraint. The easternmost 30 columns were excluded from block partitioning, as they contain very few valid terrestrial signals within China. At each monthly time step, the non-overlapping blocks were randomly partitioned into training (70%; 54 blocks) and validation (30%; 23 blocks) subsets. The split was conducted within each month rather than across time, as the objective was to learn spatial relationships rather than temporal forecasting. To increase the number of training samples and improve spatial continuity, a sliding window with a stride of 10 is applied to extract overlapping patches within the training and validation sets, respectively. Parameters are optimized using the Adam optimizer with a learning rate schedule to ensure stable convergence during training. The model is trained for 800 epochs, with the learning rate gradually reduced over time.

The training objective combines two complementary losses, which balance amplitude constraints from GRACE/-FO observations at large scales and spatial patterns from physically informed TWSA at fine

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

resolutions. We computed the loss of Absolute Error (AE) between the averaged GRACE TWSAs and averaged predicted TWSAs over each patch (corresponding to the GRACE original resolution of 3°), and then averaged all patches within a batch to obtain the aggregation error at the GRACE scale (Eq.6). To capture high-resolution spatial patterns, a loss of Mean Absolute Error (MAE) was additionally introduced at the pixel level of 0.1°. As for the MAE loss constrained by physically informed TWSA, we exclude invalid pixels, including missing values in the ocean and outliers of hydrologic simulations (1st and 99th percentiles of the data distribution within each monthly map). Because the number of valid pixels varies across patches, the MAE loss was computed as the average over all valid pixels across the batch (Eq.7).

$$L_{AE} = \frac{1}{B}\sum_{b=1}^{B} \left| \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i - \frac{1}{N}\sum_{i=1}^{N} y_{i,G} \right| \quad (6)$$

$$L_{MAE} = \frac{1}{\sum_{b=1}^{B} |\Omega_b|}\sum_{b=1}^{B}\sum_{i\in\Omega_b} |\hat{y}_i - y_{i,PI}| \quad (7)$$

where B denotes the batch size, which is set to 512 in this study; N represents the number of pixels within each patch; $\hat{y}_i$ indicates the predicted TWSA at pixel i; y_i represents the reference value at each pixel, corresponding to GRACE TWSA for the AE loss ($y_{i,G}$ in Eq.6) and physically informed TWSA for the MAE loss ($y_{i,PI}$ in Eq.7); and Ω represents the valid pixels in each patch with |Ω| denoting the number of valid values in this patch.

The AE loss constrains the spatially averaged TWSA within each 30 × 30 patch, thereby preserving the coarse-resolution signals from GRACE measurements. In contrast, the MAE loss is computed at the pixel level, and improves fine-scale spatial details by minimizing the absolute differences between predicted TWSA and the high-resolution $TWSA_{PI}$. The final loss function was defined as a weighted combination of the two losses constrained by GRACE TWSA at the patch level and physically informed TWSA at the pixel level (Eq.8).

$$L = \lambda_{AE} L_{AE} + \lambda_{MAE} L_{MAE} \quad (8)$$

Two parameters of λ_{AE} and λ_{MAE} control the trade-off between large-scale GRACE consistency and fine-scale spatial details. Increasing the ratio of λ_{AE} to λ_{MAE} forces the model to better match the GRACE statistics but may suppress fine-scale spatial variability. Conversely, decreasing this ratio enhances the model's ability to resolve local structural details but may weaken the consistency with GRACE-derived TWS variations if the fine-scale constraint becomes overly dominant.

In this study, we first calculated the average uncertainty of GRACE JPL-M across China (1.96 cm). Using this as a benchmark, we designed the loss-weighting strategy to ensure that the AE loss for both the training and validation sets converged to values below the average GRACE uncertainty across China. This criterion guarantees that the downscaled TWSA has comparable accuracy relative to GRACE observation at large scales. Within this constraint, we gradually increased the ratio of λ_{MAE} to λ_{AE} to improve the model's ability to resolve fine-scale spatial variability while preserving large-scale consistency. Through grid search over the combinations of [2.0,1.5,1.0] × [2.0,1.5,1.0], the optimal weighting was determined as λ_{AE} = 1.0 and λ_{MAE} = 1.0.

Uncertainty of downscaled TWSA was estimated from three independent training runs with different random seeds, capturing variability from initialization, data shuffling, and stochastic optimization. Despite identical training settings, stochasticity in the learning process introduces variability in model parameters and predictions. For each independent run, the trained model produced a complete set of downscaled TWSA fields, from which the evaluation metrics are computed. The final downscaled TWSA was generated by averaging TWSA outputs across the three runs, while the corresponding standard deviation was estimated as uncertainty. For each monthly map, those pixels with estimated uncertainty exceeding the 99th percentile of all pixels across China were identified as low-confidence predictions. These values were subsequently replaced using nearest-neighbour interpolation from surrounding valid grid cells to ensure spatial continuity.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

3.3 Closure of water balance and evaluation indices

The water balance serves as the water mass-conservation constraint at the catchment scale, demonstrating the relationship between TWS change (TWSC) and the flux integration of precipitation (*P*), evapotranspiration (*ET*), and runoff (*R*) (Eq.9).

$$TWSC = P - ET - R \tag{9}$$

To reduce potential high-frequency artifacts from the finite differences, both TWSC and hydrologic fluxes were smoothed, using Eq.10 and Eq.11, respectively (Gou and Soja, 2024; Landerer et al., 2010; Xiong et al., 2025).

$$TWSC(t) = \frac{TWSA(t+1) - TWSA(t-1)}{2\Delta t} \tag{10}$$

$$\tilde{X} = \frac{1}{4}X(t-1) + \frac{1}{2}X(t) + \frac{1}{4}X(t+1) \tag{11}$$

where t denotes the target month; Δt is the time interval taken as a month; and X represents the hydrologic flux of P, ET, or R. Water balance closure was used to assess whether the downscaled TWSA preserves mass conservation beyond the native GRACE resolution and captures physically consistent relationships among hydrologic fluxes. Precipitation, ET, and runoff were derived from ERA5L, consistent with the input predictors, to ensure methodological consistency. This analysis is intended to evaluate internal model consistency rather than provide an independent validation, and therefore, consistent datasets avoid additional uncertainties introduced by heterogeneous data sources. We compared TWS change derived from the storage and flux term (Eq.9) across 163 small to medium catchments in China. These basins are defined based on HydroBASINS Level 5 database, representing a medium spatial scale with the drainage area on the order of 10,000–100,000 km². We further excluded small basins (area < 10,000 km^2), desert regions (long-term annual mean precipitation < 200 mm), and basins strongly affected by intensive irrigation (irrigated area fraction > 5%) or cryospheric processes (glacier area fraction > 5%). This exclusion aims to minimize the influence of processes that are not explicitly accounted for in the water balance equation (e.g., additional storage loss due to human water use), thereby ensuring a more reliable assessment.

We selected two commonly used indices to evaluate the performance of downscaled TWSA, including the Pearson correlation coefficient (*CC*) (Eq.12) and root mean square error (*RMSE*) (Eq.13).

$$CC = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \sum_{i=1}^{n}(y_j - \bar{y})^2}} \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(x_j - y_j)^2}{n}} \tag{13}$$

where x_j represents the variable to be evaluated; y_j represents the corresponding variable taken as the baseline; an overbar denotes mean; and n is number of data pairs for evaluation.

Here *CC* was used to indicate the degree of coherence between the evaluated and reference datasets, while *RMSE* was employed to quantify the magnitude of discrepancies. The other index of Kling-Gupta efficiency (*KGE*) (Gupta et al., 2009; Kling et al., 2012) was not adopted, because it is less suitable for anomaly-based TWSA evaluation when the mean TWSA could approach zero. Specifically, *KGE* is a comprehensive indicator including correlation, ratio of means, and ratio of standard deviations of the two paired datasets. The value of ratio of means can be extremely high if its denominator (the mean of TWSA) is approaching to zero, further resulting in a very large negative value of *KGE* over the regions with relatively stable TWS changes.

4 Results

4.1 Performance of downscaled TWSA at large scales

The AE loss for both training and validation sets converged after approximately 400 epochs (Fig. 3a), representing the GRACE-based constraint. By the end of training (800 epochs), the AE loss converged to 0.84 cm for the training set and 1.50 cm for the validation set, both lower than the spatially averaged

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

uncertainty of GRACE JPL-M across China (1.96 cm). The MAE loss, associated with physically informed spatial patterns, converged more gradually and reached a stable level after approximately 600 epochs, with final values of 1.97 cm for the training set and 2.61 cm for the validation set, respectively. For both AE and MAE losses, the validation curves consistently follow the trends of the training curves without divergence, and the gap between them remains relatively stable after convergence, indicating stable model training with no evident overfitting. We upscaled the high-resolution TWSA from 0.1° to 3° using the block mean method, and find that compared to physically informed TWSA, the downscaled product substantially improved the agreement with GRACE/-FO observations (Fig. 3b). Across all 3° grid cells in China during 2002–2019, the coefficient of determination (R^2) increased from 0.34 ($TWSA_{PI}$ vs. $TWSA_{GRACE}$) to 0.70 ($TWSA_{DS}$ vs. $TWSA_{GRACE}$), indicating enhanced large-scale consistency after incorporating GRACE constraints.

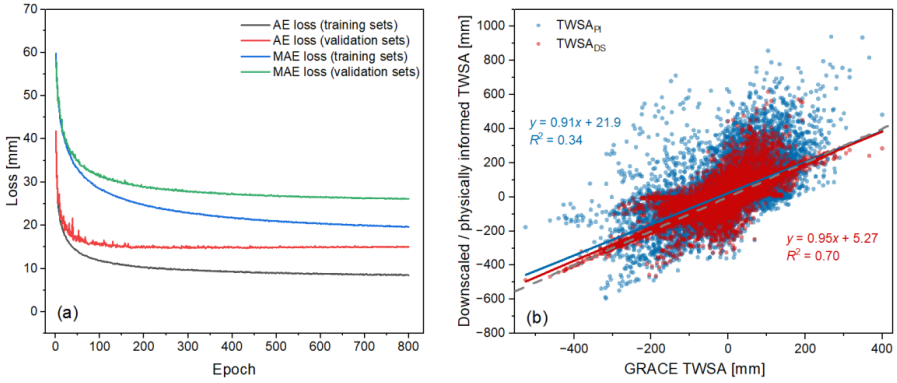

Fig. 3 (a) Loss curves of the multi-constraint U-Net model. The curves are shown separately for the AE loss and MAE loss, each including both training and validation sets. **(b)** Scatterplots between $TWSA_{GRACE}$ vs. $TWSA_{PI}$ (blue dots) and $TWSA_{GRACE}$ vs. $TWSA_{DS}$ (red dots) for all 3° grid cells in China during 2002–2019. Solid lines indicate linear regression fits, with the corresponding regression equations and coefficients of determination (R^2). Dots, regression lines, and corresponding statistics share consistent colors for each comparison. The dashed grey line indicates the 1:1 line.

We mapped the grid-level evaluation metrics across China to evaluate the ability of downscaled TWSA to preserve GRACE signals at the 3° resolution during the 2002–2019 period (Fig. 4). The close agreement between $TWSA_{DS}$ and $TWSA_{GRACE}$ is reflected by a high correlation (median *CC*: 0.95; Fig. 4a) and low discrepancy (median *RMSE*: 1.38 cm; Fig. 4b). Relatively large discrepancies of $TWSA_{DS}$ are mainly found in coastal areas, which may be attributed to both land-ocean signal leakage in GRACE observations and missing values over adjacent ocean areas in physical simulations. Nevertheless, over the majority of inland China, $TWSA_{DS}$ demonstrates consistently high agreement with GRACE/-FO observations. In comparison, the physically informed TWSA ($TWSA_{PI}$) exhibits much weaker agreement with GRACE, characterized by a lower median *CC* (0.66) (Fig. 4c) and larger *RMSE* (5.50 cm) (Fig. 4d). Low correlations between $TWSA_{PI}$ and $TWSA_{GRACE}$ are located across the arid and semi-arid regions in Northwest China, whereas large discrepancies are found in the humid Southeast China. These results show the limitation of model simulations to reveal TWSA in both terms of variability and magnitude at large scales. The improvements in both *CC* and *RMSE* highlight the effectiveness of incorporating GRACE-based constraints in enhancing large-scale statistical consistency with satellite observations.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

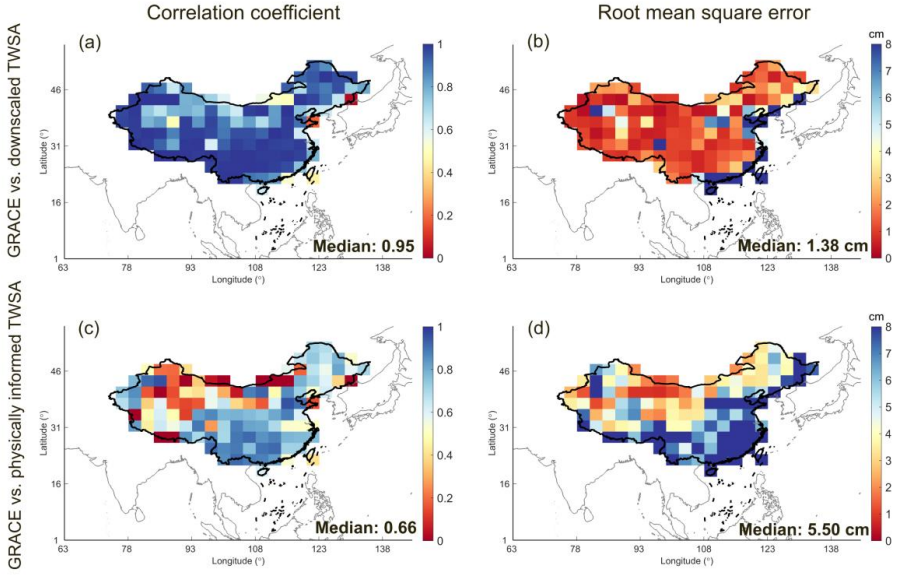

Fig. 4 Spatial distribution of evaluation indices of **(a)–(b)** downscaled TWSA and **(c)–(d)** physically informed TWSA relative to GRACE TWSA at each 3° grid cell, including correlation coefficient (left column) and root mean square error (right column). Median values of the evaluation index over all grid cells are labelled in bottom-right of each panel.

In addition to the grid-cell scale, TWS variability was assessed over the nine large basins in China to characterize regional hydrologic changes (Fig. 5). Downscaled and GRACE TWSA show good agreement for these basins, with a median *CC* of 0.94 (Fig. 5a) and *RMSE* of 1.72 cm (Fig. 5b). Performance over the Southwest, Yangtze, Yellow, Haihe, and Huaihe river basins is particularly reliable (*CC*: 0.92–1.00; *RMSE*: 0.46–2.57 cm), whereas relatively large discrepancies are found for the Southeast and Pearl river basins in the humid Southeast China (*RMSE*: 9.60 cm of the Southeast basin and 5.43 cm of the Pearl basin), which may be attributed to uncertainties from land-ocean boundary effects and large variability in precipitation. The northeastern (Songhua and Liaohe River Basin) and northwestern (Continental Basin) regions exhibit high consistency between $TWSA_{DS}$ and $TWSA_{GRACE}$, with *CC* approaching 0.90 and *RMSE* within 2 cm, indicating robust performance of the deep-learning downscaling framework across inland and continental regions in China.

Compared to $TWSA_{DS}$, the physically informed results ($TWSA_{PI}$) demonstrate lower performance in all basins, with a median *CC* of 0.80 (Fig. 5c) and median *RMSE* of 5.28 cm (Fig. 5d). These results highlight the role of GRACE constraints in enhancing both temporal variability and signal amplitude in the deep learning-based downscaling approach. The largest improvement in temporal variability is observed in the Continental Basin of Northwest China, where *CC* increases from 0.47 ($TWSA_{PI}$ vs. $TWSA_{GRACE}$) to 0.88 ($TWSA_{DS}$ vs. $TWSA_{GRACE}$). In addition, amplitude discrepancies are markedly reduced in the Yangtze and Southwest basins, where *RMSE* decreases by approximately 90%. In the Yangtze basin, *RMSE* decreases from 5.28 cm ($TWSA_{PI}$ vs. $TWSA_{GRACE}$) to 0.89 cm ($TWSA_{DS}$ vs. $TWSA_{GRACE}$), while in the Southwest basin, it declines from 9.25 cm ($TWSA_{PI}$ vs. $TWSA_{GRACE}$) to 0.98 cm ($TWSA_{DS}$ vs. $TWSA_{GRACE}$).

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

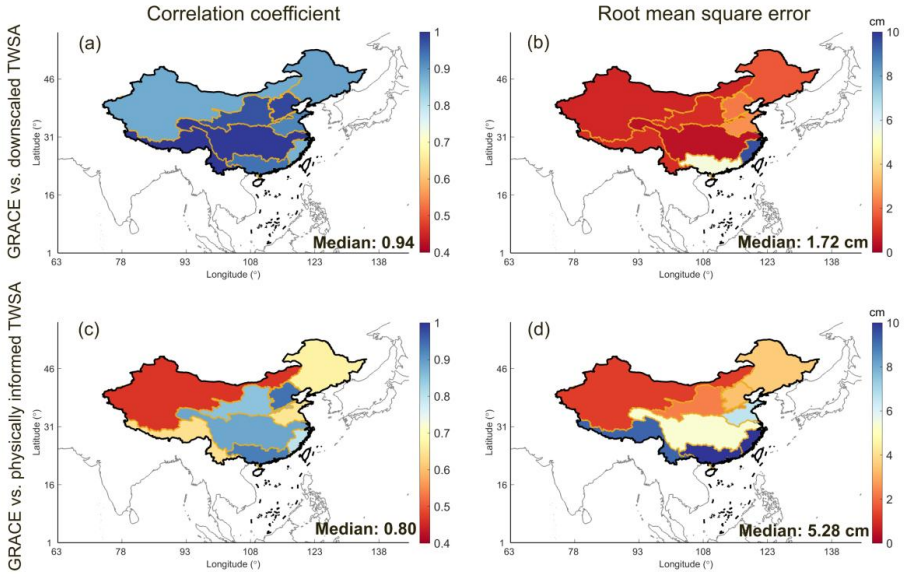

Fig. 5 Basin-scale evaluation of **(a)–(b)** downscaled TWSA and **(c)–(d)** physically informed TWSA relative to GRACE TWSA over nine river basins across China during 2002–2019, including correlation coefficient (left column) and root mean square error (right column). Median values of the evaluation index over all basins are labelled in bottom-right of each panel.

Basin-scale time series provide more details to understand the performance of downscaled and physically informed TWSA in characterizing seasonal amplitude, interannual variability, and long-term trends in TWSA (Fig. 6). Across nine large river basins in China, the Haihe, Huaihe, Yellow, and Southwest basins show significant TWS declines during 2002–2019. Compared with the physically informed simulations, the downscaled TWSA better reproduces both magnitude and temporal evolution of these declining trends. For example, in the Southwest basin (Fig. 6h), $TWSA_{PI}$ mainly captures seasonal variations but fails to represent the long-term decreasing trend observed by GRACE. In contrast, $TWSA_{DS}$ successfully reproduces this declining trend, demonstrating the benefit of incorporating GRACE constraints into the deep learning framework. As for the Haihe (Fig. 6b), Huaihe (Fig. 6c), and Yellow (Fig. 6d) river basins, $TWSA_{PI}$ shows noticeable deviations from $TWSA_{GRACE}$ after 2015, whereas the $TWSA_{DS}$ better reproduces the observed signals. Similar deviations between $TWSA_{GRACE}$ and $TWSA_{PI}$ after 2015 are also evident in the Yangtze basin with an increasing TWS trend (Fig. 6e), whereas $TWSA_{DS}$ closely follows the GRACE observations, successfully reproducing both the increasing trend and seasonal variations.

We observe an overestimation of variability amplitude in the physical simulations in the Songhua and Liaohe River basin (Fig. 6a), the Pearl River basin (Fig. 6f), and the Southeast basin (Fig. 6g), where TWS signals are dominated by seasonal and interannual variability. In these basins, the downscaled TWSA reflects the combined influence of GRACE observations and physical simulations, with its temporal variations generally lying between those of the two. Despite improvements, the downscaled product still faces challenges in fully reproducing GRACE variability in these basins, particularly the coastal regions with high precipitation variability, where both GRACE retrievals and physical simulations show large uncertainty. This suggests that improving the quality of training labels and constraints could be important for deep learning-based downscaling. Finally, the physically informed TWSA shows the weakest temporal coherence with GRACE in the Continental Basin of northwestern China (Fig. 6i), where TWS variability is primarily dominated by interannual changes. GRACE also indicates TWS decline during 2015–2017 followed by a recovery in 2018–2019 in this region. We find that the downscaled TWSA improves the temporal coherence with GRACE and captures part of these changes. However, the magnitude of TWS declines during this period is not fully represented in

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

TWSA$_{DS}$, which may be related to the challenge in linking hydrologic predictors to TWS changes in arid regions of Northwest China.

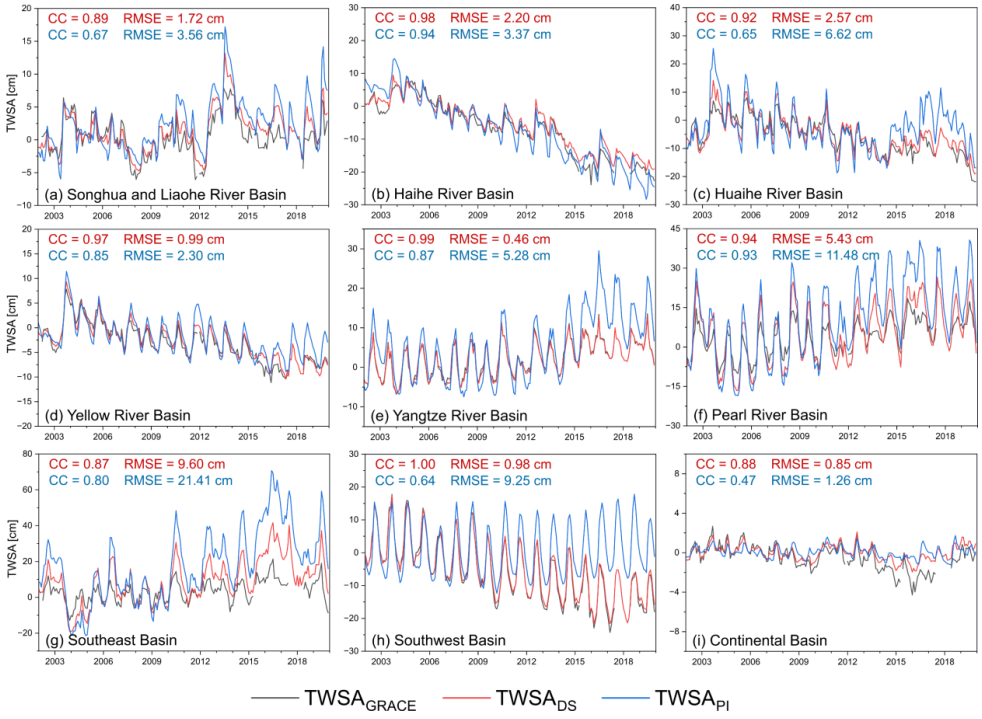

Fig. 6 Monthly time series of TWSA during 2002–2019 across nine river basins in China, estimated from GRACE (TWSA$_{GRACE}$; black line), downscaled outputs (TWSA$_{DS}$; red line), and physically informed estimates (TWSA$_{PI}$; blue line). The correlation coefficient (*CC*) and root mean square error (*RMSE*) between GRACE and two other estimations are labelled in top-left of each panel, with red text indicating TWSA$_{GRACE}$ vs. TWSA$_{DS}$ and blue text indicating TWSA$_{GRACE}$ vs. TWSA$_{PI}$.

4.2 Reliability and uncertainty estimation of downscaled TWSA at fine scales

Because accurate in-situ TWS measurements at sub-regional scales in China are not available, we cannot directly evaluate the absolute magnitude of downscaled TWSA at a high-resolution. Therefore, we focus on the reliability of downscaled TWSA indicated by the spatial correlation with physically informed TWSA, uncertainty estimation at the pixel level, and ability to close water balance across small to medium catchments beyond the native GRACE resolution. We computed the Pearson correlation between TWSA$_{DS}$ and TWSA$_{PI}$ at each 0.1° pixel over the entire time span of 2002–2019. The Pearson correlation is invariant to magnitude scaling, which reduces the impact of inaccurate magnitudes of physical informed estimates. Results show that the overall correlation between downscaled and physically informed TWSA is high with a median value of 0.91 (Fig. 7a). The relatively low correlations are mainly found in arid regions in north and northwest China, likely due to weak hydrologic signals, to which both GRACE observations and physical models have limited sensitivity.

The monthly averaged uncertainty during 2002–2019 indicates stable model performance, with a median uncertainty of 0.88 cm (Fig. 7b). Approximately 82% of pixels have uncertainty lower than the average uncertainty of GRACE JPL-M over China (1.96 cm). The pixel-scale uncertainty generally decreases from southeast to northwest across China, which is likely associated with the larger amplitude of TWS variability in the humid southeastern regions. Relatively high uncertainties (> 4 cm) are mainly found in the middle and lower Yangtze and southern China, high mountain regions of northwestern and southwestern China, areas with large lakes, and the highly irrigated regions. These high-uncertainty pixels appear as scattered and localized patches rather than forming extensive contiguous regions. In

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

addition to spatial patterns, changes in monthly uncertainty shows larger values during the gap period between GRACE and GRACE-FO missions (Jun 2017–May 2018; inserted subplot in Fig. 7b), highlighting the importance of satellite observations in constraining the deep learning framework.

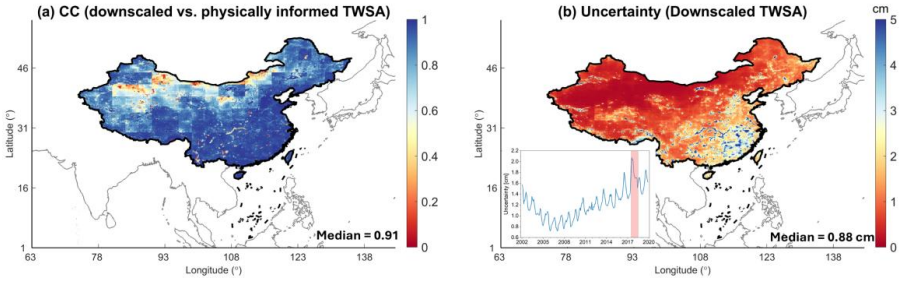

Fig. 7 (a) Correlation coefficient between downscaled and physically informed TWSA and **(b)** monthly averaged uncertainty of downscaled TWSA for each 0.1° pixel during 2002–2019. The inserted subplot in (b) shows the spatially averaged uncertainty across China for each monthly map, with the red rectangle indicating the gap period (Jul 2017–May 2018) between the GRACE and GRACE-FO missions. Missing data during the gap period have been reconstructed by the U-Net deep learning.

To evaluate the ability of downscaled TWSA to close the water balance across small to medium catchments beyond GRACE resolution, we compared TWS changes from storage and flux terms across 163 small to median catchments (10,000–100,000 km^2) in China (Section 3.3). The downscaled TWSA shows good agreement with the water budget-derived TWS changes across most catchments. The median *CC* between TWS changes derived from the water budget and the downscaled TWSA reaches 0.80 (Fig. 8a), with a low median *RMSE* of 1.44 cm (Fig. 8b). Relatively low correlations are observed in the middle reaches of the Yangtze River (Fig. 8a). This may be related to the operation of the Three Gorges Reservoir, where large-scale human water regulation causes storage changes that are not fully represented by the natural water balance. In addition, relatively large *RMSE* values are found in humid Southeast China (Fig. 8b), where monsoon precipitation leads to high variability in hydrologic fluxes and could introduce additional uncertainty in the water balance estimation. The performance of downscaled TWSA in closing the water balance is comparable to that of the physically informed TWSA, which exhibits a median *CC* of 0.79 (Fig. 8c) and a median *RMSE* of 1.36 cm (Fig. 8d). These results suggest that the deep learning-based downscaling framework generally preserves the key hydrologic relationships involved in the water balance.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

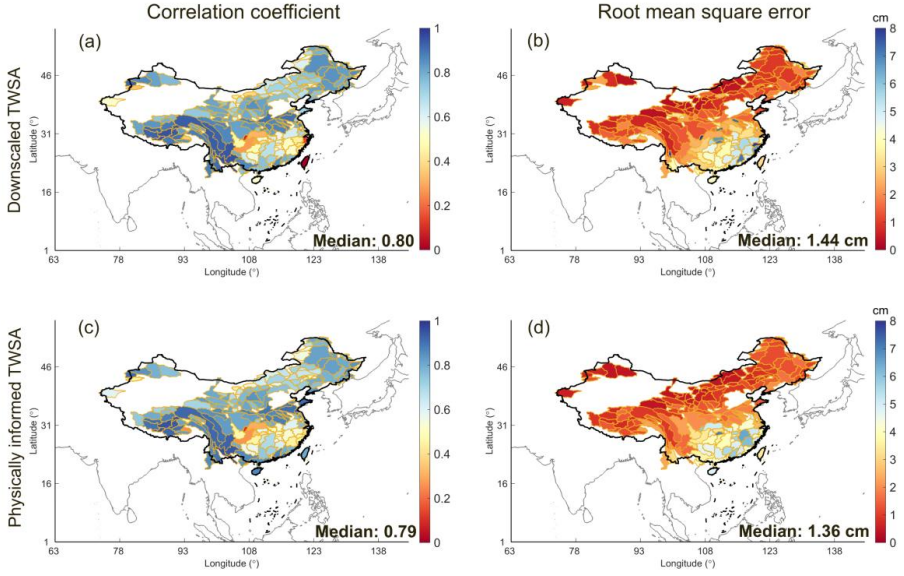

Fig. 8 Catchment-scale evaluation of monthly TWS change derived from **(a)–(b)** downscaled TWSA and **(c)–(d)** physically informed TWSA, compared with TWS change from water budget across 163 small to medium catchments in China during 2002–2019. The left column shows correlation coefficients, and the right column shows root mean square errors. Median values of the evaluation index over all basins are shown in bottom-right of each panel.

4.3 Evaluation across highly irrigated and glacierized regions

To further evaluate the physical consistency of the downscaled TWSA, we focus on regions where TWS variability is dominated by specific components, such as groundwater storage in highly irrigated areas and glacier mass balance changes in high mountain regions. In these regions, independent observations of dominant TWS components provide a valuable benchmark for assessing the performance of the downscaled TWSA. Because the comparison reference is not the observed TWS, absolute error metrics such as *RMSE* are less informative, whereas correlation is more appropriate for assessing the consistency between changes in downscaled TWSA and those in independent observations. The NCP is intensively irrigated (Fig. 1) with TWS loss dominated by groundwater consumption. The correlation between the downscaled TWSA and observed groundwater levels indicates generally good performance across 559 groundwater wells spanning 2005–2018. Approximately 80% of the wells show positive correlations, with a median *CC* of 0.41 (Fig. 9a). Notably, the agreement improves in regions with higher irrigation intensity, such as the eastern and southwestern margins of the NCP. The median *CC* increases to 0.48 and 0.65 for wells located in areas with irrigation intensity exceeding 40% and 50%, respectively. This reflects the increasing dominance of groundwater in TWS variability in highly irrigated regions, highlighting the capability of the downscaling framework to capture TWS variations under strong human intervention.

We do not find a substantial difference in overall TWS-groundwater correlation after and before downscaling: 51% of wells show an increase in *CC*, while 49% exhibit a decrease (Fig. 9b). A consistent case was reported across irrigation zones in India by Wang et al. (2024), who demonstrated that the overall correlation between variations in TWSA and those in groundwater levels did not change much after TWS downscaling. This may be related to the fact that groundwater level variations observed at individual wells may not be representative for the entire aquifer that is subject to intensive groundwater extraction. As a result, the spatial scale at which point-based groundwater observations are aggregated or compared with gridded TWSA may introduce scale-related uncertainties in the evaluation. However, the effect of downscaling is not uniform. We find a significant negative relationship between the change

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

in correlation ($CC_{\text{downscaled, obs}} - CC_{\text{GRACE, obs}}$) and the baseline correlation ($CC_{\text{GRACE, obs}}$) (Fig. 9c). This suggests that in regions with low initial agreement between GRACE TWSA and in situ groundwater levels, likely reflecting strong spatial heterogeneity in groundwater level changes, the downscaled TWSA improves the correlation with groundwater level measurements, thereby better capturing localised TWS variability. This pattern highlights the ability of the deep learning model to recover localised TWS signals in complex, human-impacted regions.

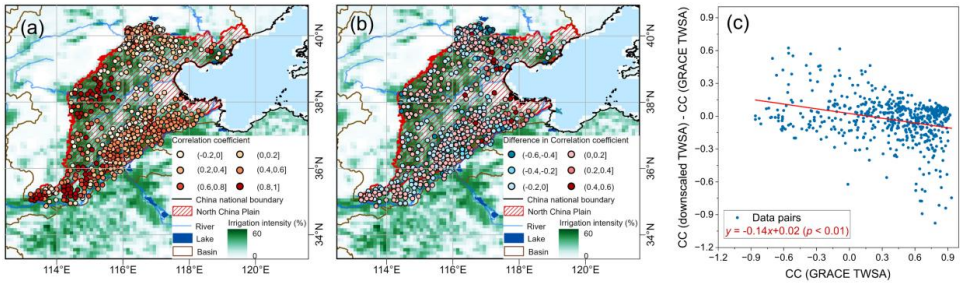

Fig. 9 In-situ validation of downscaled TWSA across the North China Plain. **(a)** Correlation coefficient (*CC*) between groundwater level observations and downscaled TWSA ($CC_{\text{downscaled, obs}}$) at 559 wells during 2005–2018. **(b)** Difference in correlation between TWSA and groundwater observations after and before downscaling ($CC_{\text{downscaled, obs}} - CC_{\text{GRACE, obs}}$). **(c)** Scatterplots of the difference in *CC* ($CC_{\text{downscaled, obs}} - CC_{\text{GRACE, obs}}$) versus the *CC* before downscaling ($CC_{\text{GRACE, obs}}$). Geographical information and irrigation intensity (%) of the NCP are shown in panels (a) and (b).

In addition to groundwater-dominated TWS variations, we evaluated the reliability of the downscaled TWSA over a representative glacierized region, the SETP in Southwest China (Fig. 1; Fig. 10a). Based on satellite altimetry observations from ICESat (2003–2009) and CryoSat-2 (2010–2019), Zhao et al. (2022) estimated regionally averaged changes in cumulative surface elevation over the SETP. We compared region-averaged TWSA from the downscaled product and cumulative glacier surface elevation changes at both seasonal and annual time scales (Fig. 10b). All estimates were converted to anomalies relative to the 2003–2019 mean to ensure consistency. We find strong agreement between the downscaled TWSA and satellite altimetry observations at the seasonal scale across both ICESat and CryoSat-2 periods (*CC*: 0.90). When aggregated to the annual scale, which reduces short-term variability and observational noise, the correlation further increases to 0.97 (Fig. 10b). This represents a substantial improvement compared to conventional high-resolution TWS estimation relied on physical models, which often poorly simulate glacierized regions due to the lack of glacier modules. Specifically, the averaged TWSA over the SETP simulated by PCR-GLOBWB 2 shows a low *CC* value of 0.12 and 0.05 at the seasonal and annual scale, respectively (Fig. 10c). These results further demonstrate that the downscaled product improves the representation of TWS variability in glacierized areas, thereby extending its applicability to cryospheric regions.

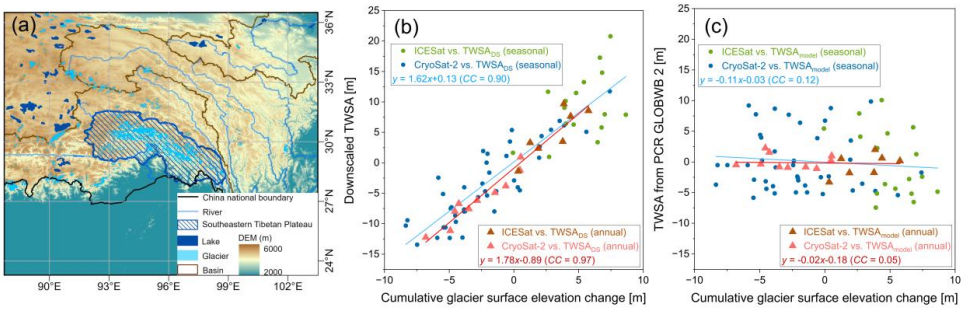

Fig. 10 Validation of downscaled TWSA over the Southeastern Tibetan Plateau (SETP) using independent satellite altimetry retrievals. **(a)** Geographical information over the SETP. **(b)–(c)** Scatterplots of cumulative glacier surface elevation change and region-mean TWSA, estimated from **(b)** the downscaled product and **(c)** PCR-

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Open Access Earth System Science Data Discussions

GLOBWB 2. In (b) and (c), seasonal mean and annual mean scatterplots are represented by dots and triangles, respectively. Data pairs during the ICESat period are shown in green (seasonal) and brown (annual), whereas those during the CrySat-2 period are shown in blue (seasonal) and red (annual). The expression of linear regression and correlation coefficient are shown in light blue (seasonal) and dark red (annual).

4.4 High-resolution spatial details

We present spatial patterns of downscaled TWSA for a representative dry month (March 2010), a wet month (July 2010), and the long-term trend during 2002–2019 to illustrate high-resolution spatial details (Fig. 11a1–a3). The downscaled outputs provide enhanced representation of fine-scale spatial variability and TWS trends, offering more intricate details and effectively pinpointing locations of extreme TWSA values. In March 2010 (Fig. 11a1), the downscaled TWSA shows pronounced negative anomalies over Southwest China, consistent with the peak of the severe 2009–2010 drought in this region (Long et al., 2014; Lu et al., 2011). In contrast, in the wet month of July 2010 (Fig. 11a2), large positive TWSA is reproduced across the middle and lower reaches of the Yangtze River Basin, associated with intense Meiyu rainfall and widespread flooding (Gao et al., 2016). The spatial pattern of TWS trends during 2002–2019 (Fig. 11a3) reveals hotspots at sub-regional scales of hundreds to thousands of square kilometres, consistent with previous studies based on in situ observations and/or model simulations. These hotspots include significant TWS declines in the Tienshan Mountains in Northwest China, the SETP and the Himalayas in Southwest China, the Hetao Irrigation District in the Yellow River Basin, the NCP in North China, and the Songhua-Nenjiang Irrigation Plain in Northeast China. In contrast, substantial TWS increases are observed in the Karakoram and western Kunlun Mountains in western China, the endorheic basins of the Tibetan Plateau, and humid southeastern China. The identified hotspots from downscaled TWSA are consistent with previously reported hydrologic processes, including glacier mass changes in high mountain regions (Farinotti et al., 2020; Farinotti et al., 2015; Hugonnet et al., 2021; Zhao et al., 2022), irrigation-induced groundwater depletion in intensively cultivated areas (Feng et al., 2013; Feng et al., 2025), lake expansion in the endorheic Tibetan Plateau (Li et al., 2019c), and flood-driven water storage increases in humid Southeast China (Slater et al., 2021). The agreement with independent studies strongly supports the physical plausibility of the downscaled TWSA patterns.

In addition, downscaled TWSA was aggregated to a coarse resolution of 3° for the selected dry and wet months as well as the long-term trend (Fig. 11b1–b3) to compare with GRACE observations. Except for a few coastal grid cells, the block-mean TWSA from the downscaled product closely reproduces GRACE estimates (Fig. 11c1–c3), capturing both spatial patterns and extreme values and demonstrating spatial reliability. To assess the extent to which fine-scale spatial variability is represented, the standard deviation of all 0.1° pixels within each 3° grid cell (900 pixels in each grid cell) was calculated as an indicator of sub-grid variability (Fig. 11d1–d3). While we acknowledge that standard deviation is not a direct measure of spatial details, a higher pixel-wise standard deviation, given a reasonable large-scale spatial pattern, generally indicates richer fine-scale spatial variability. Overall, higher sub-grid variability in both dry (Fig. 11d1) and wet (Fig. 11d2) months is observed in humid and topographically complex regions, while relatively low variability is found in arid and semi-arid regions of Northwest China. The median standard deviation of the dry month (March 2010; Fig. 11d1) is 9.31 cm, whereas that of the wet month (July 2010; Fig. 11d2) reaches 12.62 cm. The spatial pattern of variability in the long-term trend (Fig. 11d3) highlights regions with heterogeneous TWS changes, particularly in the NCP, Hetao Irrigation District in the Yellow River Basin, Tienshan Mountains and SETP. These patterns coincide with complex physical processes under sub-grid heterogeneity, indicating the potential of downscaling TWSA in identifying localized extremes and sub-regional hotspots of water storage change.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

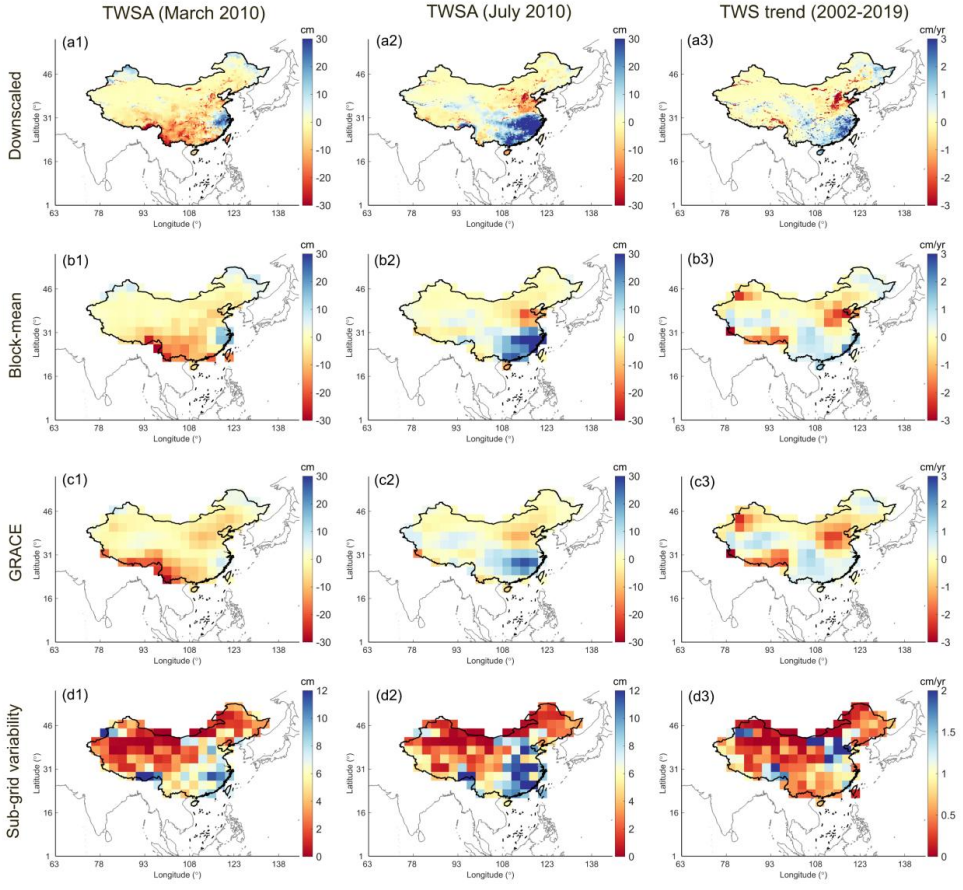

Fig. 11 Spatial pattern of TWSA in a representative dry month (March 2010; the first column), a wet month (July 2010; the second column), long-term trend during 2002–2019 (the third column), and sub-grid spatial variability in each 3° grid cell (the fourth column).

5 Discussion

5.1 Comparison with representative downscaling studies

Two representative recent studies on TWS downscaling were selected to compare with our results. Gou and Soja (2024) developed a CNN-based framework to downscale JPL-M from 3° to 0.5° at the global scale, incorporating physical simulations from the WaterGAP Global Hydrologic Model (WGHM). This study was selected because its model framework is conceptually comparable to our study, as both frameworks utilize convolutional architectures. The other representative study was conducted by Xiong et al. (2025), which employed a joint inversion-based downscaling approach to generate 0.5° TWSA over China. The downscaling results based on JPL-M were compared to be consistent with the GRACE data source of our study. This work represents a recent effort specifically designed for regional TWSA reconstruction over China. Together, these two studies provide representative benchmarks for evaluating the performance of our downscaled results.

The proposed downscaling framework in our study generally shows better performance compared to these two studies, not only improving large-scale consistency with GRACE observations, but also enhancing physical consistency and spatial detail across diverse hydrologic regimes (Table 2). At the 3° grid scale, the downscaled TWSA of our study achieves a higher correlation coefficient (*CC*: 0.95) and

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

a substantially lower *RMSE* (1.38 cm) compared with Gou and Soja (2024) (*CC*: 0.80, *RMSE*: 3.85 cm) and Xiong et al. (2025) (*CC*: 0.57, *RMSE*: 5.00 cm). Across large basins, our results maintain comparable correlation with Gou and Soja (2024) (*CC*: 0.94 for both) while further reducing *RMSE* (1.72 cm in this study vs. 2.63 cm in Gou and Soja (2024). These results outperform those from Xiong et al. (2025), which has a *CC* of 0.93 and *RMSE* of 3.07 cm for large river basins in China. In addition, the uncertainty level of the downscaled TWSA (0.88 cm) remains comparable to that of Gou and Soja (2024) (0.81 cm), demonstrating stable model performance at the pixel scale. The advantage of the proposed framework is more evident in process-based evaluations at finer scales. In particular, for water balance closure across small to medium catchments, the downscaled TWSA in this study shows higher correlation (*CC*: 0.80) and lower *RMSE* (1.44 cm) than both reference studies (*CC*: 0.70 and *RMSE*: 1.69 cm in Gou and Soja (2024); *CC*: 0.35 and *RMSE*: 2.56 cm in Xiong et al. (2025)). This improvement could partly be attributed to the higher spatial resolution of generated TWSA of this study (0.1° compared to 0.5° in the reference studies), which allows better representation of sub-basin variability.

Across the NCP, where TWS variations are dominated by groundwater variations, the results of this study are comparable to those of Gou and Soja (2024). The comparable performance is reflected in the similar correlation with in-situ groundwater levels. Gou and Soja (2024) reports a *CC* of 0.42–0.63 under increasing irrigation intensity, while our study achieves *CC* values of 0.41–0.65. Results of Xiong et al. (2025) show weaker agreement with a *CC* of 0.36–0.49. This could be related to the fact that both Gou and Soja (2024) and this study integrate outputs from multiple hydrologic simulations (WGHM and PCR-GLOBWB 2, respectively), which provide essential constraints on human-driven water storage changes, particularly relevant in the NCP. However, the absence of glacier processes in Gou and Soja (2024) results in poor performance over glacierized regions, with low correlations with glacier surface elevation changes (*CC*: 0.53 at seasonal scale and 0.63 at annual scale). Xiong et al. (2025) incorporated glacier processes in their inversion framework, and thus got a better performance (*CC*: 0.61 at seasonal scale and 0.81 at annual scale). In comparison, downscaled TWSA in our study achieves substantially higher correlations with glacier surface elevation changes (*CC*: 0.90 at seasonal scale and 0.97 at annual scale), showing clear improvements in representing glacier-dominant TWS changes. Furthermore, our results outperform these two studies in recovering fine-scale spatial variability. The standard deviation within each 3° grid cell of our downscaled TWSA is comparable to that of Xiong et al. (2025) and much higher than that of Gou and Soja (2024), indicating a stronger ability to resolve sub-grid heterogeneity than Gou and Soja (2024).

Table 2 Performance comparison of downscaled TWSA in this study with two representative studies across multiple evaluation metrics. All metrics represent the median values calculated from the evaluated samples.

Evaluation index	Gou and Soja (2024)	Xiong et al. (2025)	This study
Comparison with GRACE TWSA of 3° grid cells across China			
CC	0.80	0.57	0.95
RMSE (cm)	3.85	5.00	1.38
Comparison with GRACE TWSA of nine large river basins			
CC	0.94	0.93	0.94
RMSE (cm)	2.63	3.07	1.72
Pixel-wise uncertainty estimation			
Uncertainty (cm)	0.81	–	0.88
Comparison with TWS change derived from water budget in small to medium catchments			
CC	0.70	0.35	0.80
RMSE (cm)	1.69	2.56	1.44
Correlation with groundwater levels across the NCP			
Positive CC (%)	80	79	80
CC	0.42	0.36	0.41
CC over areas with AEI > 40%	0.46	0.39	0.48
CC over areas with AEI> 50%	0.63	0.49	0.65
Correlation with regionally averaged glacier surface elevation change across the SETP			

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

CC (seasonal)	0.53	0.61	0.90
CC (annual)	0.63	0.81	0.97
Recovery of spatial details within 3° grid cell in representative months			
SD (cm; dry month March 2010)	2.51	11.51	9.31
SD (cm; wet month July 2010)	2.79	12.31	12.62

Note: Xiong et al. (2025) does not provide uncertainty estimation associated with their downscaling product. Regarding the underlying GRACE data, this study and Gou and Soja (2024) are based on the JPL mascon solution, whereas Xiong et al. (2025) is based on the JPL spherical harmonic solution. AEI: area equipped irrigation; SD: standard deviation.

5.2 Applications, limitations and future work

Important applications of downscaled TWSA can be summarized in three main aspects. First, the high-resolution TWSA enables detailed assessment of spatial heterogeneity in TWS changes, facilitating the identification of sub-regional hotspots associated with climate variability and human activities. For example, high mountain regions are clearly resolved in the downscaled TWSA patterns, and human-induced TWS changes, such as large TWS depletion in the Hetao Irrigation District within the Yellow River Basin, can be explicitly captured compared to the smoothed patterns in the original GRACE observations (Fig. 11b3). This improved spatial representation allows for more accurate characterization of water risk and resource dynamics at sub-regional scales, tracking impacts from both climate variability and human activities on China's water availability.

Second, the enhanced spatial resolution improves the identification of hydrologic extremes by capturing TWSA variability within individual grid cells. For example, during the severe drought in Southwest China in 2009–2010, the downscaled TWSA shows a region-averaged negative anomaly of 16 cm in Yunnan Province (97°–106°E, 21°–29°N) at the drought peak in March 2010. This is comparable to the GRACE estimate (-16.6 cm) at the region-averaged scale. However, the downscaled TWSA reveals a clear northeast-southwest gradient in drought severity across Southwest China (Fig. 11a1). The most pronounced negative anomaly reaches -57 cm in southwestern Yunnan, nearly twice the corresponding GRACE estimate (-29 cm). This highlights the potential of the downscaled TWSA to resolve sub-regional heterogeneity and capture localized extremes beyond the native resolution of GRACE.

Third, the dataset is provided at 0.1° resolution, consistent with the spatial resolution of commonly used climate forcing (e.g., ERA5-Land) and hydrologic flux estimation (e.g., 0.1° precipitation derived from Integrated Multisatellite Retrievals for Global Precipitation Mission). This consistency allows direct multi-variable analysis without additional resampling. High-resolution TWSA facilitates integrated analyses of interactions between water storage and hydrologic fluxes at fine scales, whose relationships remain unclear (Tiwari et al., 2025). Further, the downscaled TWSA can serve as a storage reference for evaluating performance of hydrologic model, reducing the risk of equifinality associated with flux-only calibration (Arsenault and Brissette, 2014; Beven, 2006).

One caveat of the deep learning framework is that it primarily learns spatial relationships, with limited explicit constraint on temporal dynamics. This is inherent to the convolutional architecture, while the spatial construction of training and validation samples at each monthly time step further reinforces this tendency. In this study, the temporal continuity of input physical variables mitigates this limitation, and the downscaled TWSA still shows good performance in temporal evaluations, such as the reliability of TWSA time series within fixed spatial scale. However, temporal dependencies are not explicitly constrained in the current framework. Future work could couple the U-Net framework with temporal modelling architectures, such as ConvLSTM (Shi et al., 2015), to jointly learn spatial patterns and temporal dependencies. This extension would improve the robustness of downscaled outputs to capture the spatiotemporal dynamics of TWSA and enhance its capability for predictions and projections.

It is worth noting that the estimated uncertainty in the downscaled product refers to that introduced during the deep learning downscaling. However, it does not include uncertainties in the input labels and predictors, which may contribute to lower performance in coastal regions in Southeast China and arid regions in Northwest China. In coastal regions, GRACE retrievals are affected by signal leakage and reduced accuracy near land-ocean boundaries. Missing values over ocean areas in PCR-GLOBWB 2 limit the number of effective training patches, potentially diminishing the robustness of learning

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

processes. In addition, Southeast China is characterized by high variability in precipitation, driven by the East Asian monsoon and ocean-atmosphere coupling in the western Pacific (Ding and Chan, 2005). This may introduce additional uncertainty in climate-related predictors. In arid regions, TWS variability is often weakly coupled with hydrologic predictors. The lower signal-to-noise ratios of GRACE observations and model-based inputs further increase uncertainty in both training and validation, making it more difficult to learn reliable relationships (Gou and Soja, 2024). Future work could integrate uncertainty propagation from input data within the modelling framework, enabling a more comprehensive and systematic characterization of dataset uncertainty.

6 Conclusions

This study downscaled 3° GRACE TWSA to a much higher resolution of 0.1° using a deep learning framework. The downscaled TWSA over China during 2002–2019 demonstrates its high reliability in preserving large-scale GRACE signals (3°) while enhancing fine-scale spatial detail (0.1°). Compared with GRACE TWSA, the downscaled product achieves high *CC* values of 0.95 at the 3° grid scale and 0.94 at the basin scale, with corresponding low *RMSE* values of 1.38 cm and 1.72 cm. Independent training runs estimate a median pixel-wise uncertainty of 0.88 cm, lower than the median uncertainty of GRACE JPL-M across China (1.96 cm). Process-based evaluation further confirms the physical consistency of the downscaled TWSA. Water balance closure across 163 small to medium catchments (10,000–100,000 km^2) shows good agreement between TWS changes derived from the downscaled TWSA and ERA5L-based hydrologic fluxes (median *CC* of 0.80 and *RMSE* of 1.44 cm). In addition, downscaled TWSA shows median *CC* values of 0.48 and 0.65 with groundwater level observations over the NCP, corresponding to regions with irrigation intensity exceeding 40% and 50%, respectively. Over the glacierized SETP, the downscaled TWSA captures variations in cumulative glacier surface elevation derived from ICESat and CryoSat-2 altimetry, with high *CC* values of 0.90 and 0.97 at the seasonal and annual scales, respectively.

Results from our study generally show better performance than those from two recent representative studies on TWS downscaling in China. In addition, the spatial pattern of downscaled TWS trends reveals hotspots at sub-regional scales of hundreds to thousands of square kilometres, such as mountain ranges and irrigation zones, consistent with previous studies based on in situ observations and/or model simulations. The downscaled TWSA provides enhanced spatial detail for identifying fine-scale hotspots and sub-regional heterogeneity in TWS changes associated with hydro-climatic extremes, which are not resolved by the native resolution of GRACE. The resolution of downscaled TWSA (0.1°) aligns well with the commonly used climate forcing and hydrologic flux estimation, enabling direct analyses in storage-flux interactions and evaluation of hydrologic model performance from a storage perspective.

Estimating high-resolution TWSA is still challenging. Despite some limitations (e.g., lack of explicit constraints on temporal generalization), this study provides valuable insights for integrating multi-source information to bridge the scale gap between coarse-resolution gravity observations and fine-scale hydrologic processes. The generated datasets not only support investigation of sub-regional TWS variability and associated hydrologic processes in China, but also offer insights for advancing statistical downscaling of TWSA at broader scales.

7 Data and code availability

The new 0.1° TWSA dataset (Li and Sun, 2026a) is archived on *Zenodo* at https://doi.org/10.5281/zenodo.19502906. The uploaded data include (1) the standard version covering 2002–2019 with the most comprehensive observations for quality control, and (2) the extended version spanning 2020–2023 to support recent hydrologic analyses. This repository also provides the complete training datasets used for model development and reproducibility, consisting of training labels, dynamic predictors, and static predictors.

The U-Net model was implemented using standard Python packages. For reproducibility, the core codes used to build the model (Li and Sun, 2026b) are available on *Zenodo* at https://doi.org/10.5281/zenodo.19502631.

All datasets used in this study are readily accessible. GRACE and GRACE-FO data are provided at https://grace.jpl.nasa.gov/. Water storage simulations from PCR-GLOBWB 2 are accessed at

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

https://globalhydrology.nl/research/models/pcr-globwb-2-0/. Glacier mass balance produced by Hugonnet et al. (2021) can be accessed at https://doi. org/10.6096/13. Input predictors include: ERA5L data available at https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land-monthly-means?tab=download; NDVI derived at https://modis.gsfc.nasa.gov/data/dataprod/mod13.php; DEM data available at https://search.earthdata.nasa.gov/; irrigation maps accessed at https://www.fao.org/aquastat/en/geospatial-information/global-maps-irrigated-areas/latest-version; and population datasets derived at https://www.earthdata.nasa.gov/data/catalog/sedac-ciesin-sedac-gpwv4-popcount-r11-4.11. Independent observations of groundwater levels are accessed at https://www.geodata.cn/oldindex.html. Geographic information for the study region includes: shapefiles of China's national boundary, major rivers, and the delineation of river basins accessed from Resource and Environmental Science Data Center at https://www.resdc.cn/; glacier distribution accessed from RGI 6.0 at https://www.glims.org/RGI/randolph60.html; and location of lakes accessed from HydroLakes at https://www.hydrosheds.org/products/hydrolakes.

Author contributions

X.L. conceptualized the study. X.L. and Y.S. developed the methodology and performed the analysis. L.J.S., X.G., N.W., and B.R.S. provided critical input and contributed to the interpretation of results. X.L. drafted the manuscript. All authors discussed the results and revised the manuscript.

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors thank F. Zhao from Tsinghua University for providing cumulative glacier surface elevation over the SETP retrieved from ICESat and CryoSat-2 satellites, as well as J. Gou from ETH Zurich and B. Li from NASA Goddard Space Flight Center for valuable discussions on the U-Net model training.

Financial support

This work is supported by UK Research and Innovation (UKRI) under the Horizon Europe guarantee scheme (EP/Z002729/1), originally funded by the European Commission through the Marie Curie Programme. X.G. acknowledges support from the National Natural Science Foundation of China (Grants 42371041 & U2340230).

References

Arsenault, R. and Brissette, F.P., 2014. Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, 50(7): 6135–6153.

Arshad, A., Mirchi, A., Taghvaeian, S. et al., 2024. Downscaled-GRACE data reveal anthropogenic and climate-induced water storage decline across the Indus Basin. *Water Resources Research*, 60(7): e2023WR035882.

Bai, P., Liu, X. and Liu, C., 2018. Improving hydrological simulations by incorporating GRACE data for model calibration. *Journal of Hydrology*, 557: 291–304.

Beven, K., 2006. A manifesto for the equifinality thesis. *Journal of hydrology*, 320(1-2): 18–36.

Cao, G., Zheng, C., Scanlon, B.R. et al., 2013. Use of flow modeling to assess sustainability of groundwater resources in the North China Plain. *Water Resources Research*, 49(1): 159–175.

Chen, J., Cazenave, A., Dahle, C. et al., 2022. Applications and challenges of GRACE and GRACE follow-on satellite gravimetry. *Surveys in Geophysics*, 43(1): 305–345.

Chen, X., Long, D., Hong, Y. et al., 2017. 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 U pper Brahmaputra River basin? *Water Resources Research*, 53(3): 2431–2466.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

881 Ding, Y. and Chan, J.C., 2005. The East Asian summer monsoon: an overview. *Meteorology and*
882 *Atmospheric Physics*, 89(1): 117–142.

883 Döll, P., Müller Schmied, H., Schuh, C. et al., 2014. Global-scale assessment of groundwater depletion
884 and related groundwater abstractions: Combining hydrological modeling with information from
885 well observations and GRACE satellites. *Water Resources Research*, 50(7): 5698–5720.

886 Eicker, A., Schumacher, M., Kusche, J. et al., 2014. Calibration/data assimilation approach for
887 integrating GRACE data into the WaterGAP Global Hydrology Model (WGHM) using an
888 ensemble Kalman filter: First results. *Surveys in Geophysics*, 35(6): 1285–1309.

889 Farinotti, D., Immerzeel, W.W., de Kok, R.J. et al., 2020. Manifestations and mechanisms of the
890 Karakoram glacier Anomaly. *Nature geoscience*, 13(1): 8–16.

891 Farinotti, D., Longuevergne, L., Moholdt, G. et al., 2015. Substantial glacier mass loss in the Tien Shan
892 over the past 50 years. *Nature Geoscience*, 8(9): 716–722.

893 Feng, W., Zhong, M., Lemoine, J.M. et al., 2013. Evaluation of groundwater depletion in North China
894 using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based
895 measurements. *Water Resources Research*, 49(4): 2110–2118.

896 Feng, Z., Miao, Q., Shi, H. et al., 2025. Water Saving and Environmental Issues in the Hetao Irrigation
897 District, the Yellow River Basin: Development Perspective Analysis. *Agronomy*, 15(7): 1654.

898 Gao, T., Xie, L. and Liu, B., 2016. Association of extreme precipitation over the Yangtze River Basin
899 with global air-sea heat fluxes and moisture transport. *International Journal of Climatology*,
900 36(8): 3020–3038.

901 Gerdener, H., Kusche, J., Schulze, K. et al., 2023. The global land water storage data set release 2
902 (GLWS2.0) derived via assimilating GRACE and GRACE-FO data into a global hydrological
903 model. *Journal of Geodesy*, 97(7): 73.

904 Gou, J. and Soja, B., 2024. Global high-resolution total water storage anomalies from self-supervised
905 data assimilation using deep learning algorithms. *Nature Water*, 2(2): 139–150.

906 Gu, X., Zhang, Q., Li, J. et al., 2019. Attribution of global soil moisture drying to human activities: A
907 quantitative viewpoint. *Geophysical Research Letters*, 46(5): 2573–2582.

908 Guan, Y., Gu, X., Dai, A. et al., 2025. Anthropogenic enhancement of subsurface soil moisture droughts.
909 *Nature Climate Change*: 1–8.

910 Guan, Y., Gu, X., Slater, L.J. et al., 2023. Increase in ocean-onto-land droughts and their drivers under
911 anthropogenic climate change. *Npj Climate and Atmospheric Science*, 6(1): 195.

912 Gupta, H.V., Kling, H., Yilmaz, K.K. et al., 2009. Decomposition of the mean squared error and NSE
913 performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*,
914 377(1-2): 80–91.

915 Hugonnet, R., McNabb, R., Berthier, E. et al., 2021. Accelerated global glacier mass loss in the early
916 twenty-first century. *Nature*, 592(7856): 726–731.

917 Huss, M., 2013. Density assumptions for converting geodetic glacier volume change to mass change.
918 *The Cryosphere*, 7(3): 877–887.

919 Immerzeel, W.W., Lutz, A.F., Andrade, M. et al., 2020. Importance and vulnerability of the world's
920 water towers. *Nature*, 577(7790): 364–369.

921 Janzing, J., Wanders, N., van Tiel, M. et al., 2025. Hyper-resolution large-scale hydrological modelling
922 benefits from improved process representation in mountain regions. *Hydrology and Earth*
923 *System Sciences*, 29(23): 7041–7071.

924 Kalu, I., Ndehedehe, C.E., Ferreira, V.G. et al., 2024. Statistical downscaling of GRACE terrestrial
925 water storage changes based on the Australian Water Outlook model. *Scientific Reports*, 14(1):
926 10113.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

927 Kling, H., Fuchs, M. and Paulin, M., 2012. Runoff conditions in the upper Danube basin under an
928 ensemble of climate change scenarios. *Journal of hydrology*, 424: 264–277.

929 Landerer, F.W., Dickey, J.O. and Güntner, A., 2010. Terrestrial water budget of the Eurasian pan-Arctic
930 from GRACE satellite measurements during 2003–2009. *Journal of Geophysical Research:*
931 *Atmospheres*, 115(D23).

932 Landerer, F.W. and Swenson, S., 2012. Accuracy of scaled GRACE terrestrial water storage estimates.
933 *Water resources research*, 48(4).

934 Li, B., Rodell, M., Kumar, S. et al., 2019a. Global GRACE data assimilation for groundwater and
935 drought monitoring: Advances and challenges. *Water Resources Research*, 55(9): 7564–7586.

936 Li, F. and Kusche, J., 2026. Reproducing GRACE total water storage change at finer spatial scales.
937 *Geophysical Research Letters*, 53(1): e2025GL119881.

938 Li, X., Long, D., Han, Z. et al., 2019b. Evapotranspiration estimation for Tibetan Plateau headwaters
939 using conjoint terrestrial and atmospheric water balances and multisource remote sensing.
940 *Water Resources Research*, 55(11): 8608–8630.

941 Li, X., Long, D., Huang, Q. et al., 2019c. High-temporal-resolution water level and storage change data
942 sets for lakes on the Tibetan Plateau during 2000–2017 using multiple altimetric missions and
943 Landsat-derived lake shoreline positions. *Earth System Science Data*, 11(4): 1603–1627.

944 Li, X., Long, D., Scanlon, B.R. et al., 2022. Climate change threatens terrestrial water storage over the
945 Tibetan Plateau. *Nature Climate Change*, 12(9): 801–807.

946 Li, X., Long, D., Scanlon, B.R. et al., 2025. Retrievals and simulations of terrestrial water storage
947 changes and runoff over the Tibetan Plateau: Challenges and opportunities. *Fundamental*
948 *Research*.

949 Li, X. and Sun, Y., 2026a. A 0.1° terrestrial water storage anomaly dataset over China. In: Zenodo.

950 Li, X. and Sun, Y., 2026b. Code for: A physically guided deep learning reconstruction of terrestrial
951 water storage anomalies at 0.1° across China. In: Zenodo.

952 Liu, J., Yang, H., Gosling, S.N. et al., 2017. Water scarcity assessments in the past, present, and future.
953 *Earth's future*, 5(6): 545–559.

954 Long, D., Scanlon, B.R., Longuevergne, L. et al., 2013. GRACE satellite monitoring of large depletion
955 in water storage in response to the 2011 drought in Texas. *Geophysical Research Letters*, 40(13):
956 3395–3401.

957 Long, D., Shen, Y., Sun, A. et al., 2014. Drought and flood monitoring for a large karst plateau in
958 Southwest China using extended GRACE data. *Remote Sensing of Environment*, 155: 145–160.

959 Long, D., Xu, Y., Cui, Y. et al., 2025. Unprecedented large-scale aquifer recovery through human
960 intervention. *Nature Communications*, 16(1): 7296.

961 Long, D., Yang, W., Scanlon, B.R. et al., 2020. South-to-North Water Diversion stabilizing Beijing's
962 groundwater levels. *Nature Communications*, 11(1): 3665.

963 Loomis, B., Luthcke, S. and Sabaka, T., 2019. Regularization and error characterization of GRACE
964 mascons. *Journal of geodesy*, 93(9): 1381–1398.

965 Lu, E., Luo, Y., Zhang, R. et al., 2011. Regional atmospheric anomalies responsible for the 2009–2010
966 severe drought in China. *Journal of Geophysical Research: Atmospheres*, 116(D21).

967 Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A. et al., 2021. ERA5-Land: A state-of-the-art global
968 reanalysis dataset for land applications. *Earth System Science Data*, 13(9): 4349–4383.

969 Pascal, C., Ferrant, S., Selles, A. et al., 2022. Evaluating downscaling methods of GRACE (Gravity
970 Recovery and Climate Experiment) data: a case study over a fractured crystalline aquifer in
971 southern India. *Hydrology and Earth System Sciences*, 26(15): 4169–4186.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

Pfeffer, W.T., Arendt, A.A., Bliss, A. et al., 2014. The Randolph Glacier Inventory: a globally complete inventory of glaciers. *Journal of glaciology*, 60(221): 537–552.

Piao, S., Ciais, P., Huang, Y. et al., 2010. The impacts of climate change on water resources and agriculture in China. *Nature*, 467(7311): 43–51.

Pokhrel, Y., Felfelani, F., Satoh, Y. et al., 2021. Global terrestrial water storage and drought severity under climate change. *Nature Climate Change*, 11(3): 226–233.

Reager, J.T., Thomas, B.F. and Famiglietti, J.S., 2014. River basin flood potential inferred using GRACE gravity observations at several months lead time. *Nature Geoscience*, 7(8): 588–592.

Rodell, M., Famiglietti, J.S., Wiese, D.N. et al., 2018. Emerging trends in global freshwater availability. *Nature*, 557(7707): 651–659.

Rowlands, D.D., Luthcke, S.B., Klosko, S. et al., 2005. Resolving mass flux at high spatial and temporal resolution using GRACE intersatellite measurements. *Geophysical Research Letters*, 32(4).

Save, H., Bettadpur, S. and Tapley, B.D., 2016. High-resolution CSR GRACE RL05 mascons. *Journal of Geophysical Research: Solid Earth*, 121(10): 7547–7569.

Scanlon, B.R., Fakhreddine, S., Rateb, A. et al., 2023. Global water resources and the role of groundwater in a resilient water future. *Nature Reviews Earth & Environment*, 4(2): 87–101.

Scanlon, B.R., Zhang, Z., Save, H. et al., 2018. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *Proceedings of the National Academy of Sciences*, 115(6): E1080–E1089.

Shi, X., Chen, Z., Wang, H. et al., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Slater, L., Villarini, G., Archfield, S. et al., 2021. Global changes in 20-year, 50-year, and 100-year river floods. *Geophysical Research Letters*, 48(6): e2020GL091824.

Sun, A.Y., Save, H., Rateb, A. et al., 2024. Deciphering the role of total water storage anomalies in mediating regional flooding. *Geophysical Research Letters*, 51(16): e2023GL108126.

Sutanudjaja, E.H., Van Beek, R., Wanders, N. et al., 2018. PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model. *Geoscientific Model Development*, 11(6): 2429–2453.

Tapley, B.D., Watkins, M.M., Flechtner, F. et al., 2019. Contributions of GRACE to understanding climate change. *Nature Climate Change*, 9(5): 358–369.

Tiwari, A.D., Pokhrel, Y., Boulange, J. et al., 2025. Similarities and divergent patterns in hydrologic fluxes and storages simulated by global water models. *Nature Water*: 1–11.

Vishwakarma, B.D., Zhang, J. and Sneeuw, N., 2021. Downscaling GRACE total water storage change using partial least squares regression. *Scientific data*, 8(1): 95.

Wahr, J., Swenson, S. and Velicogna, I., 2006. Accuracy of GRACE mass estimates. *Geophysical Research Letters*, 33(6).

Wang, Y., Li, C., Cui, Y. et al., 2024. Spatial downscaling of GRACE-derived groundwater storage changes across diverse climates and human interventions with Random Forests. *Journal of Hydrology*, 640: 131708.

Wiese, D.N., Bienstock, B., Blackwood, C. et al., 2022. The mass change designated observable study: overview and results. *Earth and Space Science*, 9(8): e2022EA002311.

Wiese, D.N., Landerer, F.W. and Watkins, M.M., 2016. Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. *Water Resources Research*, 52(9): 7490–7502.

Xiong, Y., Feng, W., Bai, H. et al., 2025. High-resolution terrestrial water storage anomalies and components in China from GRACE/GFO via joint inversion downscaling. *Water Resources Research*, 61(7): e2024WR038996.

https://doi.org/10.5194/essd-2026-282
Preprint. Discussion started: 9 July 2026

1017 Yang, H., Cao, W., Zhi, C. et al., 2021. Evolution of groundwater level in the North China Plain in the
1018 past 40 years and suggestions on its overexploitation treatment. *Geology in China*, 48(4): 1142–
1019 1155.

1020 Yang, W., Long, D., Scanlon, B.R. et al., 2022. Human intervention will stabilize groundwater storage
1021 across the North China Plain. *Water Resources Research*, 58(2): e2021WR030884.

1022 Yao, T., Bolch, T., Chen, D. et al., 2022. The imbalance of the Asian water tower. *Nature Reviews Earth*
1023 *& Environment*, 3(10): 618–632.

1024 Yazdian, H., Salmani-Dehaghi, N. and Alijanian, M., 2023. A spatially promoted SVM model for
1025 GRACE downscaling: Using ground and satellite-based datasets. *Journal of Hydrology*, 626:
1026 130214.

1027 Zhang, Q., Liu, C., Xu, C.-y. et al., 2006. Observed trends of annual maximum water level and
1028 streamflow during past 130 years in the Yangtze River basin, China. *Journal of hydrology*,
1029 324(1-4): 255–265.

1030 Zhao, F., Long, D., Li, X. et al., 2022. Rapid glacier mass loss in the Southeastern Tibetan Plateau since
1031 the year 2000 from satellite observations. *Remote Sensing of Environment*, 270: 112853.