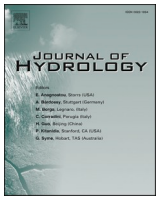
Journal of Hydrology 668 (2026) 135015

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



A dynamic soft-constrained deep learning paradigm for spatial downscaling of satellite gravimetry terrestrial water storage

Metehan Uz [a,*], Kazım Gökhan Atman [b], Orhan Akyılmaz [c], C.K. Shum [d]

[a] GFZ Helmholtz Centre for Geosciences, Potsdam, Germany
[b] School of Mathematical Sciences, Queen Mary University of London, London, England, UK
[c] Dept. of Geomatics Eng., Istanbul Technical University, Istanbul, Turkey
[d] Division of Geodetic Science, School of Earth Sciences, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

This manuscript was handled by Dan Lu, Editor-in-Chief, with the assistance of Shaoxing Mo, Associate Editor

Keywords:
Soft constraint paradigm
Mass conservation
GRACE and GRACE follow-on gravimetry
Deep learning-aided spatial downscaling
Terrestrial water storage anomalies

ABSTRACT

The Gravity Recovery and Climate Experiment (GRACE) and GRACE Follow-On (GRACE-FO) satellite gravimetry missions have contributed significantly to our knowledge of variations in Earth's Terrestrial Water Storage anomalies (TWSA) throughout the last two decades. However, the ability to quantifying hydrometeorological and other climate/weather episodes is hindered by limitations in the current TWSA spatiotemporal resolutions at monthly sampling and approximately coarser than 300 km. In this study, we used Deep Learning (DL) approach that is specifically developed for accurate and effective spatial downscaling of TWSA time series from NASA's Jet Propulsion Laboratory (JPLM). Each TWSA maps of JPLM are downscaled from 300 km to 50 km spatial resolution spanning from April 2002 through December 2022 by using inherent spatiotemporal correlations of WaterGAP Hydrology Model (WGHM) TWSA. For this purpose, a novel dynamic soft-constrained loss function is introduced and applied that adaptively balances while optimizing the TWSA signal with low-resolution JPLM observations against high-resolution spatial patterns derived from the WGHM hydrological model and ERA5 inputs. Internal validation shows that while the downscaled TWSA preserves basin-averaged temporal dynamics (trends, seasonality) from JPLM, the correlations and spectral analyses show it successfully incorporates WGHM TWSA's high-resolution spatial variability. External validation of downscaled TWSA products also demonstrates their ability to capture El Niño Southern Oscillation (ENSO)-driven interannual variability, glacial mass loss trends, spectral consistency with Soil Moisture Active Passive (SMAP) satellite-derived surface soil moisture at high-resolution band and similar predictive skill against previous studies. Furthermore, the validation against groundwater well observations indicates that the downscaled TWSA effectively represents the spatial patterns of long-term groundwater depletion in heavily stressed aquifers and significantly enhancing the spatial localization of depletion or recharging signals relative to the coarse-resolution JPLM TWSA.

1. Introduction

The monitoring of the Earth's water resources is becoming increasingly important to better understand the variability of the climate system and fresh water (Tapley et al., 2019; Humphrey et al., 2023; Rodell and Reager, 2023). TWSA has an ability to indicate the variations of Earths water cycle by defining to total amount of water that is stored in different compartments (groundwater storage (GWS), soil moisture (SMS) storage, snow water equivalent (SWE), rivers or lakes as surface water storage (SWS), glaciers and canopy water storage (CWS)), therefore, it is a key element to monitor the changes in global water resources (Scanlon et al., 2018; Girotto and Rodell, 2019; Humphrey et al., 2023). The dominance of each compartments varies spatially and temporally, for example, while SMS can be more significant on tropical climate and mid-latitude regions, GWS is also more dominant for the long-term time period (Girotto and Rodell, 2019). As a key element for water cycle, TWSA can be derived from land hydrology models (LHM) with higher spatial and temporal resolutions by simulating water fluxes or storage data from satellite-based and/or in-situ station observations (Humphrey et al., 2023; Rodell and Reager, 2023). However, there are some limitations in usage of LHM-derived TWSA due to the modeling errors that are sourced from each simulated compartments or possessing to unreliable long-term trends that is the fundamental indicator of the climate and anthropogenic changes in water cycle (Scanlon et al., 2018;

* Corresponding author.
E-mail address: metehan.uz@gfz.de (M. Uz).

https://doi.org/10.1016/j.jhydrol.2026.135015
Received 12 July 2025; Received in revised form 13 January 2026; Accepted 21 January 2026
Available online 24 January 2026

Humphrey et al., 2023, Gou and Soja, 2024).

In contrast to these indirect or model-based estimates from LHM, GRACE and GRACE-FO (GRACE/-FO) satellites are unique tools for directly monitoring changes in global TWSA with monthly temporal and ~300 km spatial resolutions (Tapley et al., 2019; Chen et al., 2022), respectively. GRACE mission ended in October 2017 primarily because of an aging battery problem on GRACE-B, however, GRACE-FO mission was initiated in May 2018 to maintain the continuity of the GRACE project and still continues to its mission to provide mass change information (Flechtner et al., 2016; Kornfeld et al., 2019; Tapley et al., 2019). Therefore, over the past two decades, GRACE/-FO observations have yielded consistent insights into not only the hydrosphere but also providing informations for cryosphere, ocean and solid Earth studies (Tapley et al., 2019; Chen et al., 2022). In addition, World Meteorological Organization has officially identified TWSA as an Essential Climate Variable within the 2022 implementation plan of the Global Climate Observing System program (Zemp, 2022). Nowadays, the importance of directly observed TWSA or uniqueness of GRACE/-FO missions is even more pronounced. However, there are particular limitations, which are originated from the inherent GRACE/-FO orbit configuration and/or observation quality, in terms of coarse spatial (~300 km) and temporal (monthly) resolutions of GRACE/-FO observation-derived TWSA (Tapley et al., 2019; Chen et al., 2022). Thus, the coarse resolutions of GRACE/-FO TWSA are considered to be insufficient for monitoring hydrologic events demanding higher spatial and temporal resolutions (Girotto and Rodell, 2019; Pascal et al., 2022; Humphrey et al., 2023; Gerdener et al., 2023a; Tourian et al., 2023; Gou and Soja, 2024; Zhang et al., 2024).

In order to overcome these limitations, there have been significant efforts to downscale GRACE/-FO TWSA time series by implementing either dynamic "model-based" or statistical "data-based" approaches (Pascal et al., 2022; Tourian et al., 2023). These two methodologies represent two distinct philosophies for integrating or downscaling GRACE/-FO TWSA with LHM. The model calibration or assimilation of GRACE/-FO TWSA data into LHM to achieve higher resolution is considered as model-based downscaling (Zaitchik et al., 2008; Houborg et al., 2012; Sun et al., 2012; Eicker et al., 2014; Schumacher et al., 2018; Li et al., 2019; Gerdener et al., 2023a) and uses GRACE/-FO dataset as a constraint to directly adjust and improve the internal physical parameters of a LHM to produce a more physically consistent forward model. On the other hand, the data-based approaches also have been emerged to achieve higher resolution TWSAs through state-of-art statistical approaches (Pascal et al., 2022; Gou and Soja, 2024; Uz et al., 2024a; Zhang et al., 2024) and treats the LR GRACE/-FO TWSA and the HR LHM dataset as separate data sources to be optimally fused to generate a new TWSA time series. However, each downscaling approach has its own specific drawbacks. For example, the assimilation approaches could have the discrepancies between GRACE/-FO TWSA and LHM input data resolutions and could lack the representation of anthropogenic processes. The most of statistical approaches are based on the assumption that the hydrological and physical processes are identical at all resolutions and ignore non-linear relationships between high-resolution (HR) input and low-resolution (LR) output. In addition, the statistical downscaling approaches pragmatically assumes that the HR LHM yields sufficiently realistic spatial patterns, even though LHMs have biases in magnitude and long-term trends that can be corrected at large scales by GRACE/-FO TWSA (Sun, 2024). These drawbacks cause simulation of unrealistic downscaled dataset and inability to calculate proper uncertainty estimates (Pascal et al., 2022; Tourian et al., 2023; Gerdener et al., 2023a). Therefore, most of the existing downscaled products have the limitations as only focusing to local/regional scale and weakened generalizability as well as the lack of constraining paradigm that is based on hydrologic process (Gerdener et al., 2023a; Gou and Soja, 2024; Zhang et al., 2024).

Addressing the critical limitation in spatial resolution of TWSA and drawbacks of former downscaling approaches, this study introduces a novel dynamic soft-constrained loss function to downscale spatial resolution of JPLM TWSA (Watkins et al., 2015). This proposed loss function is applied via a hybrid Variational U-Net architecture, which is based on combining the spatial mapping strengths of U-Net (Ronneberger et al., 2015) with the probabilistic modeling capabilities of a (VAE) Variational Autoencoder (Kingma and Welling, 2013). The main idea of our loss function is to achieve downscaling of TWSA by applying observational constraints (as soft-constrained) from LR targeted output (JPLM TWSA) with HR spatial information of input drivers such as WGHM TWSA (Müller Schmied et al., 2023) and ERA5 climate reanalysis data (Hersbach et al., 2023). Besides, it can adaptively balance the coherence between downscaled and JPLM TWSA considering the spatial variability of WGHM TWSA through a dynamic weighting factor, that is, Pearson correlations between downscaled and WGHM TWSAs and updated in each iteration of training runs. With our proposed DL framework, the monthly JPLM TWSA time series from April 2002 to December 2022 were spatially downscaled to 50 km resolution along with uncertainty estimates.

The downscaled TWSAs are internally and externally validated by rigorous comparisons. The internal validation is based on assessing the coherence of downscaled TWSA relative to the source datasets, i.e., JPLM and WGHM TWSAs, and includes comparing the basin-averaged time series across diverse climate zones from humid to arid, analyzing derived temporal components (linear trend, annual and semi-annual amplitude/phase) and evaluating spatial power spectra of TWSA signal across different spatial resolutions. On the other hand, the external validation leverages independent datasets and phenomena to assess the realism and utility of downscaled dataset and involves comparing downscaled TWSA's response to ENSO events against observed teleconnections, evaluating its representation of glacier mass change trends against the independent glaciers related dataset of Global Gravity-based Groundwater Product in key high-mountain regions, assessing spatial coherence with high-resolution SMAP surface soil moisture anomalies and benchmarking its performance using Nash-Sutcliffe Efficiency (NSE – (Nash and Sutcliffe, 1970)) against downscaled and assimilated TWSA products from previous studies relative to both JPLM and WGHM. Finally, the downscaled TWSA's ability to represent long-term groundwater variations is rigorously validated by comparing its derived GWS anomalies (GWSA) against a comprehensive network of in-situ monitoring wells across major Contiguous United States (CONUS) aquifers.

The findings robustly demonstrate the efficacy of the proposed methodology and the downscaled TWSA successfully conserves the large-scale signals, long-term trends, dominant seasonal cycles, and interannual variability (including ENSO responses and major hydrological events) that are inherent in the GRACE/-FO observations. In addition, it also effectively incorporates the fine-scale spatial variability and structural detailed characteristic of the HR WGHM TWSA as confirmed by correlation analyses (PCC and a new correlation coefficients of Chatterjee (2021)) and spectral comparisons. Notably, the downscaled dataset realistically represents glacier mass loss trends derived from GRACE/-FO, despite this signal being absent in the WGHM input. Overall, the study reveals that the proposed soft-constrained DL paradigm yields an improved HR TWSA dataset offering enhanced spatial details and robust performance across diverse hydro-climatic conditions and signal types, making it a valuable tool for regional water resource assessments.

2. Data

2.1. JPL mascon data

JPLM TWSA data sets are calculated from GRACE/-FO observations using the mascon solution techniques as described by (Watkins et al., 2015; Wiese et al., 2016; Landerer et al., 2020). The time series of the dataset consist of two separate periods that are from April 2002 to June

2017 for the GRACE mission, and from May 2018 to the present (up to December 2022 in this study) for the GRACE-FO mission. A data gap exists between the GRACE and GRACE-FO missions, spanning an 11-month period from June 2017 to May 2018. In addition, intermittent data gaps are also present within the data records of each individual mission. All standard post-processing corrections are applied and each monthly TWSA are calculated by removing the mean TWSA that is calculated as the average of monthly solutions between 2004.0 and 2009.999. The JPLM TWSA utilizes defined limitations in both spatial and temporal dimensions to calculate global, monthly gravity fields using equal-area 3° × 3° (300 km × 300 km, half wavelength) spherical cap Mascon functions. This approach aims to reduce the impact of measurement errors and the data has not undergone any extra empirical destriping filter. Thus, although the JPLM dataset has a monthly temporal resolution, it covers the entire globe and is resampled to a gridded dataset with a spatial resolution of 0.5° × 0.5° (50 km × 50 km) (Watkins et al., 2015; Wiese et al., 2023). Since GRACE/-FO have intrinsic spatial resolution ~300 km, JPLM TWSAs at 50 km spatial resolution is a fixed gridded dataset and does not represent the actual resolution of GRACE/-FO. Therefore, since we anticipated that the simulated TWSAs by our DL paradigm would experience spatial downscaling, thanks to the soft constrained structure of the algorithm and the ability of the setting up interaction between input and output data, we used these fixed JPLM TWSAs at a resolution of 50 km as the output data in our DL architecture.

2.2. ERA5 model data

The monthly ERA5 dataset is released by the European Centre for Medium-Range Weather Forecast (ECMWF – https://cds.climate.copernicus.eu/) with the native spatial resolution as 0.25° × 0.25° (25 km × 25 km) (Hersbach et al., 2023). We used these monthly single level versions of ERA5 dataset covering the entire time period of the study. The key hydrometeorological variables, which are precipitation (P), evaporation (E), runoff (R), temperature (T), and water storage related dataset (Canopy Water Storage Anomalies – CWSA, Snow Water Storage Anomalies – SnWSA and Soil Moisture Storage Anomalies – SMSA) were downloaded from the ECMWF Climate Data Store. To ensure consistency with the output dataset (JPLM TWSA), all ERA5 variables were spatially resampled from their native resolution of 25 km to 50 km using linear interpolation. Furthermore, the anomalies of these variables were calculated by removing their mean values computed over the baseline period from 2004.000 to 2009.999. Two additional variables TWSA and Cumulative Water Storage Change (CWSC) were also computed to serve as inputs for the DL framework. ERA5 TWSA was derived by aggregating anomalies in SMSA, SWE, and CWSA in accordance with Eq. (1):

$$TWSA = SMSA + SnWSA + CWSA \tag{1}$$

The CWSC was calculated using the water balance equation, which accumulates net inflow (P) and outflows (E and R) for each monthly grid cell over time, as expressed in Eq. (2):

$$CWSC_t = \sum_{i=1}^{t} (P_i - E_i - R_i) \tag{2}$$

Thus, ERA5-derived P, T, TWSA, and CWSC variables were selected as the primary input datasets for the DL model architecture developed in this study.

2.3. WaterGAP global hydrology model data

WGHM is a comprehensive hydrological model that provides a detailed description of water storage, consumption and resources in all terrestrial regions (excluding Antarctica). The recent releases of WGHM, version 2.2e, has been released through the website – https://gude.uni-frankfurt.de/. This latest version provides a monthly TWS as one of the regular outputs of the WGHM and is released with a spatial resolution of 50 km × 50 km from 1901 to December 2022. The TWS product represents many components, including canopy, snow, soil, groundwater, wetland, lake, reservoir, and river storages (Müller Schmied et al., 2023; Gerdener et al., 2023a; Gou and Soja, 2024). To achieve a consistent temporal baseline for both the JPLM and ERA5 datasets, we also calculated WGHM TWSA by subtracting from the average values computed for the period between 2004.000 and 2009.999 and used the time series from April 2002 to December 2022. Due to the absence of GRACE/-FO measurements in the WGHM model, both WGHM and GRACE/-FO TWSAs are completely independent and not influenced by each other (Müller Schmied et al., 2023). Therefore, WGHM TWSA is useful to employ as an input in our DL architecture to obtain higher spatiotemporal information. However, while WGHM TWSAs exhibit a higher level of detail in their global structures compared to GRACE/-FO products, they are nevertheless subject to two significant limitations. Initially, WGHM simulations exhibit a higher level of noise as a result of imperfect simulation methodology. Moreover, WGHM TWSAs demonstrate lower accuracy when compared to the GRACE/-FO TWSAs because they do not depend on real observations (Gou and Soja, 2024). Unlike WGHM, which lacks data on mountain glaciers (Müller Schmied et al., 2023), GRACE/-FO has the ability to observe total water mass change, i.e. the sum of the water content in all compartment that forms the TWS, to the extent of allowable spatial resolutions from satellite gravimetry observations.

2.4. Rationale for input–output data selection

The specific roles, which is defined in the proposed DL paradigm, with the strengths and limitations of each input and output dataset are defined as the selection factor of these datasets for our DL architecture. The main aim of the study is that the simulation of the spatially downscaled GRACE/-FO –like time series. JPLM TWSA are chosen as LR output, since they are designed to reduce the impact of measurement errors and minimize signal leakage without the need for empirical destriping filters by providing a high-quality and observationally-driven representation of large-scale mass changes. Therefore, its role in the DL paradigm is established as the reference point for our downscaled TWSA. WGHM is chosen as the primary input and source of HR spatial information for two key reasons. First, it is a comprehensive hydrological model that is provided at the targeted 50 km resolution by representing numerous water storage compartments. Second, it is completely independent from GRACE/-FO observations, which is crucial for avoiding repetition in the learning process. On the other hand, WGHM has limitations by including a higher noise level and a well-documented tendency to underestimate long-term trends compared to GRACE observations. However, our methodology is designed specifically to address these strength and limitations of WGHM time series. The proposed dynamic soft-constrained loss function leverages WGHM TWSA for its spatial information, i.e., HR spatial patterns and gradients, while simultaneously using the JPLM TWSA to correct for biases in magnitude and long-term trends. In addition, the variables from the ERA5 reanalysis dataset (ER5-derived TWSA, CWSC, P and T) are chosen as HR auxiliary inputs. Their purpose is to provide the HR additional and physically consistent climate information to the DL architecture. This allows the DL model to learn more complex and non-linear relationships between climate drivers and water storage that might not be captured by two primary TWSA dataset, i. e., JPLM and WGHM. As a result, DL framework is designed by strategically utilizing the unique strengths of each dataset that is explained as the observational accuracy of JPLM TWSA, the higher spatial resolution of WGHM TWSA and the climatic context of chosen ERA5 input.

3. Method

3.1. Network architecture of deep learning model

We used a hybrid DL model that combines elements of both U-NET (Ronneberger et al., 2015) and VAE (Kingma and Welling, 2013) frameworks and is called probabilistic U-Net or Variational U-Net. This type of model aims to leverage the strengths of both architectures: the U-Net's ability to learn detailed spatial mappings via image-to-image regression task by preserving fine details and the VAE's capability for generative modeling by learning latent representations and potentially providing uncertainty estimates. The comprehensive details of the architecture which integrates a U-Net structure with a VAE framework for the task of spatial downscaling are illustrated in Supplementary Fig. S1. The DL model accepts a multi-channel input tensor, $\boldsymbol{X}$, which consists of 5 channels representing various hydro-climatic inputs including WGHM TWSA and ERA5-derived P, T, CWSC, TWSA at 50 km spatial resolution. Thus, the dimension (also called batch size) of $\boldsymbol{X}$ is (360, 720, 5) and propagates through three primary stages; an encoding path, a probabilistic bottleneck (latent distribution), and a decoding path during a forward pass. These steps can be explained as follows:

- The encoder progressively down samples the input through a series of convolutional and pooling layers capturing hierarchical features while generating intermediate feature maps at multiple scales serving as the contracting path of the Variational U-Net architecture. Therefore, its primary function is to hierarchically extract and compress spatial features from the input tensor by generating a condensed representation while capturing contextual information across multiple scales. The structure of encoder comprises a sequence of 6 encoder blocks and each block executes a defined sequence of operations as the feature maps are processed by a convolution layer with 3x3 kernel size following by a Rectified Linear Unit (ReLU) activation to learn localized spatial patterns. Subsequently, the spatial dimensions are reduced via a max pooling layer operation with varying pool sizes (e.g., 2 × 2, 3 × 3, 5 × 5) and the same padding effectively increasing the receptive field for subsequent layers. Concurrently, the number of feature channels (namely filters) is increased from 8 up to 256 to capture more complex representations at coarser resolutions. In addition, a Dropout layer is applied after the max pooling layer ensuring regularization in DL framework to mitigate overfitting and to facilitate epistemic uncertainty estimation via implementation of Monte Carlo Dropout (MCDO) during inference (Gal and Ghahramani, 2016; Kendall and Gal, 2017). Finally, the encoder blocks return with two key outputs, first, a tuple containing the feature maps (namely $\boldsymbol{f}_1\boldsymbol{f}_2\cdots\boldsymbol{f}_6$) within each encoder blocks prior to pooling layers, which serve as inputs for the skip connections to the decoder. Second, maximally downsampled feature map (namely $\boldsymbol{p}_6$) resulting from the last encoder block and which is passed to the bottleneck stage.
- The probabilistic bottleneck stage forms the crucial interface between the encoder and decoder pathways and incorporates the variational inference mechanism that is central to our DL architecture, which deviates from a standard U-Net by incorporating VAE principles. It processes the most compressed feature representation from the encoder to yield the parameters (mean and log-variance) of a latent probability distribution. A stochastic latent variable is then sampled from this distribution using the reparameterization to enable gradient backpropagation. The stage starts receiving the most compressed feature map ($\boldsymbol{p}_6$) from the final encoder block. Initially, this feature map undergoes further transformation via the Bottleneck class which applies a sequence of convolutional block operations, comprising convolutional and activation layers to refine the representation. The output of this standard bottleneck processing is then projected into the latent space. This is achieved using two distinct convolutional layers by computing the mean vector, $\boldsymbol{Z}_{\boldsymbol{\mu}}$, of the approximate posterior distribution and the logarithmic of its variance vector, $\boldsymbol{Z}_{\boldsymbol{\sigma}} = \log(\sigma^2)$, respectively. Both layers utilize a 3x3 kernel and linear activation outputting feature maps with a channel depth equal to the specified latent dimension (LD). Subsequently, the reparameterization is employed via a Lambda layer executing the sampling function taking $\boldsymbol{Z}_{\boldsymbol{\mu}}$ and $\boldsymbol{Z}_{\boldsymbol{\sigma}}$ as input and generates a stochastic sample $\boldsymbol{Z}$ from the learned Gaussian distribution as $\boldsymbol{Z} = \boldsymbol{\mu} + \boldsymbol{\sigma}\cdot\boldsymbol{\varepsilon} = \boldsymbol{Z}_{\boldsymbol{\mu}} + \exp(0.5^*\boldsymbol{Z}_{\boldsymbol{\sigma}}) \bullet \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is standard Gaussian noise. As a result, this sampling process introduces the stochasticity while ensuring the entire model remains differentiable allowing the gradients to flow back through the sampling step during training. Thus, the resulting latent vector $\boldsymbol{Z}$ encapsulates the learned probabilistic representation and serves as the input to the subsequent decoder path.
- The decoder constitutes the expansive path of the Variational U-Net architecture and responsible for reconstructing the HR target map from the latent representation, $\boldsymbol{Z}$, and multi-scale features provided by the encoder. Decoder mirrors the encoder's structure symmetrically and consists of a sequence of 6 decoder blocks followed by a final convolutional layer. As in encoder blocks each decoder block performs several operations sequentially: first, transpose convolutional layer is applied to upsample the feature map that are received from the preceding layer. These steps are effectively increased the spatial resolution considering strides as 5 × 5, 3 × 3 and 2 × 2 to invert the pooling operations in the corresponding encoder stages. The upsampled feature map is then concatenated along the channel dimension with the corresponding high-resolution feature map, i.e., from $\boldsymbol{f}_6$ to $\boldsymbol{f}_1$ that are passed from the encoder via skip connections. This concatenation step is critical as it re-introduces fine-grained spatial details lost during encoding. After that, a dropout layer is also applied for regularization and MCDO inference in same manner in encoder stage. Finally, a convolutional block, which consists of convolutional and ReLU activation layers, refines the combined features. In contrast to encoder stage, the number of feature channels is typically halved starting from 256 to reaching up to 8 as the spatial resolution increases through the decoder stages. The processing is ended via final convolutional and linear activation layer to complete this regression task and employing 1x1 kernel to map the feature representation from the last decoder block to the final single-channel and high-resolution TWSA prediction, $\widehat{\boldsymbol{Y}}$.

As a result, while U-Net structure provides the powerful encoder-decoder framework with skip connections essential for spatial mapping tasks, the VAE component integrated into the bottleneck adds a probabilistic layer allowing the model to learn a distribution over downscaled outputs rather than just a single deterministic output. This facilitates the modeling total uncertainty by providing the aleatoric uncertainty via latent sampling and adds a regularization component, i.e., Kullback–Leibler (KL) divergence, to the training process. Therefore, our DL algorithm network incorporates mechanisms to quantify two distinct types of predictive uncertainty: aleatoric and epistemic uncertainties to provide a more comprehensive assessment of prediction reliability. In other words, the aleatoric uncertainty quantification is applied by VAE latent sampling stage by the parameters $\boldsymbol{Z}_{\boldsymbol{\mu}}$ and $\boldsymbol{Z}_{\boldsymbol{\sigma}}$ that define a probability distribution (assuming Gaussian) in the latent space.

According to the fundamental principle of Bayesian theorem-based uncertainty quantification (UQ) methods, they ensure transforming a deterministic model into a stochastic model to enable probabilistic simulations (Wursthorn et al., 2022). MCDO is a variational inference technique that approximates Bayesian inference by sampling from the model's posterior distribution (Gal and Ghahramani, 2016; Gawlikowski et al., 2023). Thus, it can be applied to quantify epistemic uncertainties utilizing dropout regularization in the training phase to prevent overfitting (Hinton et al., 2012; Srivastava et al., 2014) and generating multiple prediction samples for the same input data through

an iterative process (Gal and Ghahramani, 2016). These samples are generated by enabling the dropout layer during the inference timing of DL framework, especially during the prediction or simulation step. This is achieved by conducting multiple forward passes using the same input data, leading to the generation of different samples. Thus, epistemic uncertainty is estimated by calculating the standard deviation of the model's predictive distribution (Gal and Ghahramani, 2016) by averaging the sampled predictions as shown in Supplementary Fig. S1. As a result, total predictive uncertainty is calculated from the aleatoric and epistemic components via square root of the sum of variances assuming independence between the sources of these two types of uncertainty.

3.2. Spatial downscaling strategy of soft-constrained paradigm

The main objective of spatial downscaling of TWSA time series with our DL algorithm depends on two factors; (i) the recovery of the spatiotemporal variation of all compartments within the water cycle (groundwater, snow, surface water, soil moisture) with GRACE/-FO TWSA based on realistic observations, and (ii) overcoming the LR of GRACE/-FO TWSAs with the HR spatiotemporal localization information that is derived from GRACE-independent LHM TWSAs. Therefore, we used soft-constrained paradigm defining a new loss function to implement via our variational U-NET algorithm. Soft constraining has been successfully applied before to a variety of DL based modeling tasks (Beucler et al., 2021; Harder et al., 2023) and is achieved by incorporating a regularization component into the loss function. Hence, this approach is alternatively referred to as loss constrained method (Beucler et al., 2021). However, the optimization of our Variational U-Net DL algorithm is guided by a composite dynamically weighted loss function designed to synergistically integrate information from both LR target observations (JPLM TWSA) and HR model inputs (WGHM TWSA, ERA5). Therefore, we applied this soft constrained paradigm via novel inverse dynamic weighting mechanism by using the regularization term, λ, as given by,

$$\mathscr{L}_{Total} = \lambda \mathscr{L}_{RC} + (1-\lambda)\mathscr{L}_{CV} \tag{3}$$

where $\mathscr{L}_{RC}$ is defined as reconstruction loss, $\mathscr{L}_{CV}$ is the constrain violation loss. Generally, λ is considered as a fixed value and it is aimed to investigate how it affects the results according to certain values as in the study of Irrgang et al. (2020) or which acts as a hyperparameter that must be tuned at each training iteration as in the study of (Gou and Soja, 2023). Our proposed loss function also accounts for correlations between both downscaled simulation and chosen input data at each iteration ensuring that the DL algorithm maintains consistency across scales. However, these correlations are dynamically adjusted in a reverse manner when compared to previous studies and serve as regularization terms allowing the loss function to adapt based on the degree of agreement between the downscaled simulations and both LR JPLM TWSA and HR WGHM TWSA. This adaptive regularization framework is critical for refining the loss function to optimize the downscaling performance.

$\mathscr{L}_{RC}$ consists of both the low-resolution loss, $\mathscr{L}_{LR}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})$, and KL divergence loss, $\mathscr{L}_{KL}(\boldsymbol{Z}, \boldsymbol{X})$ and which can be given for each batch, $B_k (k = 1, 2, \cdots, BS)$ as,

$$\mathscr{L}^{B_k}_{RC}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}, \boldsymbol{Z}, \boldsymbol{X}) = \left[\mathscr{L}^{B_k}_{LR}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) + \mathscr{L}^{B_k}_{KL}(\boldsymbol{Z}, \boldsymbol{X})\right] \tag{4}$$

$\mathscr{L}^{B_k}_{LR}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})$ is designed to enforce mass conservation as soft-constrained manner (which is not strictly hard-constrained application, since it is only applied in loss function) between $\boldsymbol{Y}$ and $\widehat{\boldsymbol{Y}}$ by aggregating TWSAs in each 6x6 kernel using 2D average pooling function, P. We chose 6x6 kernel size to calculate mean values of each window of P function, because which corresponds to 300 km i.e. the coarse spatial resolution of JPLM TWSA. Within this operation the spatial resolution of $\widehat{\boldsymbol{Y}}$ is reduced effectively to match the coarser resolution of the LR observations $\boldsymbol{Y}$, e.g., JPLM TWSA. After that the Mean Squared Error (MSE) is calculated between these two resulting LR windows. The squared differences between corresponding grids in $P(\widehat{\boldsymbol{Y}})$ and $P(\boldsymbol{Y})$ are calculated and then averaged over all grids in each window separately across all samples within the batch samples. $\mathscr{L}^{B_k}_{LR}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}})$ is given only for batch sample B_k as

$$\mathscr{L}^{B_k}_{LR}(\boldsymbol{Y}, \widehat{\boldsymbol{Y}}) = \frac{1}{N_{pg}} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \left[P(\widehat{Y}_{ij})^{B_k} - P(Y_{ij})^{B_k}\right]^2 \tag{5}$$

where $H' = H/k$ and $W' = W/k$ are the height and width of the low-resolution pooled windows, $N_{pg} = H' \times W'$ is the total number of grids in each pooled window. On the other hand, $\mathscr{L}^{B_k}_{KL}(\boldsymbol{Z}, \boldsymbol{X})$ defines KL Divergence Loss which is a fundamental component derived from the theoretical underpinnings of VAEs. Its primary purpose within this framework is not only directly related to reconstruction accuracy but also serves as a crucial regularization term acting upon the latent space. It quantifies the dissimilarity between the probability distribution learned by the encoder as the approximate posterior assuming as Gaussian, $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$, and a predefined prior distribution over the latent variables $p(\boldsymbol{Z})$. Within this minimization process the KL divergence encourages the distributions encoded for different inputs to collectively resemble the standard Gaussian structure, effectively regularizing the complexity of the information encoded in the latent space. In our case, we make the variational assumption that the approximate posterior $\mathrm{q}_\phi(\mathbf{Z}|\mathbf{X})$ follows a multivariate Gaussian distribution with a diagonal covariance matrix. The encoder network learns the parameters of this distribution, namely the mean $\mathbf{Z}_{\boldsymbol{\mu}}$ and the standard deviation $\mathbf{Z}_\sigma$. This formulation, combined with a standard multivariate Gaussian prior $\mathrm{p}(\mathbf{Z}) = \mathcal{N}(0, \mathbf{I})$, allows the KL divergence $\mathrm{D}_{\mathrm{KL}}\left(\text{posterior}\, \mathrm{q}_\phi(\mathbf{Z}|\mathbf{X}) \| \mathrm{p}(\mathbf{Z})\right)$ to be calculated in closed form. Furthermore, to enable differentiable sampling from $\mathrm{q}_\phi(\mathbf{Z}|\mathbf{X})$ during training, we employ the reparameterization trick $\mathbf{Z} = \mathbf{Z}_{\boldsymbol{\mu}} + \mathbf{Z}_\sigma \odot \epsilon$, generating samples where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Therefore, the KL divergence across the N_l dimensions for a single latent vector (namely latent dimension that is accepted as one of hyperparameters of this study) corresponding to a spatial position (k, l) in the latent map for batch sample B_k is:

$$D_{KL}\left(q_\phi\left(Z^{B_k}_{k,l}|X^{B_k}\right) \| p\left(Z^{B_k}_{k,l}\right)\right) = -\frac{1}{2} \sum_{d=1}^{N_l} \left[1 + Z^{B_k}_{\sigma_{k,l,d}} - \left(Z^{B_k}_{\mu_{k,l,d}}\right)^2 - \exp\left(Z^{B_k}_{\sigma_{k,l,d}}\right)\right] \tag{6.1}$$

The final KL divergence loss LKL is the average of this quantity over all latent spatial positions (k, l),

$$\mathscr{L}^{B_k}_{KL}\left(q_\phi(\boldsymbol{Z}|\boldsymbol{X}) \| p(\boldsymbol{Z})\right) = \frac{1}{N_{lg}} \sum_{k,l} D_{KL}\left(q_\phi\left(Z^{B_k}_{k,l}|X^{B_k}\right) \| p\left(Z^{B_k}_{k,l}\right)\right) \tag{6.2}$$

$$\mathscr{L}^{B_k}_{KL}\left(q_\phi(\boldsymbol{Z}|\boldsymbol{X}) \| p(\boldsymbol{Z})\right) = \frac{1}{N_{lg}} \sum_{k=1}^{H^L} \sum_{l=1}^{W^L} \left[-\frac{1}{2N_l} \sum_{d=1}^{N_l} \left[1 + Z^{B_k}_{\sigma_{k,l,d}} - \left(Z^{B_k}_{\mu_{k,l,d}}\right)^2 - \exp\left(Z^{B_k}_{\sigma_{k,l,d}}\right)\right]\right] \tag{6.3}$$

where H^L and W^L are the height and width of the latent feature map, $N_{lg} = H^L \times W^L$ is the total number of spatial positions in the latent map.

On the other hand, $\mathscr{L}_{CV}$ consists of both high-resolution loss, $\mathscr{L}_{HR}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right)$, and gradient difference loss, $\mathscr{L}_{GR}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right)$ and is expressed for sample batch as

$$\mathscr{L}^{B_k}_{CV}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right) = \left[\mathscr{L}^{B_k}_{HR}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right) + \mathscr{L}^{B_k}_{GR}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right)\right] \tag{7}$$

with WGHM TWSA, $\dot{\boldsymbol{X}}$. $\mathscr{L}^{B_k}_{HR}\left(\dot{\boldsymbol{X}}, \widehat{\boldsymbol{Y}}\right)$ term serves as a direct constraint

within the composite loss function compelling the model's HR simulation to closely match the grid values of a provided HR WGHM TWSA. In other words, the DL algorithm is forced to learn the spatial variations/patterns that are derived from WGHM TWSA at targeted resolution for simulated TWSA. $\mathscr{L}_{HR}^{B_k}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)$ is calculated as MSE,

$$\mathscr{L}_{HR}^{B_k}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)=\frac{1}{N_g}\sum_{i=1}^{H}\sum_{j=1}^{W}\left[\dot{X}_{i,j}^{B_k}-\widehat{Y}_{i,j}^{B_k}\right]^2 \tag{8}$$

and is applied considering all grids in height (H) and width (W) of both $\dot{\boldsymbol{X}}$ and $\widehat{\boldsymbol{Y}}$ with $N_g=H\times W$ is the total number of grids in one of $\dot{X}$ or $\widehat{Y}$. $\mathscr{L}_{GR}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)$ term is designed to enforce similarity in spatial patterns between $\widehat{\boldsymbol{Y}}$ and $\dot{\boldsymbol{X}}$ by minimizing the difference between the spatial gradients. Thus, we encourage our DL model to learn the patterns of spatial variations that are found in WGHM TWSA. For this purpose, the vertical (G_x) and horizontal (G_y) gradients are calculated using Sobel Filters for both $G(\widehat{\boldsymbol{Y}})=\left[G_x(\widehat{\boldsymbol{Y}})G_y(\widehat{\boldsymbol{Y}})\right]$ and $G\left(\dot{\boldsymbol{X}}\right)=\left[G_x\left(\dot{\boldsymbol{X}}\right)G_y\left(\dot{\boldsymbol{X}}\right)\right]$. After that element-wise squared differences between the two gradient tensors are computed as $\Delta G\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)^2=\left[G\left(\dot{\boldsymbol{X}}\right)-G(\widehat{\boldsymbol{Y}})\right]^2$. The intermediate pooling step is then applied to these squared differences of each gradient components using 2D average pooling function, P, as in $\mathscr{L}_{LR}$. Thus, while $\mathscr{L}_{LR}$ aims to optimize mass conservation in soft-constrained manner with JPLM TWSA, $\mathscr{L}_{GR}$ aims to take into account the spatial variations within each windowed 6x6 kernels. $\mathscr{L}_{GR}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)$ is given as,

$$\mathscr{L}_{GR}^{B_k}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)=\frac{1}{2N_{pg}}\sum_{i=1}^{H'}\sum_{j=1}^{W'}\left[P\left((\Delta G_y)^2\right)_{i,j}^{B_k}+P\left((\Delta G_x)^2\right)_{i,j}^{B_k}\right] \tag{9}$$

with H' and W' of the pooled windows as well as the total number of grids in each pooled window, N_{pg}.

As a result, total loss can be given for each batch samples as,

$$\mathscr{L}_{Total}\left(\boldsymbol{Y},\widehat{\boldsymbol{Y}},\boldsymbol{Z},\boldsymbol{X},\dot{\boldsymbol{X}}\right)=\frac{1}{N_B}\sum_{B_k=1}^{N_B}\left[\lambda\mathscr{L}_{RC}^{B_k}(\boldsymbol{Y},\widehat{\boldsymbol{Y}},\boldsymbol{Z},\boldsymbol{X})+(1-\lambda)\mathscr{L}_{CV}^{B_k}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right)\right] \tag{10}$$

considering the total number of samples in the batch, N_B. In addition, the regularization parameter, λ is determined by calculating using Eq. (11), as the overall average of 2D Pearson correlations (which is calculated as 2 dimensional considering the size H x W) between $\widehat{\boldsymbol{Y}}$ and $\dot{\boldsymbol{X}}$ TWSAs in each batch samples,

$$\lambda=\frac{1}{N_B}\sum_{B_k=1}^{N_B}\rho^{B_k}\left(\dot{\boldsymbol{X}},\widehat{\boldsymbol{Y}}\right) \tag{11}$$

It is anticipated that increasing the resolution of the simulations at each iteration will enhance the similarity to the $\dot{\boldsymbol{X}}$. Thus, in order to establish a connection definition with both $\dot{\boldsymbol{X}}$ and $\boldsymbol{Y}$ TWSA, the regularization parameter is incorporated as a factor and denoted as $(1-\lambda)$, in the constraint violation component of Eq. (10). Thus, as λ increases throughout each iteration, its similarity to the $\dot{\boldsymbol{X}}$ is regulated to a certain degree. However, to retain a similarity to the $\boldsymbol{Y}$, a straight multiplication with the regularization parameter is used to increase the contribution of $\boldsymbol{Y}$ when λ is increased at each step. This enables the implementation of our enhanced DL strategy, which allows the downscaling using both LR GRACE/-FO TWSAs considered as observed data and the integration of HR spatial variations of TWSA derived from hydrological models. The efficiency of the soft constrained paradigm via this proposed loss function is based on the inverse dynamic balancing of these two different loss terms.

4. Results and discussions

Monthly TWSA time series for terrestrial regions were simulated excluding Greenland and Antarctica from April 2002 to December 2022 with our proposed DL framework. In addition, the uncertainties were determined as discussed in Section 3. There are totally 249 months from the beginning to the end of study time period, 216 out of which has monthly solutions available from GRACE/-FO. Out of the total 216 months of JPLM TWSA, 180 samples are randomly chosen to use in training, while the other 36 samples are allocated for testing. Thus, approximately 15 percent of the total output dataset corresponds to the selected months for testing. Additionally, there are 33 months of data gaps, which include an 11-month gap between GRACE and GRACE-FO throughout the study time period. The DL algorithm is also employed to simulate these data gap of 33 months. On the other hand, 249 months of ERA5 data and WGHM TWSAs are chosen to be used as input. Within these data architecture, our processing chain consists of three phases that are training, testing, and prediction. During training, DL algorithm optimizes its parameters to minimize the errors considering $\mathscr{L}_{Total}$ values at each iteration. The model learns to extract and exploit the spatiotemporal dependencies present in the 5-channel input data, which includes WGHM TWSA and ERA5 hydro-climatic variables. This iterative estimation procedure is shown in Supplementary Fig. S1 and is carried out repeatedly until the training criteria are satisfied. Upon convergence of the iterative training process the final set of weights is utilized to simulate downscaled TWSA across a 249-month period extending from April 2002 to December 2022. This simulation period includes 36 months designated for testing and 33 months characterized by data gaps.

Furthermore, our DL algorithm possesses certain hyperparameters that are either directly or indirectly tuned by the algorithm itself. Some of the hyperparameters, which are dropout (DO) rate and latent dimension (LD), of the Variational U-NET are chosen to tune for choosing the best configuration of DL architecture. These parameters are crucial since they are directly related to uncertainty quantification. DO rate and the iteration number of the MCDO approach (N_{iter}) during the inference step are critical variables for determining the epistemic uncertainties of the simulations that can be also considered as model uncertainties of the DL algorithm when implementing the MCDO approach (Gal and Ghahramani, 2016; Kendall and Gal, 2017). On the other hand, LD defines the size of the bottleneck and is also important parameter to quantify aleatoric uncertainties. A smaller LD forces DL algorithm to learn a more compressed and separated representation of the input data facilitating the capture of the most prominent features. However, even the greater LD provides more capacity to learn more details in representation of input data, it might lead to a less interpretable latent space. In addition to these, batch size (BS) is a significant hyperparameter to capture the temporal correlations among input data points, especially when dealing with time-dependent inputs within each batch (Uz et al., 2024a). Thus, BS determines the temporal correlation distance of the input data, which significantly affects simulation results.

The results of the experiments on hyperparameter tuning and model configuration selection for Variational U-NET algorithm are analyzed applying grid search to test different configurations. For this purpose, the grid search is designed considering DO (which is chosen as 0.05, 0.15 and 0.25) and LD parameter (which is chosen as 16,32 and 64). This range for the DO was selected based on our experience from a related previous study (Uz et al., 2024a), which demonstrated its effectiveness in regularizing similar DL architectures for hydrological applications. A total of 9 different configurations were tested separately and the details are given in Supplementary Section 1. The evaluation led to the following configuration for the final hyperparameter selection as 0.05 and 32 for DO and LD, respectively. Moreover, BS = 12 and N_{iter} = 30 is chosen as constant values for each configuration, since they are enough to provide significant result as in our previous study (Uz et al., 2024a). In addition to these specific hyperparameters, optimizer,

activation function and the learning rate are also essential components of the model (Bengio, 2012; Bergstra and Bengio, 2012; Smith, 2018). Accordingly, the Adamax optimizer was selected with an initial learning rate of 0.001. In addition, learning rate is reduced by a factor of 0.8 when the learning process becomes stagnant since there is no improvement in validation loss among in each iteration. This progress is accomplished via an early-stopping mechanism automatically to avoid overfitting. The total loss has been selected as the metric for monitoring performance during iterative training. The training process is terminated upon 15 consecutive iterations without improvement. The ReLU activation function is consistently utilized throughout the network to alleviate the vanishing or exploding gradients as discussed in Section 3. As the hyperparameter of proposed loss function, λ regularization parameter, which is discussed in detail in Section 3.2, was determined dynamically at each step of the training phase.

Within this ultimate setup, we employed DL algorithm to simulate and spatially downscale the monthly TWSAs and associated uncertainties (Hereafter these downscaled simulations are called as DWSC TWSA). The general overview of training stage for the chosen configuration and quantified uncertainties of DWSC TWSA are evaluated and shown in Fig. 1. First, the relevant metrics of the training and testing stages that are calculated at each epoch of the training are given in Fig. 1a which also demonstrates the changes in the learning rate that are utilized by the Adamax optimizer as well as the overall mean of the regularization parameters (λ and its inverse, i.e., $1-\lambda$) that is used in loss function. The early-stopping mechanism automatically terminated the training process at the 235'th epoch as the model's architecture had reached its maximum learning capacity. Hence, the training process required 235 epochs of iteration and took approximately 45 min. During this period, learning rate values decreased from 1e-03 to 2.6214e-04, while utilizing the Adamax optimizer to update the weights through gradient evaluation. The square root of the total loss, $\sqrt{\mathscr{L}_{Total}}$, for training and testing exhibit a monotonic decrease, rapidly converging during the initial epochs and subsequently stabilizing until reaching to final epoch. Thus, the DL model optimization and learning is successfully completed. The close proximity and parallel trajectories of the training and testing total loss curves suggest robust generalization capabilities and there is no overfitting phenomenon. In addition, the overall mean of the regularization hyperparameter, λ, of total loss function increased from approximately 0.27 to 0.94 demonstrating a steady increase throughout the training process. This trend confirms the intended functionality of the dynamic weighting mechanism within the total loss function (as in Eq. (10)). Thus, DWSC TWSA achieved higher structural correlation with the HR WGHM TWSA by increasing λ, greater weight is placed on the reconstruction loss ($\mathscr{L}_{RC}$), while the weight on the constraint violation loss ($\mathscr{L}_{CV}$) diminishes. It is determined that the

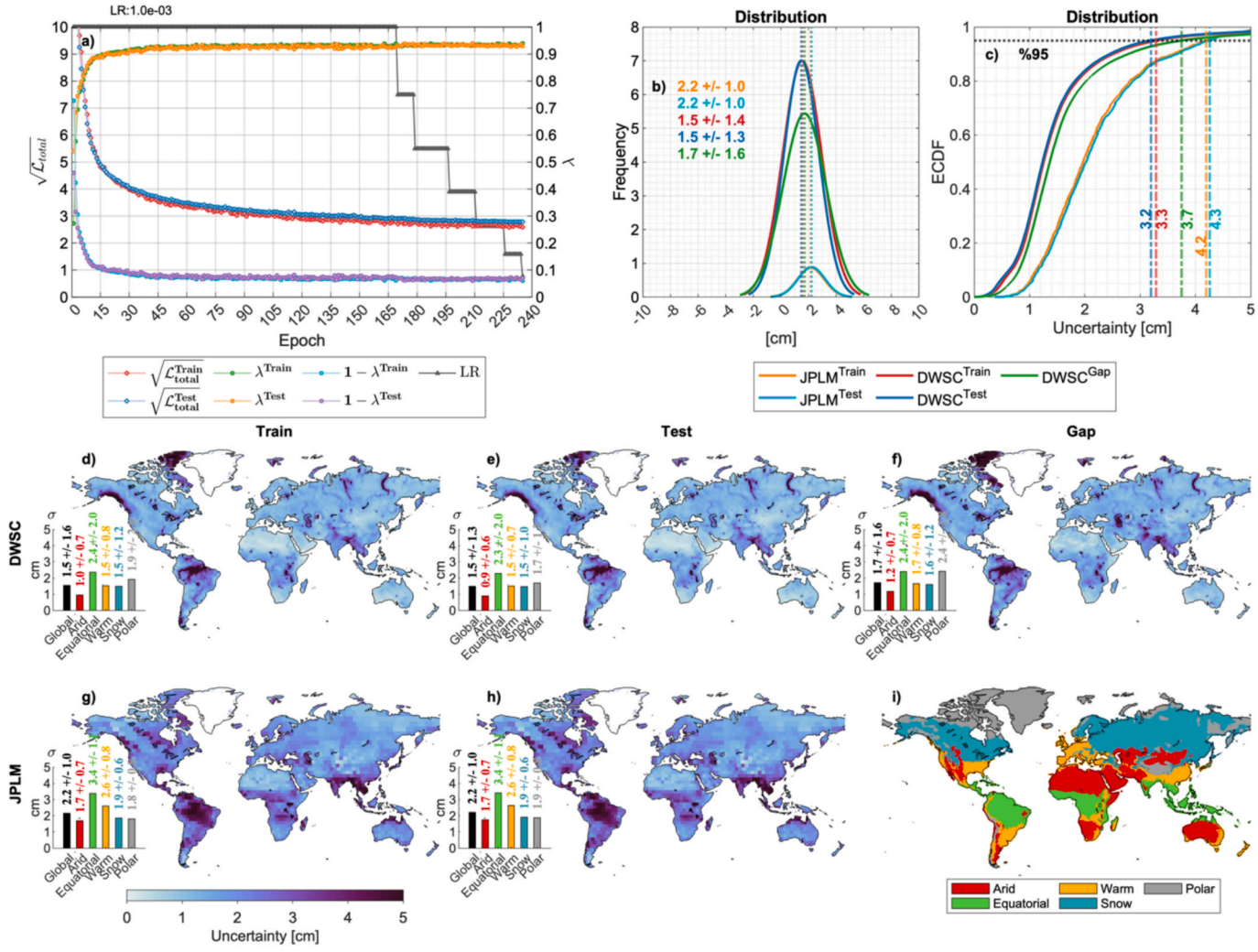


Fig. 1. Variational U-Net training performance and uncertainty analysis comparing the downscaled product (DWSC) and input JPLM data. (a) Training loss convergence and dynamic weight (λ) evolution. (b and c) Distributional comparison (histogram, ECDF) of predictive uncertainty across training, testing, and gap periods. (d–h) Spatial maps comparing median uncertainty [cm] between DWSC (d-f) and JPLM (g–h), including standard deviation stratified by climate zones shown in (i).

correlations to calculate λ are in the same direction and that positive correlations are calculated since they are set up between the same hydrologic variable, i.e., TWSA, from downscaled (DWSC TWSA) and input (WGHM TWSA). Fig. 1a confirms that there are no negative values from first to final iterations. Furthermore, even if the initial iteration steps begin with negative values, the next stages typically have the same directional correlations, which are also proportionate to how well our deep learning algorithm extracts features.

Further on, the spatial distributions of the overall uncertainties for both JPLM and DWSC TWSAs throughout the full GRACE/-FO time series are shown in Fig. 1d–h. These figures display the median values, categorized by training, test, and data-gap months (this only exist for DWSC TWSA). The histogram of these median values in Fig. 1b show that the means of DWSC TWSA uncertainties are lower and standard deviations are higher than JPLM. In addition, the frequencies of DWSC uncertainties are significantly higher than JPLM. DWSC uncertainties have more realistic spatial variations in each grid, since JPLM TWSA has same uncertainty values in each mascon's that are defined as 300 km spatial resolution. This result is supported by the empirical cumulative distribution function (ECDF) of these maps as shown in Fig. 1c. The 95th percentile values confirm this while DWSC uncertainties are around 3.2–3.3 cm, while JPLM uncertainties are around 4.2–4.3 cm for training and testing time periods, separately. The uncertainty during the gap period is slightly higher (~3.7 cm at 95%) than during Train/Test for DWSC, which is expected and reasonable as the model relies solely on the input data without direct GRACE/-FO constraints during these times. However, ECDF of DWSC uncertainties of each time periods are converged to JPLM ECDF's in higher percentile than 95%.

From Fig. 1d to f, the areas with significant uncertainty in the DWSC TWSAs are characterized by river basins or glacier regions where the TWSA signal is strong and a similar spatial pattern is also observed in the JPLM TWSA (as in Fig. 1g and h). The uncertainties are evaluated over the different regions that have different hydro-climatic conditions using World map of the Köppen-Geiger climate classification for major climate zones (excluding Greenland and Antarctica). For this purpose, we utilized a global climate classification dataset, which is named as World Maps of the Köppen-Geiger Climate Classification 1976–2000 and is downloaded from World Bank Data Catalog (https://datacatalog.worldbank.org/search/dataset/0042325). Köppen-Geiger classes are aggregated into five broad climate zones that are Arid, Equatorial, Warm Temperate, Snow, and Polar regions. The spatial distribution of these aggregated zones is illustrated in Fig. 1i. In order to evaluate spatial distribution of uncertainties, the mean and standard deviations that are calculated considering each region boundaries separately are given as inset bar charts in Fig. 1d–h. These charts show consistently lower for mean and higher for standard deviations for the DWSC uncertainties when compared to the JPLM uncertainties within corresponding climate zones and time periods (training and testing). This suggests that the downscaling process not only reduces the overall point-wise uncertainty but also yields a more spatially homogeneous uncertainty field relative to the original low-resolution data, i.e., JPLM TWSA. As a result, these finding suggest that DWSC TWSAs exhibit comparable attributes of modeling uncertainty via variational U-Net to JPLM TWSAs and they are simulated including the targeted informations from both JPLM and WGHM TWSAs with a higher spatial resolution, i.e., 50 km, from April 2002 to December 2022. The downscaled TWSA time series can be downloaded from (Uz et al., 2025).

4.1. Internal validation of downscaled TWSA

4.1.1. Evaluating model's learning through correlation metrics

In our loss function, since we used Pearson correlation (PCC) as hyperparameters to dynamically regularize the relation between DWSC products and JPLM / WGHM TWSAs, we assume that DWSC has a linear relationship with both JPLM and WGHM TWSA. The aim of this assumption via soft-constrained paradigm of proposed loss function is that not only DWSC TWSA has spatial variability as WGHM TWSA but also the TWSA signal of original JPLM is conserved in DWSC. Therefore, the spatial PCC maps are calculated comparing DWSC TWSA to both WGHM and JPLM TWSAs at the beginning (epoch 01) and end (final epoch) of training. Plotting these maps and ECDFs (globally and different climate zones) allows you to visualize how the linear similarity between DWSC TWSA and the input/output data changes spatially and distributionally as the model learns. This demonstrates how well the model incorporates linear patterns from both sources over time. Same analysis is applied by using the new correlation coefficient (hereafter called as XIC) that is proposed by Chatterjee (2020), since this metric has ability to track the development of functional relationships and potentially capture non-linear dependencies. In other words, it ensures assessing how well DWSC TWSA becomes representable as a function of WGHM or JPLM TWSAs during training. This analysis is important because our goal is to downscale GRACE solutions statistically to higher spatial resolution. Therefore, DWSC TWSA time series must be a function of JPLM TWSA, however, it must also include higher spatial variations as in WGHM TWSA.

Fig. 2 evaluates the learning process by tracking the evolution of linear (PCC) and functional (XIC) correlations between the DWSC and JPLM / WGHM TWSAs comparing the initial and final training epochs. PCC's between DWSC and both JPLM (from Fig. 2b1 to d1) and WGHM (from Fig. 2b2 to d2) from initial to final epochs are significantly increased. This shows that the DL algorithm successfully optimizes for linear similarity encouraged by the loss function. The PCC values between WGHM and JPLM in Fig. 2a1 is accepted as reference values for DWSC correlations. It is observed that the DWSC TWSA at the final epoch exhibit a higher correlation with both JPLM and WGHM, when compared to this reference values. This result is also supported by the ECDF plot in Fig. 2f1 as PCC values for both DWSC − JPLM (from 0.70 to 0.93) and DWSC − WGHM (from 0.89 to 0.98) are significantly increased from initial to final epoch w.r.t. the 95′th percentile threshold of JPLM-WGHM (0.76). When comparing the final epoch PCC maps for JPLM-DWSC (Fig. 2d1) and WGHM-DWSC (Fig. 2e1), the linear correlations of JPLM-DWSC tends to be higher than WGHM-DWSC correlations in Arid and Polar climate zones. Conversely, WGHM-DWSC correlation significantly exceeds the JPLM-DWSC correlation within Equatorial, Warm Temperate, and Snow regions. This is also shown in ECDF illustration and their 95′th percentile values for each climate zones as in from Fig. 2g1 to k1. These spatial differences suggests that DWSC TWSA potentially influenced by the characteristics of the hydrological signals within different climate regimes. In Equatorial, Warm Temperate, and Snow regions, which often exhibit strong, well-defined, and spatially variable hydrological signals (e.g., intense precipitation, river dynamics, snow accumulation/melting), the high-resolution WGHM likely provides dominant and informative spatial patterns. The DL algorithm effectively learns these patterns and resulting in a high WGHM-DWSC correlation by reflecting the successful transfer of spatial variability as intended. In contrast, Arid and Polar regions may be characterized by signals that are either less variable at high frequencies and are dominated by large-scale processes (e.g., extensive aquifer systems, peripheral cryosphere changes) or potentially less accurately represented by the WGHM at the target 50 km resolution. In such cases, the coarser but observationally constrained JPLM signal might represent the dominant variability more reliably.

Consequently, DL model is guided by the loss function's objective to conserve the JPLM signal (via $\mathscr{L}_{RC}$) and potentially influenced by a slower development of the WGHM-DWSC correlation (λ) in these zones. Therefore, DWSC product is aligned more strongly in a linear sense with JPLM TWSA signals. This highlights a varying balance achieved by the soft-constrained paradigm prioritizing the strongly informative spatial structure. On the other hand, XIC correlations are given from Fig. 2a2–k2 as in comparison of PCC values. XIC values supported the results of PCC values providing crucial insights into the functional relationships learned by the DL algorithm. From initial to final epoch, the

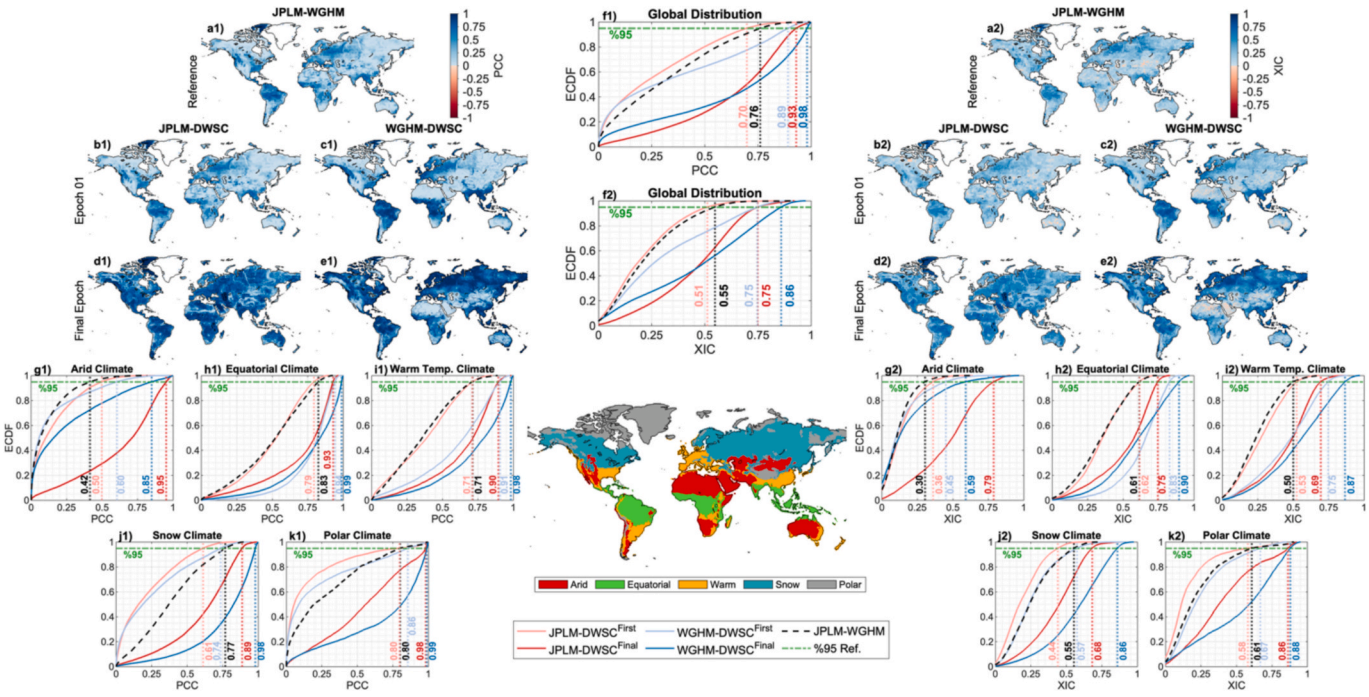


Fig. 2. Spatial maps of Pearson Correlation Coefficient (PCC) and a new correlation coefficient (XIC) that are derived from Chatterjee (2020) between JPLM and WGHM as reference for evolution of linear and functional correlations during model training (a1, a2). Spatial maps of PCC between JPLM-DWSC and WGHM-DWSC at Epoch 1 (b1, c1) and Final Epoch (d1, e1). Corresponding spatial maps for XIC at Epoch 1 (b2, c2) and Final Epoch (d2, e2). Global Empirical Cumulative Distribution Functions (ECDFs) comparing PCC (f1) and XIC (f2) for JPLM-DWSC and WGHM-DWSC pairs at Epoch 1 (dashed lines) and Final Epoch (solid lines), relative to the JPLM-WGHM reference (black dashed line). Regional ECDFs of PCC (g1–k1) and XIC (g2–k2) w.r.t the climate zones that are namely Arid, Equatorial, Warm temperature, Snow and Polar and defined in the central map inset.

XIC values show a significant increase in both JPLM-DWSC (when compared Fig. 2b2 to d2, light to dark red lines in Fig. 2f2) and WGHM-

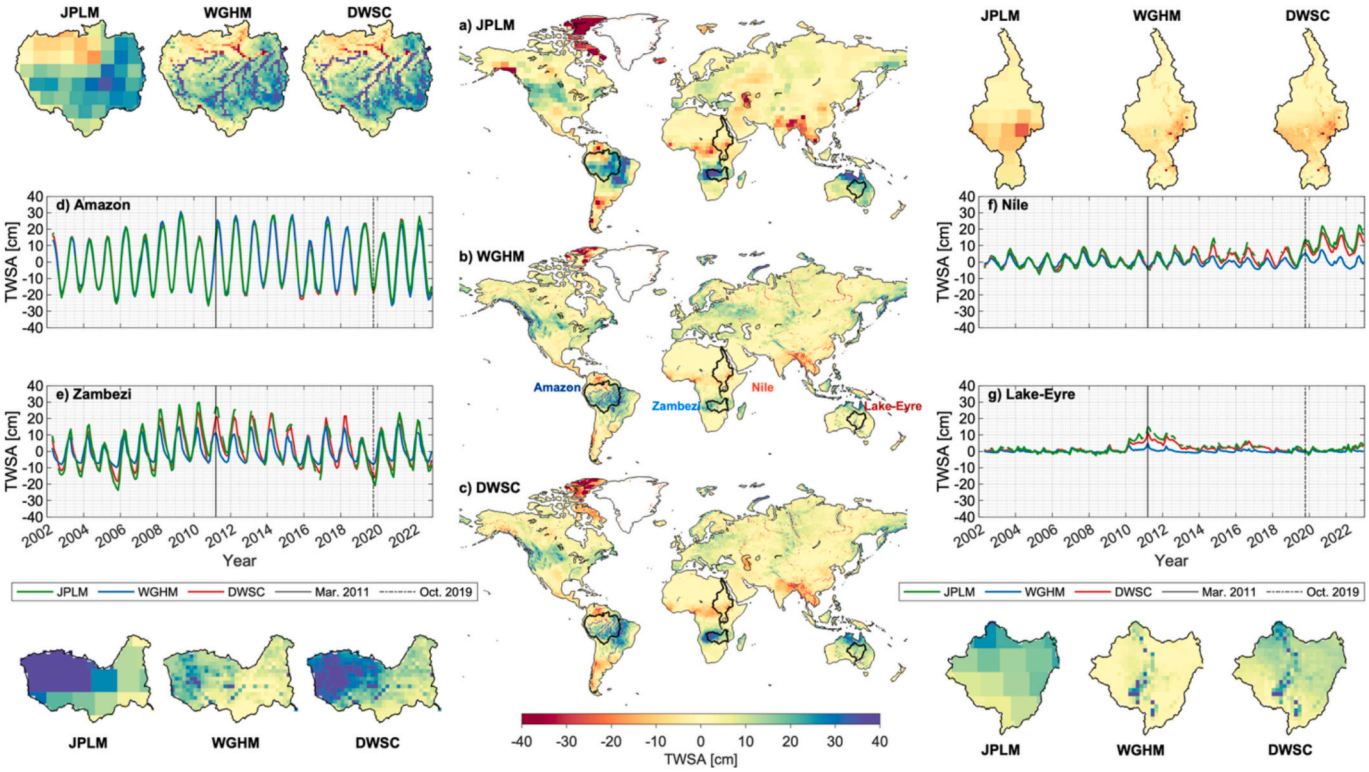


Fig. 3. Global spatial distribution of TWSA [cm] for JPLM (a), WGHM (b) and DWSC (c) in March 2011. Basin-averaged TWSA time series comparing JPLM (green), WGHM (blue) and DWSC (red) for the Amazon (d), Zambezi (e), Nile (f), and Lake Eyre (g) basins from April 2002 to December 2022. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

DWSC (when compared Fig. 2c2 to e2, light to dark blue lines in Fig. 2f2). XIC is substantially greater than zero and the 95′th percentiles of ECDF curves demonstrate that a considerable portion of grid cells exhibit moderate-to-strong (e.g., XIC > 0.5) functional dependence for both JPLM-DWSC and WGHM-DWSC. This confirms that DL algorithm successfully learns to represent the DWSC simulations as a function of the both JPLM and WGHM TWSAs. As shown in Fig. 2d2, DWSC simulations on arid regions is directly dependent to JPLM TWSAs, which is proven by the significantly highest XIC values (with 95′th percentile values of JPLM-DWSC in Fig. 2g2 are obtained up to 0.79) that are obtained from JPLM-DWSC.

4.1.2. Comparison in spatial and temporal domain

Due to the low spatial resolution of GRACE/-FO TWSAs, analyzing the average TWSA signal within basin boundaries is a common practice. The objective of this comparison in our study is to evaluate whether the spatially downscaled DWSC TWSA signals conserve the basin-averaged time series characteristics of the original GRACE/-FO TWSA signal (JPLM). This assesses potential systematic biases to see whether there is any over- or underestimations and verifies the integrity of the TWSA signal. Therefore, the spatiotemporal behaviors of DWSC, JPLM and WGHM TWSA were examined using basin-wise averages for the Amazon (humid), Zambezi (sub-humid), Nile (semi-arid), and Lake Eyre (arid) basins representing diverse climate regions (Oki and Sud, 1998; Lehner and Grill, 2013) as shown in Fig. 3. Consistent with hydro-climatic expectations, the TWSA signal amplitudes of JPLM, WGHM, and DWSC gradually decrease from the humid Amazon to the arid Lake Eyre basin (Fig. 3d to g). Critically, the DWSC TWSA signals closely match the JPLM signals within these basins exhibiting the minimal systematic differences relative to the JPLM reference output that are used in the DL algorithm. In contrast, the WGHM TWSA signal often differs substantially from both JPLM and DWSC particularly in terms of amplitudes and trends; this finding is also noted by Müller Schmied et al. (2023). This divergence is quantitatively supported by Table 1 presenting the estimated trends, annual, and semi-annual amplitudes for each basin. Unlike WGHM, Table 1 clearly shows that the amplitudes, phases and trends of the DWSC TWSA time series demonstrate high coherence to those derived from the JPLM TWSA data.

From Fig. 3a to c TWSA maps for JPLM, DWSC and WGHM TWSAs are given with inset snapshots for each chosen basins for March 2011, whose time epoch is also shown as vertical solid black line in Fig. 3d–g, and the most significant visual difference between JPLM and DWSC is the spatial resolution. While JPLM TWSA consistently exhibit the characteristic smoothness and blockiness associated with its native low resolution (~300 km), DWSC maps display significantly higher spatial detail resolving finer features within the basins similar to WGMH TWSA. This clearly demonstrates the successful spatial resolution enhancement achieved by the downscaling process. The spatial patterns within DWSC TWSAs often show strong coherences to those in the WGHM TWSAs. For example, the features like river networks appear to be effectively transferred from WGHM to DWSC TWSAs. This also indicates that the model successfully learns and incorporates the spatial patterns and variability present in the WGHM input data fulfilling another key objective of the downscaling paradigm. While DWSC TWSA inherits spatial patterns from WGHM, its magnitudes often appear more consistent with JPLM especially during significant events or in areas with strong signals. Because of the heavy precipitation in Australia during the La Nina phase in 2010–2011 (Boening et al., 2012), the snapshots for Lake-Eyre basin are particularly revealing that while JPLM shows a strong positive anomaly covering a large portion of the basin reflecting the major flood event, WGHM shows a much weaker and spatially different signal for the same event. DWSC TWSA map clearly captures the strong positive anomaly as seen in JPLM in terms of both spatial extent and magnitude, while rendering it with the higher spatial detail characteristics of WGHM. This strongly supports the idea that DWSC conserves the crucial event signal from the GRACE observations. In fall season, a different point in the seasonal cycle compared to March 2011 is observed, often corresponding to drier conditions in many Southern Hemisphere regions like Zambezi and Lake Eyre and post-peak flow for rivers like the Nile. Therefore, TWSA samples for October 2019 is also represented in Supplementary Fig. S3 and similar findings are found in this example as well.

Fig. 4 provides a detailed comparison of key temporal components specifically linear trend, annual amplitude, and semi-annual amplitude derived from the JPLM, DWSC, and WGHM TWSA time series. The figure includes spatial maps of these components for each dataset (in Fig. 4a–c, Fig. 4e–g, Fig. 4i–k) illustrated with the zoom view as insets of chosen key regions, i.e., Central Valley, North China Plain, Amazon and Congo basins, and ECDFs of the differences relative to the JPLM reference (JPLM − DWSC and JPLM − WGHM that are illustrated with panels Fig. 4d, h, l). This analysis aims to quantitatively assess the coherence of the DWSC TWSA in representing these fundamental signal characteristics compared to the original GRACE observations (i.e., JPLM TWSA) and DL model input (i.e., WGHM TWSA). Moreover, the evaluation of DWSC TWSA is conducted through inset illustrations of chosen regions emphasizing TWSA signal localization and amplification and through the comparative analysis with JPLM and WGHM TWSA. The Central Valley and the North China Plain, which are identified as critical hotspots of groundwater depletion (Uz et al., 2024a; Ali et al., 2024), provide useful information for assessing the long-term trend components. In contrast, the Amazon and Congo basins are good examples to evaluate seasonal TWSA variability (Kitambo et al., 2023; Tourian et al., 2023).

Visual comparison of the trend maps reveals that the DWSC trend map (Fig. 4b) closely mirrors the large-scale spatial patterns and magnitudes seen in the JPLM trend map (Fig. 4a), capturing significant features like depletion trends in the Middle East, India, and North America, as well as positive trends in other regions on Earth. In contrast, WGHM trend map (Fig. 4c) displays considerably weaker and spatially less coherent trends that are particularly failing to represent the strong negative signals observed by GRACE/GRACE-FO. When DWSC TWSA trends in Fig. 4b insets for Central Valley and North China Plain basins are compared to JPLM TWSA trends in Fig. 4a, it can be concluded that DWSC TWSA are localized almost as stronger as JPLM TWSA at the center of highly groundwater depleted sub-regions for both Central Valley (which is defined as southern part of basin (Uz et al., 2024a)) and for North China Plain (which is located as southwestern part of basin (Ali

Table 1
TWSA time series trend (TR), annual amplitude (AA), annual phase (AP), semi-annual amplitude (SSA) and semi-annual phase (SAP) components of Amazon, Zambezi, Nile and Lake-Eyre basin that are calculated for JPLM, WGHM and DWSC TWSAs.

Basin	Model	Trend [mm/yr]	Annual amplitude [mm]	Annual phase [°]	Semi-annual amplitude [mm]	Semi-annual phase [°]
Amazon	JPLM	2.5	202.3	112.8	12.6	314.4
	WGHM	1.5	203.0	109.8	5.0	274.9
	DWSC	2.2	206.6	113	9.5	321.6
Zambezi	JPLM	1.1	140.1	94.2	32.9	127.3
	WGHM	1.3	85.7	66.5	26.8	99.0
	DWSC	1.6	123.7	90.0	29.2	116.5
Nile	JPLM	7.0	53.3	274.9	8.7	156.4
	WGHM	−0.1	31.9	250.7	4.5	161.1
	DWSC	4.6	47.0	267.7	6.2	167.5
Lake-Eyre	JPLM	0.7	5.4	141.7	5.1	122.9
	WGHM	0.0	5.6	78.5	2.4	92.8
	DWSC	0.6	5.1	132.6	3.8	137.7

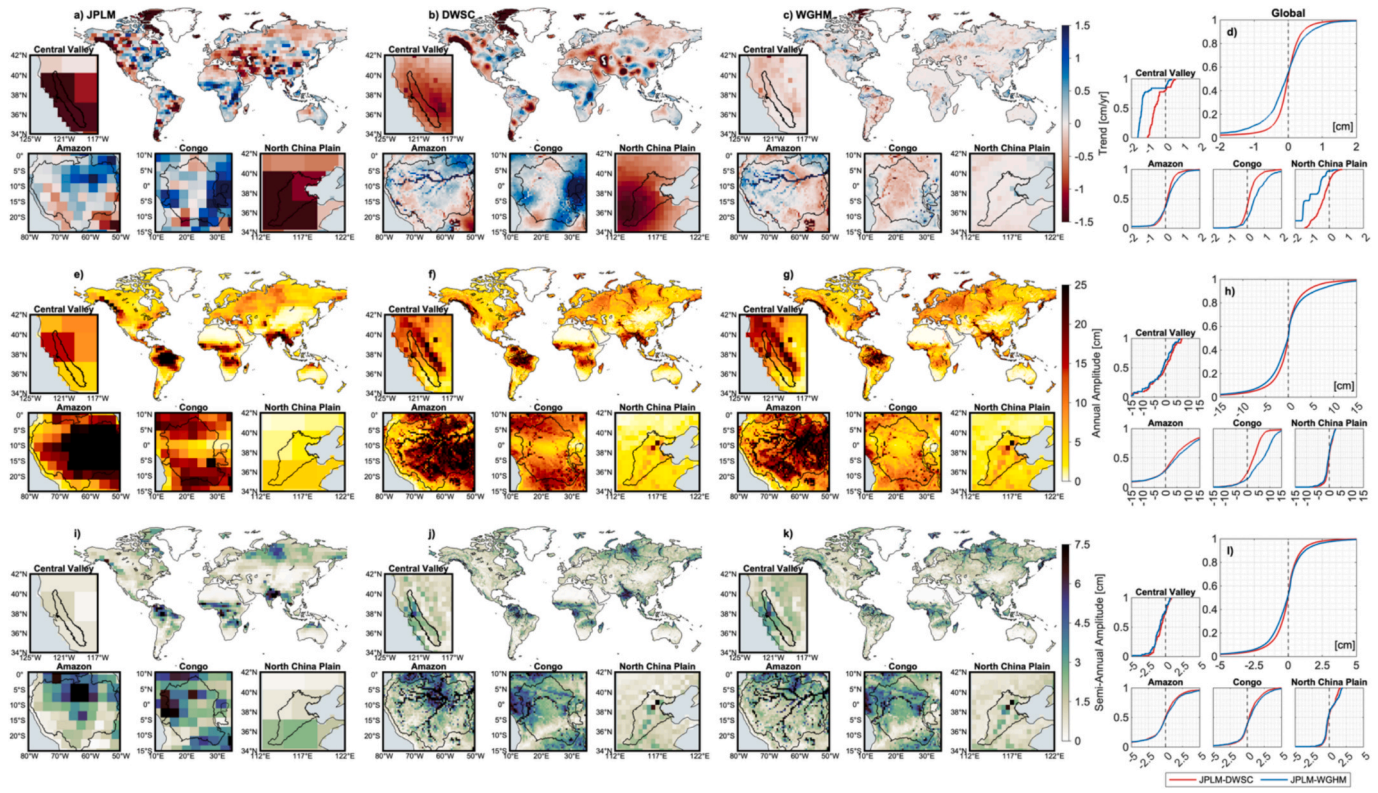


Fig. 4. Comparison of derived temporal components from JPLM, DWSC, and WGHM TWSA time series to evaluate signal localization and amplification in the downscaled product. Spatial maps of linear trend (a–c), annual amplitude (e–g) and semi-annual amplitude (i–k) including inset maps for key regions (Central Valley, Amazon, Congo, North China Plain) to provide a detailed view of high-resolution signal reconstruction. Empirical Cumulative Distribution Functions (ECDFs) of the spatial differences between JPLM and DWSC (red line) and between JPLM and WGHM (blue line) for trend (d), annual amplitude (h) and semi-annual amplitude (l) with inset regional ECDFs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2024)). In addition, the similar result is also obtained for increasing or decreasing trends in sub-regions for Amazon and Congo basins. However, the trend of WGHM TWSA differs from both JPLM and DWSC TWSA, since WGHM underestimates TWSA trends when compared to GRACE/-FO TWSAs (Müller Schmied et al., 2023). The ECDF of differences between JPLM and DWSC trends (JPLM–DWSC as red line in Fig. 4d) is sharply centered around zero indicating that the differences are minimal across most grid cells globally and regionally. Conversely, ECDF of differences between JPLM and WGHM trends is much wider and less centered signifying substantial discrepancies between the observed GRACE trends and the WGHM model trends. This clearly demonstrates that the DWSC product successfully conserves the long-term trend signal present in the JPLM data with high similarity and significantly outperforming the WGHM.

The annual and semi-annual amplitude maps show strong signals in similar regions for all three datasets (e.g., Amazon, Central Africa). The DWSC maps (in Fig. 4f and j) visually aligns well with the JPLM map (Fig. 4e and i) in terms of the magnitude and location of major amplitude centers. While DWSC also shares similar spatial patterns with WGHM (when compared to Fig. 4g and k), its overall intensity appears closer to JPLM as in shown in ECDF illustrations in Fig. 4h and l. In addition, the localization and amplification of the seasonal signals of DWSC TWSA for Amazon and Congo basins as given in Fig. 4f and j are not only stronger as JPLM TWSA but also reflect the HR features of river networks. But, even if WGHM TWSA has also HR features, the amplitude of the seasonal signals in Fig. 4g and k are less intense than JPLM and DWSC, especially for Amazon and Congo basins. As a result, the comparative analysis presented in Fig. 4 strongly supports the hypothesis that DWSC TWSA effectively conserves the fundamental temporal characteristics of the original JPLM GRACE/-FO signal. For linear trend, annual amplitude, and semi-annual amplitude, the spatial maps and particularly the ECDF difference plots demonstrate that DWSC achieves significantly better agreement with JPLM TWSA than WGHM. The differences between JPLM and DWSC for these components are consistently small and centered around zero globally indicating minimal systematic bias and higher coherence in representing both long-term changes and dominant seasonal cycles observed by GRACE.

4.1.3. *Comparison through spectral domain*

An analysis of the radially averaged power spectra (RAPS) is presented in Fig. 5 for selected months to provide valuable insights into how the DWSC simulations represent TWSA signals across different spatial scales (from 50 km to 5000 km resolutions) relative to the JPLM observations and WGHM model input. The RAPS of TWSA samples are illustrated for two different months during the GRACE period in March 2005 (Fig. 5a) and the GRACE-FO period in October 2019 (Fig. 5b). In addition, August 2013 (Fig. 5c) and January 2018 (Fig. 5d) are also illustrated, since these months are in the gap time period where GRACE/-FO data was not available. The spectra are compared across three key bands as LR (>600 km averaging radius), intermediate resolution (IR) corresponding roughly to GRACE/GRACE-FO resolution (300–600 km averaging radius) and HR (<300 km averaging radius). According to Fig. 5a and b, a consistent pattern emerges as the power spectra of DWSC (red lines) closely align with those of JPLM (green lines) at LR band for months where JPLM data are available. This indicates that the downscaled simulations successfully conserve the power associated with large-scale hydrological features and processes captured by the GRACE observations. In contrast, the power spectra of WGHM (blue lines) sometimes deviate from JPLM and DWSC at these LR bands.

Within the IR band, the DWSC spectra generally track the JPLM spectra suggesting that the downscaling still preserves the TWSA signal characteristics near the native resolution limit of GRACE/-FO. Even if

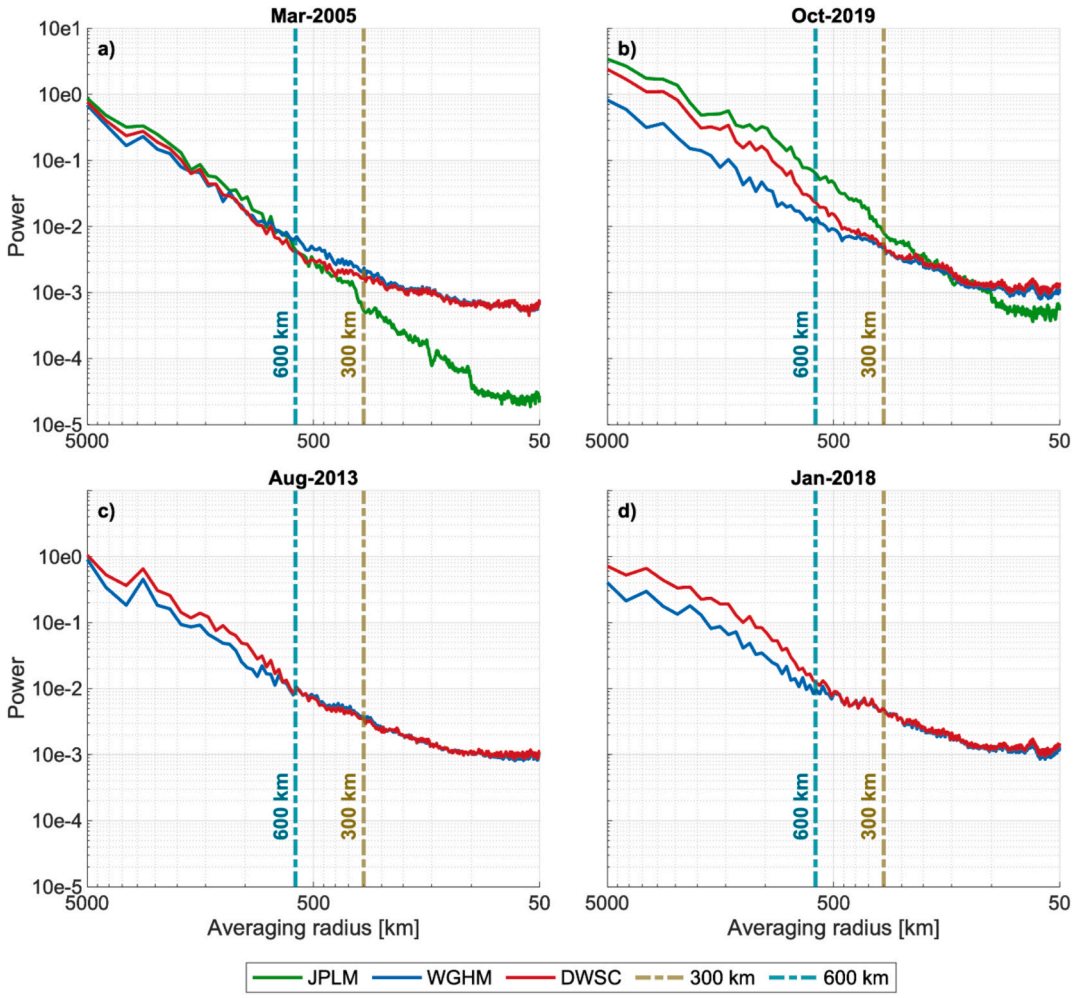


Fig. 5. Radially averaged power spectrum of the average signal of TWSA samples display during the GRACE period in March 2005 (a) and the GRACE(−FO) period in October 2019 (b) within the native or intrinsic full (approximately 600 km) and half-wavelength (approximately 300 km) spatial resolution limits of the GRACE/-FO solutions. Additionally, TWSA samples for August 2013 (c) and January 2018 (d) where GRACE/-FO data was not available.

WGHM power spectra often shows slightly different power levels in this range, DWSC TWSA also initiate to converge to power spectra of WGHM. On the other hand, JPLM spectra exhibit significantly attenuated power at HR band beyond the effective resolution of GRACE/-FO. Conversely, both WGHM and DWSC power spectra display substantially higher power at this band and DWSC spectra closely follow the WGHM spectra. This demonstrates that DWSC effectively incorporates the high-frequency spatial variability and power of HR WGHM TWSA by adding realistic small-scale details that are absent in the original JPLM TWSA. Furthermore, the power spectrum calculated for August 2013 and January 2018 (Fig. 5c and d) show the DWSC spectrum aligning to WGHM spectrum particularly higher resolutions than ~600 km similarly as in Fig. 5a and b. However, although DL model relies solely on the WGHM and ERA5 inputs during gap periods without the observational constraint from JPLM, DWSC power spectra is similar to JPLM as in Fig. 5a and b and has higher power than WGHM at LR bands. This result is based on the fundamental theory of DL approaches which makes predictions w.r.t. to the weights that are determined only by the data provided during training. Thus, data gaps are estimated utilizing these weights with a certain level of accuracy and without the need for any constraints. As a result, the spectral analysis confirms that the downscaling methodology successfully integrates information across different resolution bands. DWSC retains the LR band spectral power characteristic of the JPLM GRACE observations while effectively incorporating the higher spatial frequency power and variability present in the WGHM model TWSA at HR band.

4.2. External validation

4.2.1. Comparison by El Niño Southern Oscillation

As a result of large-scale ocean–atmosphere interactions over the equatorial Pacific, the El Niño Southern Oscillation (ENSO) is the most active interannual time-scale episodic climate phenomenon that affects long-term mass anomaly both over continents and oceans (Bjerknes, 1969; Uz et al., 2024b). Therefore, analyzing how these interactions are inherited in our DWSC products during periods exhibiting extreme water storage anomalies and assessing whether the downscaled TWSA signal demonstrates these variations constitutes a valuable external validation to test the ability of our downscaling paradigm. For this purpose, the relationship between interannual TWSA variations (after removing trends and seasonal cycles) and the ENSO are evaluated using the Niño 3.4 Sea Surface Temperature Anomaly (SSTA, https://www.cpc.ncep.noaa.gov/data/indices/) index and the results are given Fig. 6. This index is defined in the region between 120°W–70°W longitudes and 5°N–5°S latitudes which is shown with magenta area in Fig. 6c. SSTA anomalies represent the deviation of monthly SSTs from their average over a long period of time. Within this concept, while El Niño event is considered to occur when the Nino3.4 index exceeds +0.5°C for a minimum of five consecutive months, La Niña event is also considered when the Niño3.4 index falls below −0.5°C (Uz et al., 2024b). First, the

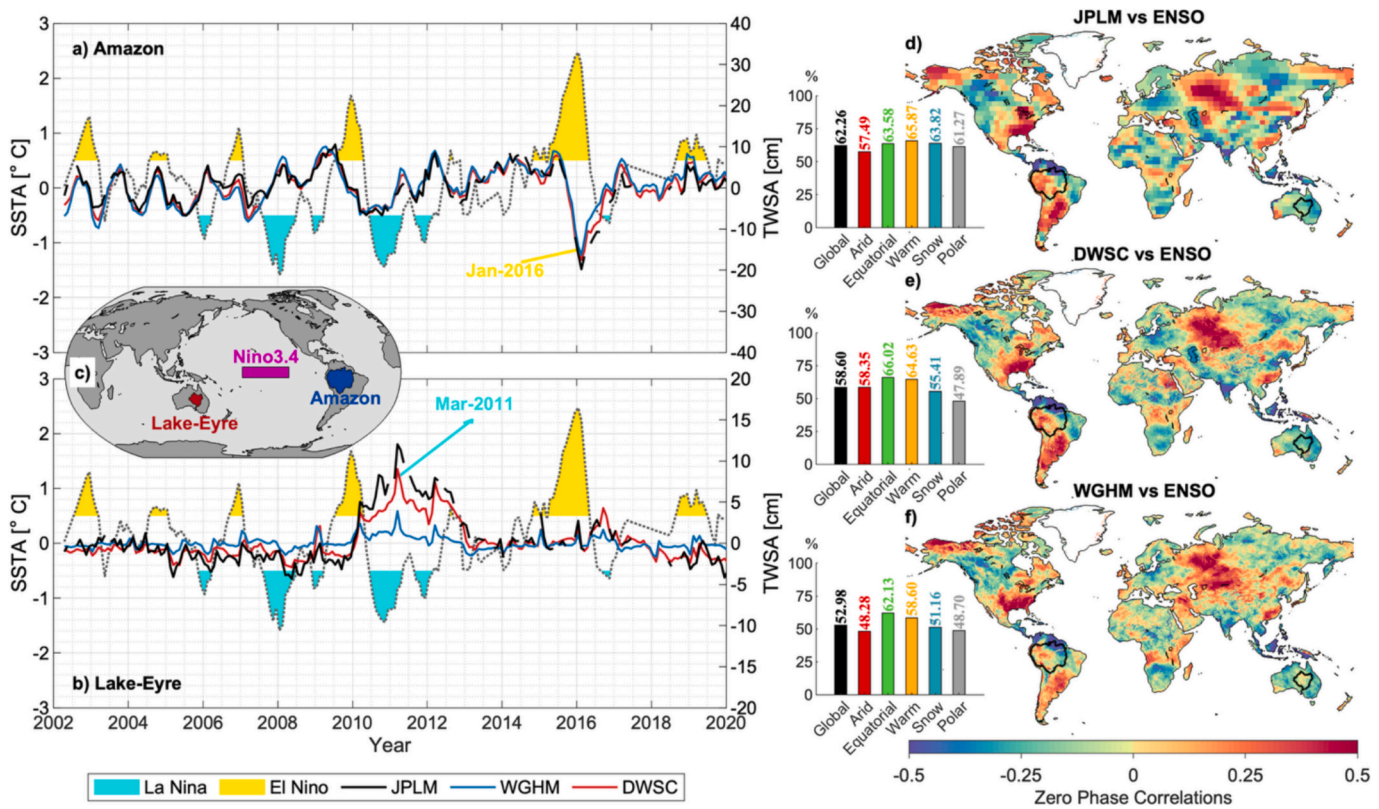


Fig. 6. Basin-averaged detrended-deseasoned time series for JPLM (black), WGHM (blue), and DWSC (red) compared with the Niño 3.4 SSTA index for the (a) Amazon and Lake Eyre (b) basins and Map showing basin locations and the Niño 3.4 region (c). Spatial maps of zero-lag correlation between nonseasonal TWSA and the Niño 3.4 SSTA index for JPLM (d), DWSC (e) and WGHM (f). Inset bar charts showing the percentage of land area with statistically significant ($p < 0.05$) zero-lag correlations within global and specific climate zone categories for each dataset. (Arrows indicate specific ENSO events (Jan 2016 El Niño and Mar 2011 La Niña). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

analysis focuses on comparing basin-averaged and detrended – deseasoned (nonseasonal) time series TWSA from JPLM, DWSC, and WGHM for the Amazon and Lake Eyre basins to ENSO index as given in Fig. 6a and b, since the time series of nonseasonal TWSA variations for each model specifically connected to the ENSO interaction in the Amazon and Lake-Eyre regions. In addition, hydrologic conditions of these regions, while Amazon is humid region, Lake Eyre is arid, are opposite.

There were two significant ENSO activity as the Amazon drought induced by the El Niño period in 2015–2016 (Xie and Fang, 2020) or the heavy precipitation in Australia during the La Niña phase in 2010–2011 (Boening et al., 2012). These events are also good example to externally validate whether nonseasonal TWSA variations can be captured by model time series as drought event in a humid region or flood event in an arid region. In Fig. 6a, Amazon region generally experienced drier conditions during strong El Niño events, however, the time series highlights the strong 2015–2016 El Niño as positive SSTA peak. In January 2016, the Amazon region noted one of the lowest values of TWSA associated with drought. During this period, all three JPLM (black line), WGHM (blue line) and DWSC (red line) exhibit a pronounced negative nonseasonal TWSA anomaly consistent with drought conditions often associated with El Niño in this region. Throughout the time period, nonseasonal TWSA variations of DWSC closely track both JPLM and WGHM capturing the interannual fluctuations linked to ENSO phases. On the other hand, Lake Eyre arid basin was experienced to the significant flooding during strong La Niña events in 2010–2011 (negative SSTA anomalies) as shown in Fig. 6b. Correspondingly, both JPLM and DWSC show a very large positive nonseasonal TWSA anomaly capturing the documented major flooding event around March 2011. WGHM signal, however, shows a much weaker positive anomaly during this period failing to fully represent the magnitude of the La Niña impact observed by GRACE/-FO. Thus, DWSC time series demonstrates high coherence and capability to retain the amplitude of significant interannual variations observed by GRACE outperforming WGHM in representing ENSO-driven extremes.

The spatial analysis of zero-phase correlations in Fig. 6d–f between nonseasonal TWSA and the ENSO index is also supporting these findings. The map derived from JPLM data (Fig. 6d) reveals established large-scale ENSO teleconnection patterns including negative correlations across the tropics (e.g., Amazon, Indonesia) and positive correlations in regions like southern Africa and parts of Australia. The corresponding map for DWSC (Fig. 6e) reproduces these large-scale patterns with higher similarity preserving the location, sign, and general strength of the correlations observed in JPLM, while rendering them at a higher spatial resolution. However, WGHM correlation map (Fig. 6f) shows slight differences in the spatial extent and intensity of these teleconnections compared to JPLM and DWSC. Quantitative assessment using the percentage of land area with statistically significant correlations, which are given in bar charts inset in Fig. 6d-6f and quantify the percentage of both globally and different climate zones exhibiting statistically significant ($p < 0.05$, based on critical $r \approx 0.134$) zero-lag correlations with the ENSO index, supports the visual map comparison. The percentage of significant correlations for DWSC closely matches with JPLM and, thus, both observational (JPLM) and DWSC datasets consistently show a higher percentage of significant ENSO correlations, especially in Arid, Warm temperature and Equatorial regions, compared to the WGHM model. This demonstrates that DWSC not only captures the spatial structure of ENSO's influence but also preserves the statistical significance of these climate-hydrology links to a degree comparable to the original GRACE observations.

4.2.2. Comparison by mountain glaciers

The WGHM TWSA does not include glacier mass changes (Müller Schmied et al., 2023). Therefore, since our DWSC TWSA time series include contributions from both the WGHM and JPLM TWSAs, it is important to test how realistically the mountain glacier mass changes could be inherited from the JPLM TWSAs. In other words, testing whether the DWSC TWSA samples include glacier mass changes is one of the quality indicators of our GRACE-like simulations or how realistically the JPLM time series has been downscaled. For this purpose, time series of GWSA, SWSA, SMSA and SWE are removed from DWSC and each model TWSAs to calculate only the glacier mass changes time series. To this end, we used the GWSA, SWSA, SMSA and SWE datasets that are released by Global Gravity-based Groundwater Product (G3P) project (Zemp et al., 2019; Güntner et al., 2024). In addition, the glacier mass change time series (hereafter named as G3PG), which are also released by G3P Project and are independent from GRACE/-FO observations (Zemp et al., 2019; Güntner et al., 2024), are used to compare each model-derived glacier mass change time series. In order to make a fair comparison, each time series are calculated only from January 2003 to December 2022 excluding the time period covering the data gap between GRACE and GRACE-FO. In Fig. 7e, the trend map of the G3PG time series at glacier regions (excluding Greenland and Antarctica) of the Earth is shown, defining the dominant glacier mass loss regions. These glacier regions are also defined in the Randolph Glacier Inventory (RGI – (RGI 7.0 Consortium, 2023)) database (shown in Supplementary Fig. S4). Here we chose four different sub-regions from these dominant glacier mass loss regions to calculate glaciers mass change time series. Therefore, the 3° × 3° degree grids were defined for the glacier regions of Alaska (ALA), Arctic Canada North (ACN), Southern Andes (SAN) and South East Asia (ASE) (see Fig. 7e) and the glacier mass change time series from both DWSC and JPLM TWSAs for these selected sub-regions are calculated (after converting to mass unit) and shown in Fig. 7a - 7d along with their long-term trends. The trend of the DWSC time series for all regions closely aligns with the trends observed in both the JPLM and G3PG time series. It is anticipated that the trend values of WGHM are underestimated because the WGHM TWSA time series data lacks information on glaciers. Consequently, the regions for WGHM time series exhibit an unrealistic trend value.

The time series plots (Fig. 7a–d) consistently demonstrate that the DWSC (red lines) capture the pronounced negative trends indicating glacier mass loss closely aligning with both the JPLM (black lines) and the independent G3PG (green lines) time series in all four regions. The calculated linear trends for DWSC (−13.2 Gt/yr for ALA, −5.7 Gt/yr for ACN, −6.8 Gt/yr for SAN, −2.2 Gt/yr for ASE) are quantitatively similar

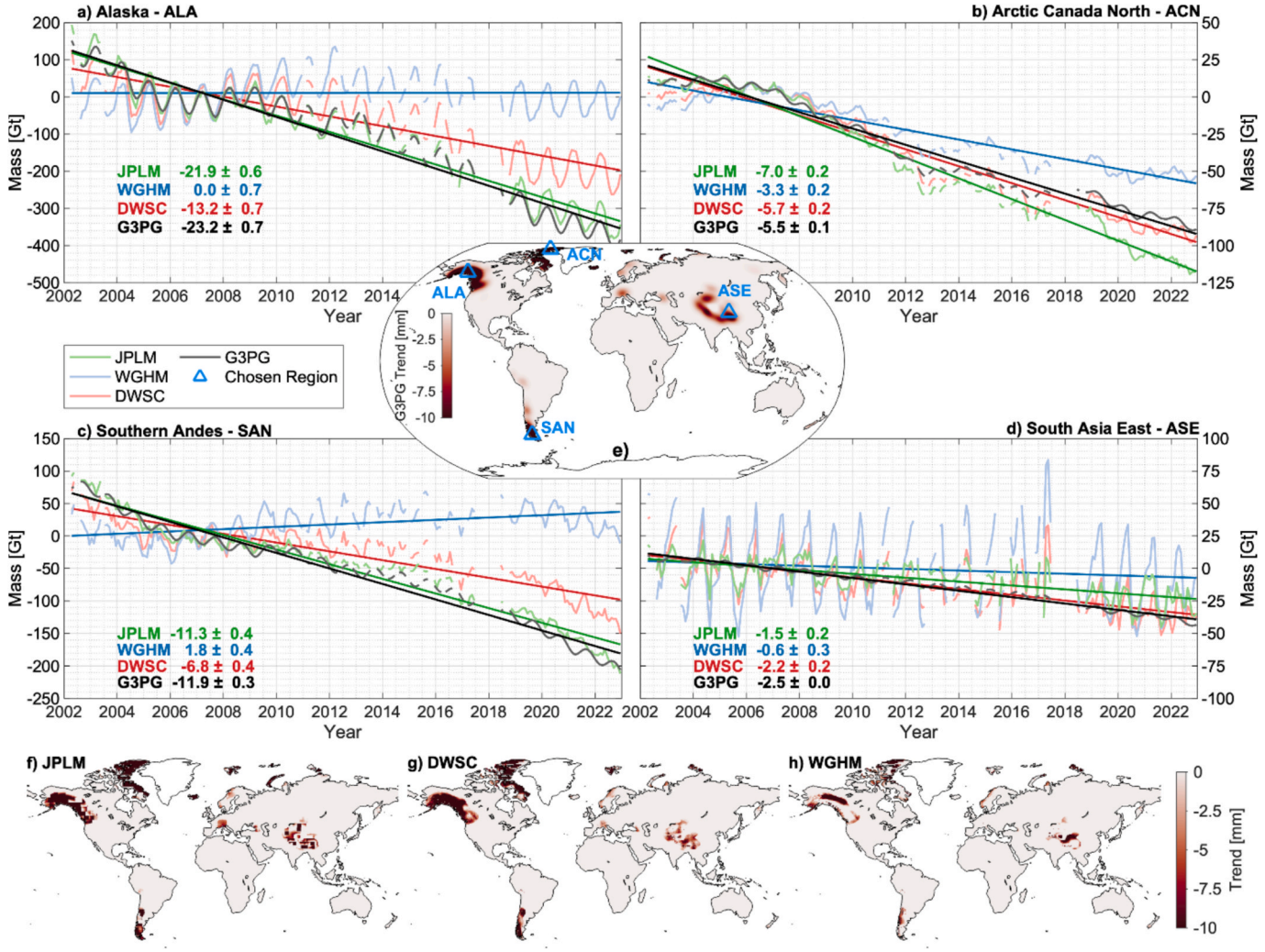


Fig. 7. Time series of estimated glacier mass change [Gt] for JPLM (black), WGHM (blue), DWSC (red) and independent G3PG data (green) averaged over 3°x3° regions in (a) Alaska (ALA), (b) Arctic Canada North (ACN), (c) Southern Andes (SAN), and (d) South East Asia (ASE). Calculated linear trends [Gt/yr] ± standard deviation is indicated for each dataset. Map (e) showing the locations of the four chosen regions and the spatial distribution of G3PG trends. Spatial maps of derived long-term trends for JPLM (f), DWSC (g) and WGHM (h). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and crucially show reasonable agreement with those from JPLM (−21.9, −7.0, −11.3, −1.5 Gt/yr respectively) and the independent G3PG trends (−23.2, −5.5, −11.9, −2.5 Gt/yr respectively). In contrast, WGHM time series (blue lines) exhibit significantly weaker or even slightly positive trends (0.0, −3.3, 1.8, −0.6 Gt/yr respectively) confirming the expected lack of glacier signal representation in this dataset. The close agreement between DWSC and JPLM / G3PG trends coupled with the clear divergence from WGHM strongly suggests that the DWSC product successfully inherits the glacier mass change signal primarily from the JPLM observations during the downscaling process. This conclusion is further supported by the spatial trend maps in Fig. 7f–h. The JPLM trend map (Fig. 7f) clearly shows strong negative trends concentrated in major glacier regions globally and DWSC trend map (Fig. 7g) effectively reproduces these spatial patterns of glacier mass loss with high coherence to JPLM, however, it is potentially rendered with finer spatial detail influenced by the downscaling architecture. Conversely, the WGHM trend map (Fig. 7h) lacks these distinct negative trends over glacier areas displaying much weaker and spatially inconsistent signals. In summary, the comparison against both JPLM and the independent G3PG dataset provides compelling evidence that the DWSC product realistically incorporates glacier mass change signals. Despite WGHM lacking this information, the soft-constrained downscaling paradigm allows DWSC to effectively inherit and represent the long-term glacier mass loss trends present in the JPLM GRACE observations demonstrating its suitability for applications in cryospheric regions alongside hydrological areas.

4.2.3. *Comparison by SMAP satellite-derived surface soil moisture data product*

The Soil Moisture Active Passive (SMAP) satellite mission was launched by the National Aeronautics and Space Administration (NASA) in January 2015 (Entekhabi et al., 2010) in order to monitor surface soil moisture (SSM) and root zone soil moisture (RZSM) using a L-band active radar and passive radiometer. Here, we aim to further investigate the reliability of our downscaled TWSA by comparing the Level-4 (L4) data products that are obtained from integrating SMAP observations into the GEOS-5 catchment land surface model (Blank et al., 2023). With a temporal resolution of 3 h and a spatial resolution of 9 km, the assimilation provides estimates for the SSM of 5 cm and the RZSM of up to 1 m below surface. The SMAP L4 dataset, as described by Reichle et al. (2022), is available for download from the National Snow and Ice Data Center (NSIDC) at https://nsidc.org/data/spl4smgp/versions/7. For this validation, we only utilized SSM L4 products starting from April 2015 to December 2022. We resampled the data to monthly and 50 km resolutions by averaging from the 3 hourly and 9 km dataset. Besides, SSM anomalies (SSMA) are calculated by removing the average value of the time period that is used. While performing this validation, it is crucial to emphasize two fundamental considerations. Firstly, it is important to note that the SMAP L4 data products should not be regarded as purely obtained from satellite soil moisture observations. Instead, they should

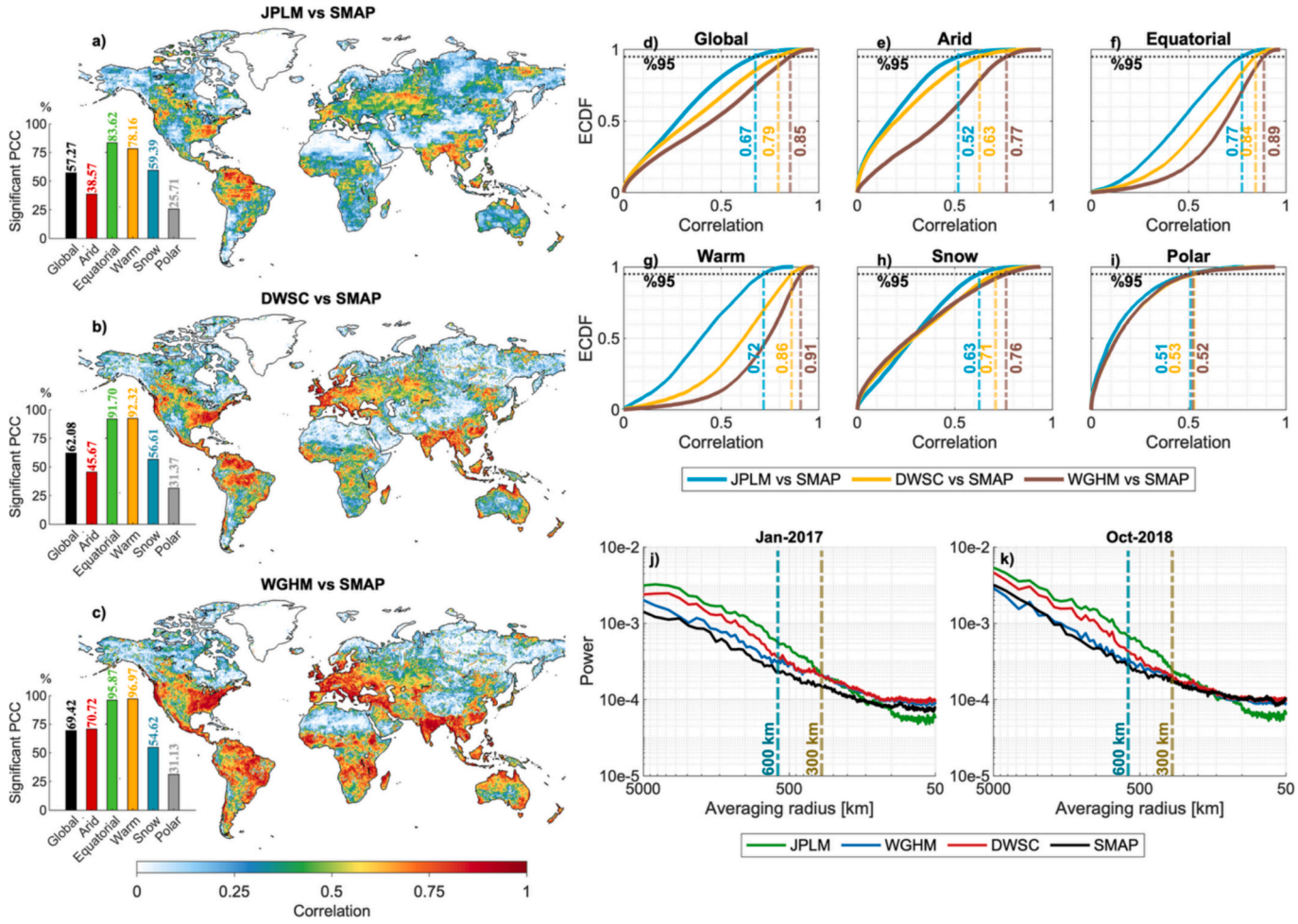


Fig. 8. Spatial maps of Pearson Correlation Coefficient (PCC) between TWSA (JPLM, DWSC, WGHM) and SMAP SSMA from April 2015 to Dec 2022 (a–c). Inset bar charts show the percentage of statistically significant correlations globally and by climate zone. Empirical Cumulative Distribution Functions (ECDFs) of the PCC values for globally (d) and by climate zone (e–i). Radially Averaged Power Spectra (RAPS) comparing JPLM, WGHM, DWSC, and SMAP for January 2017 (j) and October 2018 (k).

be seen as a simulated dataset that is subject to significant influence from data assimilation, namely from the climate data used as input for the land surface model (Blank et al., 2023). Moreover, while SSMA is restricted to the top few centimeters of the soil, TWSA considers water storage variations in all compartments, including both the soil and depths. Nevertheless, GRACE-like observations of TWSA are unable to differentiate between storage variations across different compartments or soil depths. Thus, SSMA time series, although inherently possessing a higher spatial resolution, can only represent a part of the TWSA. Conversely, our DL paradigm enables DWSC TWSAs to accurately localize the sum of all compartments of the JPLM TWSA, while also acquiring higher spatial resolution information from WGHM TWSAs. Therefore, validating the correspondence of DWSC TWSAs with SSMA time series shows how much information is gained in our simulations, separately from JPLM and WGHM TWSAs.

The spatial correlation maps in Fig. 8a–c illustrate the PCC between each TWSA dataset and SMAP SSMA. A clear progression in spatial detail and correlation strength is visible from Fig. 8a to c. JPLM map shows broad correlation patterns consistent with its low resolution. WGHM exhibits the strongest and most detailed correlation patterns with surface soil moisture particularly highlighting river networks and humid regions, since it is a high-resolution hydrological model focused on land surface processes. However, DWSC map presents an intermediate case displaying significantly more spatial detail and generally higher correlations than JPLM, while incorporating many of the high-resolution correlation features seen in WGHM, such as enhanced correlations along major river systems like Amazon or Lena. This visual evidence suggests that DWSC successfully harmonize the large-scale coherence with enhanced spatial detail. In addition, ECDFs in Fig. 8d–8i quantify the correlation distributions both globally and across different climate zones. In Fig. 8d, ECDF curves confirm the visual progression as WGHM shows the highest correlations with SMAP followed by DWSC and then JPLM. Thus, the 95th percentile correlation values increase from JPLM (~0.67) to DWSC (~0.79) and further to WGHM (~0.86) indicating that DWSC achieves substantially higher correlation with SSMA than the original GRACE data. Besides, the regional ECDFs (in Fig. 8e–i) show variations consistent with expected hydrological behavior and the correlations are generally highest in Equatorial regions (Fig. 8f) and lowest in Arid (Fig. 8e) and Polar (Fig. 8i) regions for all datasets. Especially, DWSC curve lies between the JPLM and WGHM curves in most zones reinforcing its intermediate character but showing a significant improvement over JPLM.

The percentage of statistically significant correlations increases progressively from JPLM (lowest) to DWSC and then WGHM (highest) both globally and within most climate zones. This indicates that the enhanced spatial resolution and incorporation of hydrological model structure in DWSC lead to a larger extent of statistically meaningful correlation with SSMA compared to the original LR GRACE data. The RAPS plots are given in Fig. 8j and k and compared the spatial power content of both TWSAs and SSMA for two specific months as Jan. 2017 and Oct. 2018, respectively. At LR bands (> 600 km averaging radius), DWSC (red line) and JPLM (green line) spectra show higher power compared to WGHM (blue line) and SMAP (black line) consistent with TWSA representing the full water column while SMAP represents only the surface. In the IR bands (300–600 km), DWSC continues to track JPLM, however, the power in JPLM drops off significantly due to its resolution limitations at HR bands (<300 km radius). In contrast, the DWSC spectra closely follow both WGHM and the independent and high-resolution SMAP spectra in this high-frequency band. These findings collectively support the conclusion that the DWSC successfully enhances spatial resolution and incorporates realistic hydrological patterns while preserving the integrity of the LR TWSA signal.

4.2.4. Comparison with previous studies

DWSC TWSAs are also externally assessed utilizing the independent GRACE/-FO like TWSA data products available from other studies which include similar downscaled or assimilated TWSA time series. Thus, we compared our solutions with (i) the GLWS 2.0 time series (hereafter called as GLWS), which was obtained by assimilating GRACE/-FO data into the WGHM model using the ensemble Kalman filter method (Gerdener et al., 2023a, b), and (ii) the GRACE-SeDA v1.0 (hereafter called as SeDA) time series, which was also calculated using both WGHM and JPLM TWSAs within a self-supervised deep learning approach (Gou and Soja, 2023; Gou and Soja, 2024). Fig. 9 provides a comparative assessment of these different TWSA products using the NSE metric to evaluate the predictive skill of each TWSA time series relative to both the JPLM and the WGHM TWSAs. According to the foundation of the NSE evaluation that if the NSE values of the simulated model are negative, then the mean observation is better than these simulations (Krause et al., 2005). Thus, an NSE value greater than 0 indicates that the simulated or downscaled products are better than simply using the mean of the compared reference dataset signifying meaningful predictive skill (called as significant NSE). The analysis reveals the different performance characteristics for each product. DWSC demonstrates the excellent predictive skill when compared to JPLM GRACE data (Fig. 9a) that is evidenced by predominantly high positive NSE values globally and a high percentage (~76%) of significant skill (NSE > 0), indicating strong coherence to the observational reference. Its performance against WGHM (Fig. 9b), while still showing widespread significant skill (~66%) exhibits greater spatial variability and generally lower NSE magnitudes reflecting the intended balance of the soft-constrained approach which integrates WGHM structure while prioritizing GRACE signal conservation. The SeDA product also displays significant skill relative to both JPLM (Fig. 9c, ~66% significant NSE) and WGHM (Fig. 9d, ~58% significant NSE) potentially indicating a slightly weaker adherence to the WGHM model structure compared to DWSC based on the relative NSE scores. In contrast, the GLWS assimilation product shows markedly lower skill in matching JPLM and WGHM temporal variations (Fig. 9e, ~30% and Fig. 9f, ~40% significant NSE) compared to both DL methods and exhibits very low NSE scores. Overall, the NSE comparison suggests that the DL approaches, especially DWSC, achieve a more effective balance in integrating observational constraints from JPLM with HR structural information from WGHM than GLWS data assimilation product based on this performance metric. This could be due to the fact that the GLWS employs the ITSG-Grace2018 time series (Mayer-Gürr et al., 2018; Kvas et al., 2019) as the GRACE/-FO solution which involves the different gravity inversion process when compared to JPLM. In addition, NSE indicates whether there are signals that are different and meaningful from the reference data used as the basis for hydrological analysis. Thus, GLWS TWSA are most similar to WGHM TWSA with lowest significant NSEs and it could be concluded that it is the result of assimilation efforts.

4.2.5. Comparison with groundwater well time series

DWSC TWSA time series are also validated by using a direct comparison with independent observations from monitoring wells. To provide this external validation, we compared our GWSA products against an extensive set of in-situ groundwater level anomaly (GWLA) time series across the CONUS with a specific focus on two major and heavily stressed aquifer systems: the Central Valley of California and the High Plains aquifer (Rateb et al., 2020; Scanlon et al., 2021). We utilized a comprehensive well observation dataset that is an annual median GWLA time series from (Jasechko, 2023; Jasechko et al., 2024). We selected all wells in the CONUS that provided at least 10 years of data between 2002 and 2022, since the dense well network exists to ensure more accurate comparison. First, GWSA time series are calculated by removing the sum of canopy water, snow water storage and root zone soil moisture from the JPLM, DWSC, and WGHM TWSAs using outputs from the GLDAS NOAH model (Rodell et al., 2004; Beaudoing et al., 2020). To ensure consistency of the comparison between point-wise GWLA and gridded GWSA time series, we first calculated the annual median values of GWSA time series from 2002 to 2022 and then one-dimensional annual GWSA

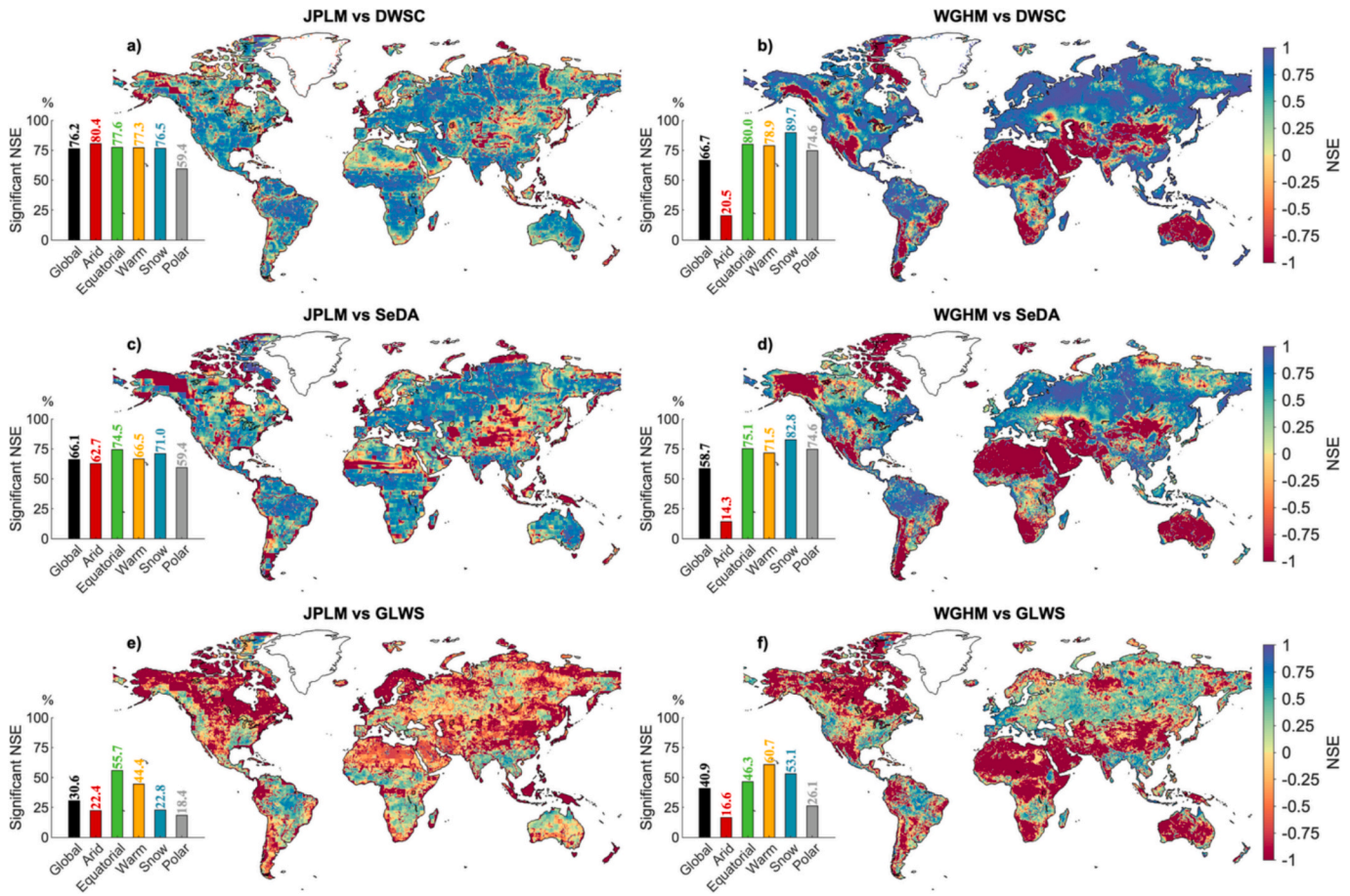


Fig. 9. Spatial maps show NSE calculated between reference datasets (JPLM or WGHM) and evaluated products (DWSC, SeDA, GLWS): JPLM vs DWSC (a), WGHM vs DWSC (b), JPLM vs SeDA (c), WGHM vs SeDA (d), JPLM vs GLWS (e) and WGHM vs GLWS (f). Inset bar charts quantify the percentage of significant NSE (>0) globally and different climate zone for each respective comparison.

time series for each well location is calculated by selecting the nearest grid point.

Long-term trend is a critical indicator of groundwater stress that affects the groundwater recharge, storage, and discharge (Taylor and Alley, 2001). We calculated trends for the in-situ GWLA and for each of the three GWSA time series at every well location using the robust Theil-Sen slope estimator. The spatial distribution of these trends is shown in Fig. 10a–d. The trends of the wells in Fig. 10a reveal intense and highly localized depletion trends particularly concentrated in the southern part of the Central Valley (the San Joaquin and Tulare Basins) and the Central and Southern High Plains. These regions are well-established as global hotspots of groundwater depletion due to intensive agricultural irrigation being amplified during severe droughts (Scanlon et al., 2021; Liu et al., 2022). As shown in Fig. 10c, the trend of our downscaled product has also the similar pattern with well observation when compared to Fig. 10a. In addition, the detailed views of the Central Valley (in Fig. 10e3) and High Plains (in Fig. 10g3) show a spatial pattern of trends that closely matches the in-situ observations. In contrast, the trends of low-resolution JPLM GWSA in Fig. 10b are correctly identifying the general regions of depletion or recharge which are the expected results since GRACE/-FO observations are unique to monitor total amount of water that is stored in different compartments. However, JPLM is suffered from the signal leakage and less accurate spatial localization due to coarse resolution as shown in Fig. 10e2 and g2. For instance, when Fig. 10e2 is compared to both Fig. 10e1 and e3, in contrast to both DWSC GWSA and well observations, the trends of JPLM GWSA in the northern part of Central Valley (the Sacramento Basin) is calculated as much as the magnitude of groundwater depletion in both San Joaquin and Tulare Basins and this result does not match the Central Valley hydrologic characteristics. On the other hand, JPLM GWSA trends have the same magnitude with an overly smooth representation for the local depletion or recharge regions (especially between Northern and Central-Southern High Plains) due to the coarse resolution as shown in Fig. 10g2. In addition, since the WGHM GWSA (in Fig. 10d, e4, g4) suffer from underestimated trends and the lacking a complete representation of human intervention (Müller Schmied et al., 2023), which fails to capture the observed depletion. Moreover, the trends are calculated as recharging or neutral directions even in the severe depletion areas.

To evaluate how well each GWSA time series captures the interannual variability, we calculated the Pearson correlations between the GWLA time series and each corresponding GWSA time series at each well location. The spatial patterns of these correlations are shown for the Central Valley (Fig. 10f1–f3) and High Plains (Fig. 10h1–h3). In addition, the distribution of correlations is given as the violin plots in Fig. 10i–k to provide a quantitative summary of temporal performance. Across the entire CONUS (Fig. 10i), Central Valley (Fig. 10j) and High Plains (Fig. 10k), the distribution of correlation coefficients for the DWSC and JPLM GWSAs are statistically superior than WGHM GWSA. The median values of correlations in these regions for both JPLM (0.39, 0.48 and 0.60) and DWSC (0.33, 0.42 and 0.53) are almost the same values and higher, when compared to WGHM median values (0.00, 0.12, 0.00). Besides, the distributions of JPLM and DWSC correlations are significantly and tightly more clustered at positive values.

We anticipate that our deep learning paradigm is enable the simulation of a GRACE/-FO-like TWSA/GWSA signal by incorporating more

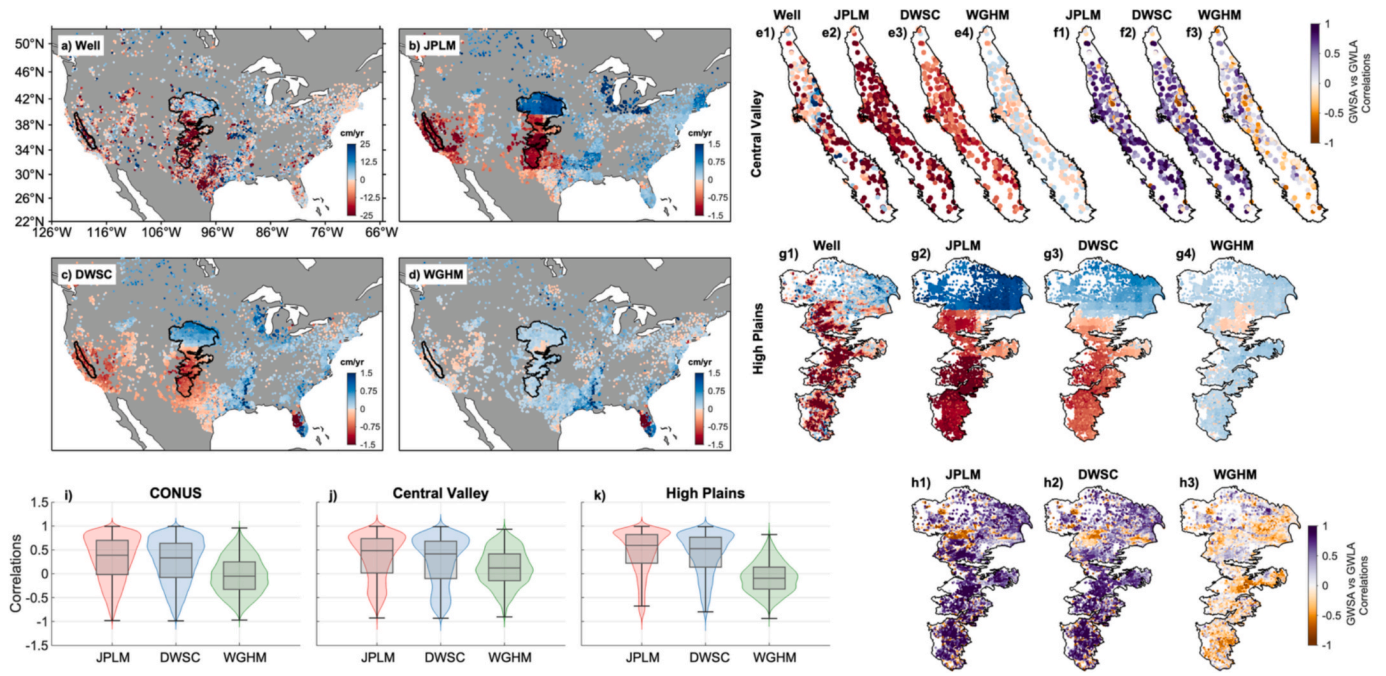


Fig. 10. External validation of downscaled simulation against in-situ groundwater well observations. Spatial maps of long-term trends (cm/yr) over the CONUS for in-situ well data (GWLA) and the three GWSA time series of JPLM, DWSC and WGHM (a–d). Detailed comparisons of trend maps for each time series on the Central Valley (e1–e4) and High Plains aquifers (g1–g4). Spatial distribution of Pearson correlations between each GWSA time series and the corresponding in-situ well time series for Central Valley (f1–f3) and High Plains aquifers (h1–h3). Violin plots of correlation coefficients for all wells across the CONUS (i), Central Valley (j) and High Plains (k).

localized features via downscaling framework. This comparison illustrates that DWSC GWSA trends are not only more localized than those of JPLM but also exhibit spatial patterns most similar to GWLA trends; furthermore, the interannual correlations of DWSC GWSA are similar to that of JPLM. This direct validation against groundwater level measurements demonstrates the downscaling capability of our deep learning paradigm by combining the LR observational consistency of GRACE with the enhanced spatial resolution of a land hydrology model. Consequently, DWSC TWSA is assessed as a more accurate and spatially detailed representation of actual groundwater dynamics.

5. Conclusions and future work

This study introduced a novel Deep Learning framework that is based on a Variational U-Net architecture combined with a novel dynamic soft-constrained loss function for spatially downscaling of GRACE/GRACE-FO TWSA. The primary challenge addressed was enhancing the spatial resolution of observation-based TWSA from the native ~300 km of GRACE/GRACE-FO to a finer 50 km, while simultaneously conserving the crucial LR signals inherent in the satellite observations and incorporating realistic HR spatial variability derived from the WGHM hydrological model TWSA. The innovative loss function dynamically balances coherence to the low-resolution JPLM GRACE mascon product and the HR WGHM structure using a data-driven weighting factor based on evolving Pearson correlations. The resulting downscaled TWSA products demonstrate significant improvements and robust performance across multiple validation experiments. Internal validation confirmed that DWSC successfully preserves the basin-averaged temporal dynamics including long-term trends and seasonal (annual and semi-annual) amplitudes observed in JPLM significantly outperforming WGHM in this regard. Correlation analyses (PCC and XIC) revealed that DWSC effectively learns both linear and functional relationships from the source datasets developing a particularly strong functional dependence on the spatial patterns inherent in WGHM, thus achieving enhanced spatial variability. Spectral analysis further supports these findings showing that DWSC TWSA matches the spectral power of JPLM TWSA at low resolution (>600 km) while aligning with the power spectra of HR WGHM TWSA at high resolutions (<300 km).

External validation against independent datasets underscored the reliability and added value of the DWSC product. Comparisons with ENSO indices demonstrated DWSC's ability to capture climate-driven interannual variability and extreme hydrological events (droughts, floods) with high fidelity to the GRACE record. Despite the absence of glacier components in the WGHM input, DWSC successfully inherited and represented glacier mass loss trends consistent with both JPLM and independent glacier datasets (G3PG) highlighting its applicability in cryospheric regions. Furthermore, DWSC exhibited significantly improved correlation and spectral consistency with HR SMAP surface soil moisture compared to the original JPLM data. Comparative analysis using NSE showed that DWSC achieves a strong balance in predictive skill relative to both JPLM and WGHM generally surpassing the performance of alternative assimilated or DL-aided studies results. The most notable finding is that a direct comparison with in-situ groundwater well observations in major aquifers of CONUS demonstrated that the downscaled TWSA-derived GWSA time series effectively represents the spatial patterns of long-term groundwater depletion. This achievement addresses the issues of signal leakage and the inaccuracies in spatial localization that are characteristic of the coarse-resolution GRACE/-FO data. This validation further confirms that our downscaled TWSAs can be utilized in monitoring groundwater stress in critical aquifer systems. Additionally, the probabilistic nature of the Variational U-Net allowed for the quantification of predictive uncertainty yielding similar uncertainty estimates for DWSC compared to the input JPLM data.

Although these strong results, it is crucial to note the limitations of our proposed paradigm, which also guide potential directions for future research. First, it must be noted that proposed DL paradigm has a potential underestimation of the trend components of simulated or downscaled TWSA when compared to trends of GRACE/-FO time series. This expected result is sourced from the nature of proposed soft-constrained paradigm, since it is only based on constraining through a

loss function which enforces for a balance between the TWSA from GRACE/-FO and that from WGHM in this study. Specifically, in order to overcome this limitation, the hard-constrained DL paradigm can be applied by designing a constrain layer in the framework algorithm that can be based on water budget equation or strict mass conservation rule. However, unlike in downscaling other climate variables such as precipitation, surface temperature, soil moisture etc., high resolution TWSA versus GRACE/-FO TWSA is not available for training such a hard-constrained deep learning model. This makes it more challenging, however we believe it may be feasible in the future when more GRACE-like TWSA data is available which could help to overcome redundancy problem of the hard-constrained deep neural network models for downscaling satellite gravimetry solutions. On other approach could be priorly removing the trend signal from both the input and the output data and training the model using de-trended input–output data sets, and finally after training process restoring the trend signal back in the predictions of the deep learning model. While applying such an approach, the final trend signal would be consistent with those of the LR GRACE/-FO TWSA, the resulting downscaled TWSA would still be questionable because the assumption of the same trend signal in all the downscaled grids within the original GRACE/-FO TWSA grid may not be realistic and requires further investigation. The above ideas are left for separate research works in the future. In addition, the accuracy of our downscaled TWSA mainly depends on the accuracy of the input data, since inaccuracies in the HR spatial patterns of the input dataset inevitably propagate to the downscaled simulations. While the model training is computationally intensive, the inference is implemented rapidly. Even if the simulations are suitable for analysis, the significant computational resources could be required for retraining with additional input data. The external validation revealed the strong performance for capturing significant hydrological events; however, the ability of the model to generalize to extreme events during the study period requires additional investigation. Consequently, further improvements may concentrate on integrating a wider range of HR inputs to minimize dependence on one particular hydrological model, investigating physics-based neural network architectures that are capable of directly learning water cycle related physical laws and developing techniques to more effectively quantify and propagate uncertainty from the input data to the final downscaled product.

In conclusion, the proposed soft-constrained deep learning paradigm provides an effective and robust method for enhancing and downscaling the spatial resolution of GRACE/GRACE-FO TWSA. The resulting DWSC dataset successfully integrates the observational power of satellite gravimetry including signals like groundwater depletion and glacier melt with the fine spatial detail of hydrological models offering a valuable resource for regional hydrological studies, water resource management, and climate change impact assessments requiring high-resolution, observationally-constrained TWSA information with associated uncertainty estimates.

CRediT authorship contribution statement

Metehan Uz: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Kazım Gökhan Atman:** Writing – review & editing, Software, Methodology, Formal analysis. **Orhan Akyılmaz:** Writing – review & editing, Methodology. **C.K. Shum:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

JPL RL06.1 v03 Mascon solutions are available at http://grace.jpl.nasa.gov. ERA5 datasets are available on European Centre for Medium-Range Weather Forecast website (ECMWF - https://cds.climate.copernicus.eu). WaterGAP Global Hydrology Model version 2.2e is available at https://gude.uni-frankfurt.de/. The HydroBasin data are available at https://www.hydrosheds.org/products/hydrobasins. GLWS 2.0 dataset is available at PANGAEA - Data Publisher for Earth & Environmental Science database https://doi.pangaea.de/10.1594/PANGAEA.954742. GRACE-SeDA dataset is available at https://www.research-collection.ethz.ch/handle/20.500.11850/648738. Global Gravity-based Groundwater Product (G3P) v1.12 are available at GFZ Data Services, https://doi.org/10.5880/g3p.2024.001. Randolph Glacier Inventory database is available at NSIDC: National Snow and Ice Data Center, https://doi.org/10.5067/F6JMOVY5NAVZ. The monthly Niño3.4 index time series is available National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) database https://www.cpc.ncep.noaa.gov/data/indices/sstoi.indices. Soil Moisture Active Passive L4 (SMAP L4) surface soil moisture dataset is available at National Snow and Ice Data Center (NSIDC) https://nsidc.org/data/spl4smgp/versions/7. GLDAS Noah Land Surface Model v2.1 is available at https://disc.gsfc.nasa.gov/datasets/. World Maps of the Köppen-Geiger Climate Classification 1976-2000 dataset available at World Bank Data Catalog (https://datacatalog.worldbank.org/search/dataset/0042325).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2026.135015.

Data availability

I have shared the data link in my manuscript.

References

Ali, S., Ran, J., Luan, Y., Khorrami, B., Xiao, Y., Tangdamrongsub, N., 2024. The GWR model-based regional downscaling of GRACE/GRACE-FO derived groundwater storage to investigate local-scale variations in the North China Plain. Sci. Total Environ. 908, 168239. https://doi.org/10.1016/j.scitotenv.2023.168239.

Beaudoing, H., Rodell, M., NASA/GSFC/HSL, 2020. GLDAS Noah Land Surface Model L4 monthly 0.25 x 0.25 degree, Version 2.1. https://doi.org/10.5067/SXAVCZFAQLNO.

Bengio, Y., 2012. Practical Recommendations for Gradient-Based Training of Deep Architectures. In: Montavon, G., Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade Lect. Notes Comput. Sci., vol 7700. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_26.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. Retrieved from J. Mach. Learn. Res. 13 (10), 281–305. http://jmlr.org/papers/v13/bergstra12a.html.

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., Gentine, P., 2021. Enforcing analytic constraints in neural networks emulating physical systems. Phys. Rev. Lett. 126, 098302. https://doi.org/10.1103/PhysRevLett.126.098302.

Bjerknes, J., 1969. Atmospheric teleconnections from the equatorial pacific. Mon. Weather Rev. 97 (3), 163–172. https://doi.org/10.1175/1520-0493(1969)097<0163:ATFTEP>2.3.CO;2.

Blank, D., Eicker, A., Jensen, L., Güntner, A., 2023. A global analysis of water storage variations from remotely sensed soil moisture and daily satellite gravimetry. Hydrol. Earth Syst. Sci. 27 (13), 2413–2435. https://doi.org/10.5194/hess-27-2413-2023.

Boening, C., Willis, J.K., Landerer, F.W., Nerem, R.S., Fasullo, J., 2012. The 2011 La Niña: so strong, the oceans fell. Geophys. Res. Lett. 39, 2012GL053055. https://doi.org/10.1029/2012GL053055.

Chatterjee, S., 2021. A new coefficient of correlation. J. Am. Stat. Assoc. 116, 2009–2022. https://doi.org/10.1080/01621459.2020.1758115.

Chen, J., Cazenave, A., Dahle, C., Llovel, W., Panet, I., Pfeffer, J., Moreira, L., 2022. Applications and challenges of GRACE and GRACE follow-on satellite gravimetry. Surv. Geophys. 43, 305–345. https://doi.org/10.1007/s10712-021-09685-x.

Eicker, A., Schumacher, M., Kusche, J., Döll, P., Schmied, H.M., 2014. Calibration/data assimilation approach for integrating GRACE data into the WaterGAP global hydrology model (WGHM) using an ensemble Kalman filter: first results. Surv. Geophys. 35, 1285–1309. https://doi.org/10.1007/s10712-014-9309-8.

Entekhabi, D., Njoku, E.G., O'Neill, P.E., Kellogg, K.H., Crow, W.T., Edelstein, W.N., Entin, J.K., Goodman, S.D., Jackson, T.J., Johnson, J., Kimball, J., Piepmeier, J.R., Koster, R.D., Martin, N., McDonald, K.C., Moghaddam, M., Moran, S., Reichle, R., Shi, J.C., Spencer, M.W., Thurman, S.W., Tsang, L., Van Zyl, J., 2010. The soil moisture active passive (SMAP) mission. Proc. IEEE 98 (5), 704–716. https://doi.org/10.1109/JPROC.2010.2043918.

Flechtner, F., Neumayer, K.-H., Dahle, C., Dobslaw, H., Fagiolini, E., Raimondo, J.-C., Güntner, A., 2016. What can be expected from the GRACE-FO laser ranging interferometer for earth science applications? Surv. Geophys. 37, 453–470. https://doi.org/10.1007/s10712-015-9338-y.

Gal, Y., Ghahramani, Z., 2016: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, Vol. 48, Association for Computing Machinery, 1050–1059, https://dl.acm.org/doi/10.5555/3045390.3045502.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., Shahzad, M., Yang, W., Bamler, R., Zhu, X.X., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56, 1513–1589. https://doi.org/10.1007/s10462-023-10562-9.

Gerdener, H., Kusche, J., Schulze, K., Döll, P., Klos, A., 2023a. The global land water storage data set release 2 (GLWS2.0) derived via assimilating GRACE and GRACE-FO data into a global hydrological model. J. Geod. 97, 73. https://doi.org/10.1007/s00190-023-01763-9.

Gerdener, H., Schulze, K., Kusche, J., 2023b: GLWS 2.0: A global product that provides total water storage anomalies, groundwater, soil moisture and surface water with a spatial resolution of 0.5° from 2003 to 2019. PANGAEA, Accessed: [01-06-2024], https://doi.org/10.1594/PANGAEA.954742.

Girotto, M., Rodell, M., 2019. Terrestrial water storage, in: Extreme Hydroclimatic Events and Multivariate Hazards in a Changing Environment. Elsevier, pp. 41–64. https://doi.org/10.1016/B978-0-12-814899-0.00002-X.

Gou, J., Soja, B., 2023. GRACE-SeDA: A global total water storage anomaly product with a spatial resolution of 0.5 degrees from self-supervised data assimilation. ETH Research Collection; Accessed: [01-06-2024], https://doi.org/10.3929/ethz-b-000648738.

Gou, J., Soja, B., 2024. Global high-resolution total water storage anomalies from self-supervised data assimilation using deep learning algorithms. Nat. Water 2, 139–150. https://doi.org/10.1038/s44221-024-00194-w.

Güntner, Andreas; Sharifi, Ehsan; Haas, Julian; Boergens, Eva; Dahle, Christoph; Dobslaw, Henryk; Dorigo, Wouter; Dussaillant, Inés; Flechtner, Frank; Jäggi, Adrian; Kosmale, Miriam; Luojus, Kari; Mayer-Gürr, Torsten; Meyer, Ulrich; Preimesberger, Wolfgang; Ruz Vargas, Claudia; Zemp, Michael, 2024. Global Gravity-based Groundwater Product (G3P). V. 1.12. GFZ Data Services, Accessed: [01-07-2024], https://doi.org/10.5880/G3P.2024.001.

Harder, P., Hernandez-Garcia, A., Ramesh, V., Yang, Q., Sattegeri, P., Szwarcman, D., Watson, C., Rolnick, D., 2023. Hard-constrained deep learning for climate downscaling. J. Mach. Learn. Res. 24 (365), 1–40. http://jmlr.org/papers/v24/23-0158.html.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N., 2023: ERA5 monthly averaged data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), Accessed: [01-02-2024], https://doi.org/10.24381/cds.f17050d7.

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv, 1207.0580v1, https://doi.org/10.48550/arXiv.1207.0580.

Houborg, R., Rodell, M., Li, B., Reichle, R., Zaitchik, B.F., 2012. Drought indicators based on model-assimilated gravity recovery and climate experiment (GRACE) terrestrial water storage observations. Water Resour. Res. 48, 2011WR011291. https://doi.org/10.1029/2011WR011291.

Humphrey, V., Rodell, M., Eicker, A., 2023. Using satellite-based terrestrial water storage data: a review. Surv. Geophys. 44, 1489–1517. https://doi.org/10.1007/s10712-022-09754-9.

Irrgang, C., Saynisch-Wagner, J., Dill, R., Boergens, E., Thomas, M., 2020. Self-validating deep learning for recovering terrestrial water storage from gravity and altimetry measurements. Geophys. Res. Lett. 47, e2020GL089258. https://doi.org/10.1029/2020GL089258.

Jasechko, S., 2023. Groundwater level data, aquifer system boundaries, and supplementary tables associated with Jasechko, S. et al. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature, doi.org/10.1038/s41586-023-06879-8 (2024). https://doi.org/10.5281/ZENODO.10003697.

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R.G., Fallatah, O., Kirchner, J.W., 2024. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature 625, 715–721. https://doi.org/10.1038/s41586-023-06879-8.

Kendall, A., Gal, Y., 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Proceedings of the 31st International Conference on Neural Information Processing Systems, 5580–5590. Presented at the Long Beach, California, USA. Red Hook, NY, USA: Curran Associates Inc.

Kingma, D.P., Welling, M., 2013. Auto-Encoding Variational Bayes. https://doi.org/10.48550/ARXIV.1312.6114.

Kitambo, B.M., Papa, F., Paris, A., Tshimanga, R.M., Frappart, F., Calmant, S., Elmi, O., Fleischmann, A.S., Becker, M., Tourian, M.J., Jucá Oliveira, R.A., Wongchuig, S., 2023. A long-term monthly surface water storage dataset for the Congo basin from 1992 to 2015. Earth Syst. Sci. Data 15, 2957–2982. https://doi.org/10.5194/essd-15-2957-2023.

Kornfeld, R.P., Arnold, B.W., Gross, M.A., Dahya, N.T., Klipstein, W.M., Gath, P.F., Bettadpur, S., 2019. GRACE-FO: the gravity recovery and climate experiment follow-on mission. J. Spacecr. Rocket. 56, 931–951. https://doi.org/10.2514/1.A34326.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. https://doi.org/10.5194/adgeo-5-89-2005.

Kvas, A., Behzadpour, S., Ellmer, M., Klinger, B., Strasser, S., Zehentner, N., Mayer-Gürr, T., 2019. ITSG-Grace2018: overview and evaluation of a new GRACE-only gravity field time series. J. Geophys. Res. Solid Earth 124 (8), 9332–9344. https://doi.org/10.1029/2019JB017415.

Landerer, F.W., Flechtner, F.M., Save, H., Webb, F.H., Bandikova, T., Bertiger, W.I., Bettadpur, S.V., Byun, S.H., Dahle, C., Dobslaw, H., Fahnestock, E., Harvey, N., Kang, Z., Kruizinga, G.L.H., Loomis, B.D., McCullough, C., Murböck, M., Nagel, P., Paik, M., Pie, N., Poole, S., Strekalov, D., Tamisiea, M.E., Wang, F., Watkins, M.M., Wen, H., Wiese, D.N., Yuan, D., 2020. Extending the global mass change data record: GRACE follow-on instrument and science data performance. Geophys. Res. Lett. 47, e2020GL088306. https://doi.org/10.1029/2020GL088306.

Lehner, B., Grill, G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrol. Process. 27 (15), 2171–2186. https://doi.org/10.1002/hyp.9740.

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

Liu, P.-W., Famiglietti, J.S., Purdy, A.J., Adams, K.H., McEvoy, A.L., Reager, J.T., Bindlish, R., Wiese, D.N., David, C.H., Rodell, M., 2022. Groundwater depletion in California's Central Valley accelerates during megadrought. Nat. Commun. 13, 7825. https://doi.org/10.1038/s41467-022-35582-x.

Mayer-Gürr, T., Behzadpur, S., Ellmer, M., Kvas, A., Klinger, B., Strasser, S., Zehentner, N., 2018. ITSG-Grace2018 - Monthly, Daily and Static Gravity Field Solutions from GRACE. GFZ Data Services. http://doi.org/10.5880/ICGEM.2018.003.

Müller Schmied, H., Trautmann, T., Ackermann, S., Cáceres, D., Flörke, M., Gerdener, H., Kynast, E., Peiris, T.A., Schiebener, L., Schumacher, M., Döll, P., 2023. The global water resources and use model WaterGAP v2.2e: description and evaluation of modifications and new features. https://doi.org/10.5194/gmd-2023-213.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10 (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

Oki, T., Sud, Y.C., 1998. Design of total runoff integrating pathways (TRIP)—a global river channel network. Earth Interact 2 (1), 1–37. https://doi.org/10.1175/1087-3562(1998)002<0001:DOTRIP>2.3.CO;2.

Pascal, C., Ferrant, S., Selles, A., Maréchal, J.-C., Paswan, A., Merlin, O., 2022. Evaluating downscaling methods of GRACE (gravity recovery and climate experiment) data: a case study over a fractured crystalline aquifer in southern India. Hydrol. Earth Syst. Sci. 26, 4169–4186. https://doi.org/10.5194/hess-26-4169-2022.

Rateb, A., Scanlon, B.R., Pool, D.R., Sun, A., Zhang, Z., Chen, J., Clark, B., Faunt, C.C., Haugh, C.J., Hill, M., Hobza, C., McGuire, V.L., Reitz, M., Müller Schmied, H., Sutanudjaja, E.H., Swenson, S., Wiese, D., Xia, Y., Zell, W., 2020. Comparison of groundwater storage changes from GRACE satellites with monitoring and modeling of major U.S. Aquifers. Water Resour. Res. 56, e2020WR027556. https://doi.org/10.1029/2020WR027556.

Reichle, R., De Lannoy, G., Koster, R. D., Crow, W. T., Kimball, J. S., Liu, Q., Bechtold, M. 2022. SMAP L4 Global 3-hourly 9 km EASE-Grid Surface and Root Zone Soil Moisture Geophysical Data, Version 7 [Data Set]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. Accessed: [09-07-2024], https://doi.org/10.5067/EVKPQZ4AFC4D.

RGI 7.0 Consortium, 2023. Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 7.0. Boulder, Colorado USA. NSIDC: National Snow and Ice Data Center. Accessed: [09-06-2024], https://doi.org/10.5067/f6jmovy5navz.

Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The global land data assimilation system. Bull. Amer. Meteor. Soc. 85, 381–394. https://doi.org/10.1175/BAMS-85-3-381.

Rodell, M., Reager, J.T., 2023. Water cycle science enabled by the GRACE and GRACE-FO satellite missions. Nat. Water 1, 47–59. https://doi.org/10.1038/s44221-022-00005-0.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.

Scanlon, B.R., Rateb, A., Pool, D.R., Sanford, W., Save, H., Sun, A., Long, D., Fuchs, B., 2021. Effects of climate and irrigation on GRACE-based estimates of water storage changes in major US aquifers. Environ. Res. Lett. 16, 094009. https://doi.org/10.1088/1748-9326/ac16ff.

Scanlon, B.R., Zhang, Z., Save, H., Sun, A.Y., Müller Schmied, H., Van Beek, L.P.H., Wiese, D.N., Wada, Y., Long, D., Reedy, R.C., Longuevergne, L., Döll, P., Bierkens, M. F.P., 2018. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. PNAS 115. https://doi.org/10.1073/pnas.1704665115.

Schumacher, M., Forootan, E., Van Dijk, A.I.J.M., Müller Schmied, H., Crosbie, R.S., Kusche, J., Döll, P., 2018. Improving drought simulations within the Murray-Darling

basin by combined calibration/assimilation of GRACE data into the WaterGAP global hydrology model. Remote Sens. Environ. 204, 212–228. https://doi.org/10.1016/j.rse.2017.10.029.

Smith, L.N., 2018. A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay. arXiv [Cs.LG]. Retrieved from http://arxiv.org/abs/1803.09820.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15 (1), 1929–1958. https://doi.org/10.5555/2627435.2670313.

Sun, A., 2024. Learning to downscale satellite gravimetry data through artificial intelligence. Nat. Water 2, 110–112. https://doi.org/10.1038/s44221-024-00199-5.

Sun, A.Y., Green, R., Swenson, S., Rodell, M., 2012. Toward calibration of regional groundwater models using GRACE data. J. Hydrol. 422–423, 1–9. https://doi.org/10.1016/j.jhydrol.2011.10.025.

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

Taylor, C.J., Alley, W.M., 2001. Ground-water-level monitoring and the importance of long-term water-level data (Vol. 1217). Denver, CO, USA: US Geological Survey. https://doi.org/10.3133/cir1217.

Tourian, M.J., Saemian, P., Ferreira, V.G., Sneeuw, N., Frappart, F., Papa, F., 2023. A copula-supported Bayesian framework for spatial downscaling of GRACE-derived terrestrial water storage flux. Remote Sens. Environ. 295, 113685. https://doi.org/10.1016/j.rse.2023.113685.

Uz, M., Atman, K.G., Akyılmaz, O., Shum, C.K., 2025. Global high-resolution terrestrial water storage anomalies through a dynamic soft-constrained deep learning paradigm. GFZ Data Services. https://doi.org/10.5880/GFZ.DQTO.2025.001.

Uz, M., Akyilmaz, O., Shum, C.K., 2024a. Deep learning-aided temporal downscaling of GRACE-derived terrestrial water storage anomalies across the Contiguous United States. J. Hydrol. 645, 132194. https://doi.org/10.1016/j.jhydrol.2024.132194.

Uz, M., Akyılmaz, O., Shum, C.K., Atman, K.G., Olgun, S., Güneş, Ö., 2024b. High-resolution temporal gravity field data products: monthly mass grids and spherical harmonics from 1994 to 2021. Sci. Data 11, 71. https://doi.org/10.1038/s41597-023-02887-5.

Watkins, M.M., Wiese, D.N., Yuan, D., Boening, C., Landerer, F.W., 2015. Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. JGR Solid Earth 120, 2648–2671. https://doi.org/10.1002/2014JB011547.

Wiese, D.N., Landerer, F.W., Watkins, M.M., 2016. Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution. Water Resour. Res. 52, 7490–7502. https://doi.org/10.1002/2016WR019344.

Wiese, D.N., Yuan, D.-N., Boening, C., Landerer, F.W., Watkins, M.M., 2023. JPL GRACE and GRACE-FO Mascon Ocean, Ice, and Hydrology Equivalent Water Height JPL. Ver. RL06.1Mv03. PO.DAAC, CA, USA. Accessed: [01-02-2024], https://doi.org/10.5067/TEMSC-3MJ63.

Wursthorn, K., Hillemann, M., Ulrich, M., 2022. Comparison of Uncertainty Quantification Methods for Cnn-Based Regression. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLIII-B2-2022, 721–728, https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-721-2022.

Xie, R., Fang, X., 2020. The unusual 2014–2016 El Niño events: dynamics, prediction and enlightenments. Sci. China Earth Sci. 63, 626–633. https://doi.org/10.1007/s11430-019-9561-2.

Zaitchik, B.F., Rodell, M., Reichle, R.H., 2008. Assimilation of GRACE terrestrial water storage data into a land surface model: results for the Mississippi River Basin. J. Hydrometeorol. 9, 535–548. https://doi.org/10.1175/2007JHM951.1.

Zemp, M., 2022. GCOS 2022 Implementation Plan. In: Chao, Qingchen; Han Dolman, Albertus Johannes;Herold, Martin; Krug, Thelma; Speich, Sabrina; Suda, Kazuto; Thorne, Peter; Yu, Weidong; Zemp, Michael. The 2022 GCOS Implementation Plan. Geneva: World Meteorological Organization, 85. https://doi.org/10.5167/uzh-224271.

Zemp, M., Huss, M., Thibert, E., Eckert, N., McNabb, R., Huber, J., Barandun, M., Machguth, H., Nussbaumer, S.U., Gartner-Roer, I., Thomson, L., Paul, F., Maussion, F., Kutuzov, S., Cogley, J.G., 2019. Global glacier mass changes and their contributions to sea-level rise from 1961 to 2016. Nature 568 (7752), 382–386. https://doi.org/10.1038/s41586-019-1071-0.

Zhang, G., Xu, T., Yin, W., Bateni, S.M., Jun, C., Kim, D., Liu, S., Xu, Z., Ming, W., Wang, J., 2024. A machine learning downscaling framework based on a physically constrained sliding window technique for improving resolution of global water storage anomaly. Remote Sens. Environ. 313, 114359. https://doi.org/10.1016/j.rse.2024.114359.