



Machine-learning-based reconstruction of long-term global terrestrial water storage anomalies from observed, satellite and land-surface model data

Nehar Mandal¹, Prabal Das², and Kironmala Chanda^{1,3}

¹Department of Civil Engineering, Indian Institute of Technology (Indian School of Mines), Dhanbad, JH 826004, India

²Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX 76019, USA

³Centre for Water Resource Management, Indian Institute of Technology (Indian School of Mines), Dhanbad, JH 826004, India

Correspondence: Kironmala Chanda (kironmala@iitism.ac.in)

Received: 28 March 2024 – Discussion started: 17 May 2024

Revised: 21 February 2025 – Accepted: 4 March 2025 – Published: 13 June 2025

Abstract. Understanding long-term terrestrial water storage (TWS) variations is vital for investigating hydrological extreme events, managing water resources and assessing climate change impacts. However, the limited data duration from the Gravity Recovery and Climate Experiment (GRACE) and its follow-on mission (GRACE-FO) poses challenges for comprehensive long-term analysis. In this study, we reconstruct TWS anomalies (TWSAs) for the period from January 1960 to December 2022, thereby filling data gaps between the GRACE and GRACE-FO missions and generating a complete dataset for the pre-GRACE era. The workflow involves identifying optimal predictors from land surface model (LSM) outputs, meteorological variables and climatic indices using a novel Bayesian network (BN) technique for raster-based TWSA simulations. Climate indices, like the Oceanic Niño Index and Dipole Mode Index, are selected as optimal predictors for a large number of grid cells globally, along with TWSAs from LSM outputs. The most effective machine learning (ML) algorithms among convolutional neural network (CNN), support vector regression (SVR), extra trees regressor (ETR) and stacking ensemble regression (SER) models are evaluated at each grid cell to achieve optimal reproducibility. Globally, ETR performs best for most of the grid cells; this is also noticed at the river basin scale, particularly for the Ganga–Brahmaputra–Meghna, Godavari, Krishna, Limpopo and Nile river basins. The simulated TWSAs (BNML_TWSA) outperformed the TWSAs from LSM outputs when evaluated against GRACE datasets. Improvements are particularly noted in river basins such as the Godavari, Krishna, Danube and Amazon, with median correlation coefficient, Nash–Sutcliffe efficiency, and RMSE values for all grid cells in the Godavari Basin, India, being 0.927, 0.839 and 63.7 mm, respectively. A comparison with TWSAs reconstructed in recent studies indicates that the proposed BNML_TWSA outperforms them globally as well as for all of the 11 major river basins examined. Furthermore, the uncertainty of BNML_TWSA is assessed for each grid cell in terms of the standard error. Results show smaller standard error magnitudes in grid cells in arid regions compared to other regions. The presented gridded dataset is published at <https://doi.org/10.6084/m9.figshare.25376695> (Mandal et al., 2024), featuring a spatial resolution of $0.50^\circ \times 0.50^\circ$ and offering global coverage.

1 Introduction

Terrestrial water storage (TWS) refers to the storage of water on or above the surface of the Earth, including hydrological elements such as groundwater, soil moisture, snow, ice and surface water (Yu et al., 2021; Yang et al., 2021). The fluctuations in TWS in both space and time have been comprehensively simulated by employing physically based land surface models (LSMs) and global hydrological models (GHMs) (Humphrey et al., 2017; Felfelani et al., 2017; Sun et al., 2021). These models have significant biases due to inherent uncertainty and the inadequate representation of some physical processes, such as the lack of modeling of human water resource interventions within LSMs (Bibi et al., 2024). Furthermore, in snow-dominated basins, LSMs often underestimate peak terrestrial water storage anomalies (TWSAs), whereas GHMs tend to overestimate them. Similarly, in temperate, arid and tropical basins, both model types generally underestimate TWSA peaks (Bibi et al., 2024), which limits the application of these model outputs for long-term analyses of climate change impact assessment, water resource management and hydrological extreme event forecasting. The Gravity Recovery and Climate Experiment (GRACE) mission and its successor GRACE Follow-on (GRACE-FO) have been providing unprecedented accurate measurements of TWSAs since April 2002 (Mo et al., 2022). These TWSA observations have been widely used in hydrological studies to assess the impacts of climate change and human activities on large-scale water balance, droughts and floods, and groundwater storage (Rodell et al., 2018). Similar to other satellite observations, the GRACE mission has data gaps, such as an 11-month gap between the end of the GRACE mission in June 2017 and the start of the GRACE-FO mission in May 2018 as well as an additional few months of data gap during each mission.

A reliable, long-term, continuous TWSA dataset is imperative for the assessment of basin-scale water balance and local hydrological extremes. Different studies have employed various techniques to reconstruct long-term TWSAs beyond the GRACE period. In a study by Becker et al. (2011), the authors integrated spatial patterns of TWS derived from GRACE data with long-term in situ river level records to recreate the TWS for the Amazon Basin from 1980 to 2008. Forootan et al. (2014) developed an autoregressive model with exogenous variables (ARX), a statistical data-driven approach, to reconstruct TWSAs for West Africa using the GRACE dataset; rainfall data from the Tropical Rainfall Measuring Mission (TRMM); and sea surface temperature (SST) information spanning the Atlantic, Indian and Pacific oceans. Ahmed et al. (2019) employed an ARX model to establish a relationship between GRACE TWS and meteorological variables, as well as vegetation indices. Nie et al. (2016) utilized the Global Land Data Assimilation System (GLDAS) products and GRACE-based TWSA data to reconstruct TWSAs using the water balance approach from 1948

to 2012 over the Amazon Basin. Humphrey et al. (2017) established a statistical data-driven model that linked GRACE TWSAs with deviations in both temperature and precipitation to recreate TWSAs from 1985 to 2015 for the entire globe.

Machine-learning-based algorithms have gained popularity over the past 2 decades, presenting new opportunities in hydrology and related fields, including the reconstruction of TWSAs. Long et al. (2014) made one of the first attempts to hindcast TWSAs for the period from February 1979 to September 2012 by developing an artificial neural network (ANN) model using GRACE data and other in situ modeling data to study extreme climate events in the Yungui Plateau, China. Sun et al. (2019) used a deep convolutional neural network (CNN) with three model architectures to predict the spatiotemporal variations in TWSAs over India. Similarly, Li et al. (2020) and Sun et al. (2020) utilized multiple linear regression (MLR) and neural-network-based models, in conjunction with ARX, to reconstruct TWSAs globally. Jing et al. (2020) generated a GRACE-like TWSA prior to the GRACE period, dating back to 1979, over the Nile Basin using the Random Forest and eXtreme Gradient Boosting ensemble learning algorithms. Yu et al. (2021) used three deep learning models to hindcast TWSAs over Canada from 1972 to 2002 based on land surface model (LSM) output as a predictor. Satish Kumar et al. (2023) reconstructed GRACE-like time series of TWSAs from 1960 to 2016 across four river basins in southern India.

The reconstruction of TWSAs has been a significant area of research, with studies such as that by Sun et al. (2020) emphasizing the need to enhance the performance accuracy of machine learning (ML) and statistical algorithms. Most global TWSA reconstruction studies primarily employ a single ML algorithm to model TWSAs (Sun et al., 2020; Mo et al., 2022; Li et al., 2020). However, some studies, including those by Li et al. (2020) and Sun et al. (2020), also employ statistical algorithms in conjunction with ML. Consequently, the reliance on a single ML algorithm for each grid cell presents a limitation that needs to be addressed to improve the robustness and accuracy of TWSA reconstructions. Different algorithms have varying strengths with respect to handling the nonlinearities and complexities inherent in regional hydrological systems. Testing multiple ML algorithms can ensure methodological robustness and help identify the appropriate approach for TWSA reconstruction.

The ML models used in hydrological studies so far can be broadly divided into two main categories: single-algorithm usage and multiple-algorithm usage. Most ML investigations predominantly belong to a single-algorithm-usage category, where the performance of a specific algorithm is evaluated against the baseline performance (Raghavendra and Deka, 2014; Sun et al., 2014; Mo et al., 2022; Khan and Maity, 2020). Other studies compare the performance of different ML models with the aim of finding the best algorithm, which performs well across a wide range of situations (Mandal and

Chanda, 2023; Sun et al., 2021). Using a single-algorithm approach could prove satisfactory when dealing with a compact research area or multiple study areas with similar hydroclimatic characteristics. However, it becomes insignificant for large study areas, multi-site analysis with different hydroclimatic conditions and in cases where the relative importance of predictors can vary spatially. This may greatly affect the final performance of ML models. Previous studies on grid-cell-scale reconstruction of GRACE TWSAs indicate that there is no individual algorithm that consistently outperforms others across all global basins (Sun et al., 2020, 2021; Li et al., 2020). Sun et al. (2020) found that a deep neural network model performed better than the other two data-driven methods for reconstructing TWSAs over global river basins at a grid cell scale. Mo et al. (2022) employed Bayesian convolutional neural networks (BCNNs) to reliably interpolate the TWSA data gap between GRACE and GRACE-FO globally. Deep convolutional autoencoders outperform CNNs and BCNNs when filling the gaps between GRACE and GRACE-FO globally (Uz et al., 2022). The integrated CNN-based support vector machine has been found by Kalu et al. (2023) to outperform other regression models in the Congo Basin, Africa.

In recent times, a relatively new approach has been adopted by many studies, wherein the optimal features are selected before applying a single and/or multiple ML algorithms to evaluate the prediction accuracy (Das and Chanda, 2020; Das et al., 2022). The significance of predictor selection has been emphasized by several studies, which point out the inadequacies of current algorithms for identifying efficient or optimal predictors (Mo et al., 2022; Sun et al., 2020; Li et al., 2020). Sun et al. (2020) specifically raises concerns about determining which set of predictors or predictor groups play a more significant role, which is crucial for enhancing the interpretability and explainability of ML solutions. To the best of our knowledge, the selection of optimal predictors for the reconstruction of TWSAs using Bayesian networks is exercised for the first time in this study, which also highlights the novelty of the study. Selection of optimal predictors is not only a methodological novelty but also a critical step to ensure that the model prioritizes the most relevant and physically meaningful predictors. This approach reduces noise, minimizes overfitting and enhances interpretability, making the final product more scientifically robust and practically useful (Das and Chanda, 2024).

In the present study on the reconstruction of TWSAs, the potential of Bayesian networks is utilized for the selection of optimal predictors from a broad set of inputs comprising observed, satellite and land-surface-based data products. The input and target datasets are used without prior interpolation of intermittent gaps, detrending, deseasoning or decomposing signals, in order to prevent the introduction of bias. We applied different ML models to each global grid cell in this study and selected the most appropriate model for each grid cell based on their performance to ensure optimal

reproducibility. Furthermore, we conduct the analysis at the basin scale across 11 global river basins with varied hydroclimatic characteristics from six different continents: Amazon, Danube, Ganga–Brahmaputra–Meghna (GBM), Godavari, Indus, Krishna, Limpopo, Mississippi, Murray–Darling, Nile and Zambezi. Among these rivers, the Amazon, GBM and Mississippi exhibit humid hydrologic characteristics, while the Nile is semiarid and the Zambezi is semi-humid (Uz et al., 2022). Furthermore, a diverse range of basin sizes has been taken into account, spanning from the vast Mississippi Basin, which covers 2 918 820 km², to the relatively small Krishna Basin, with an area of 258 948 km². The study aims to achieve three objectives. First, it aims to specifically select the optimal predictors for TWSAs from a number of meaningful inputs, including LSM outputs, meteorological variables and climate indices, for each grid cell, utilizing the potential of Bayesian networks. Second, it aims to select a leader model for each grid cell from a number of ML models, including kernel-based, network-based and ensemble models, based on their performance. Finally, it aims to simulate GRACE-like TWSAs and reconstruct a global TWSA datasets for the historical period starting from 1960 and including the data gap periods of the GRACE and GRACE-FO missions.

In the rest of the paper, Sect. 2 describes the data used and the processing of data products. In Sect. 3, the methodological details of the predictor selection process, the description of ML algorithms and the overall workflow are presented. The results and discussions are presented in Sect. 4, followed by a summary of the conclusions in Sect. 7.

2 Data and processing

The complete set of predictors used in this study includes TWSA data products from GLDAS LSMs, climate forcing data (precipitation and temperature) and a number of climate indices. Brief descriptions of data products used in this study and their sources are discussed in the subsections below. The entire period of analysis spans from 1960 to 2022.

2.1 GRACE terrestrial water storage anomalies (TWSAs)

This study makes use of the Coastline Resolution Improved version of the GRACE mascon product (RL06.1Mv03) downloaded from the Jet Propulsion Laboratory (JPL RL06) website (<https://grace.jpl.nasa.gov>, last access: 27 May 2023). The JPL RL06 has a $0.5^\circ \times 0.5^\circ$ (latitude \times longitude) spatial resolution; however, it naturally symbolizes $3^\circ \times 3^\circ$ equal-area caps, which match the mass concentration (mascon) functions used to estimate and parameterize the monthly gravity fields globally (Wiese et al., 2016; Watkins et al., 2015). Compared to classic solutions that use traditional spherical harmonic basis functions, mascon solutions offer significant improvements. Furthermore, mascon solutions do

not require postprocessing filters to mitigate errors, unlike spherical harmonics. JPL mascon is employed in this study, as this dataset is widely utilized and validated in the literature (Bibi et al., 2024; Scanlon et al., 2021; Watkins et al., 2015). The GRACE datasets are provided as anomalies with respect to the 2004 to 2009 mean terrestrial water storage (TWS). Two missions covered the observation period of the GRACE dataset: from April 2002 to June 2017, during the GRACE mission, and from June 2018 to the present day, under the GRACE-FO mission, with an 11-month gap between the two missions. Additionally, there are intermittent data gaps within each mission. In this study, short data gaps (1–2 months) are also filled using trained ML models, unlike some previous studies where the data from neighboring months are used to fill in these gaps (Mo et al., 2022; Yang et al., 2021).

2.2 GLDAS-simulated TWS

GLDAS-LSM-simulated TWS data from two different models, the Catchment Land Surface Model (CLSM) (Li et al., 2019) and NOAH (Rodell et al., 2004), are used in this study. Both LSM data products are retrieved from GES DISC, a NASA Goddard Earth Sciences Data and Information Services Center (<https://disc.gsfc.nasa.gov>, last access: 28 July 2023). CLSM GLDAS CLSM025 v2.0 covers the period from January 1948 to December 2014, while GLDAS CLSM025 v2.2 spans from February 2003 to the present; variables in these versions include TWS as an output (Sun et al., 2021). On the other hand, NOAH (GLDAS_NOAH025) TWS is estimated as an aggregate of soil moisture content (in all four layers, ranging in depth from 0 to 200 cm), canopy water storage and snow depth water equivalent. These three variables are available from January 1948 to the present and consist of two data versions (GLDAS_NOAH025 v2.0 and GLDAS_NOAH025 v2.1). The spatiotemporal resolution and other details of all data products used in this study are presented in Table 1. To obtain the GLDAS TWSAs, the long-term TWS mean from January 2004 to December 2009 is subtracted from the corresponding GLDAS TWS data (Sun et al., 2021). The corresponding TWSAs from NOAH and CLSM are denoted by NTWSA and CTWSA, respectively. To hindcast the TWSAs for the period from January 1960 to March 2002, older versions (v2.0) of the GLDAS LSM data products are used.

2.3 Meteorological data

Meteorological data, such as precipitation (P) and temperature (T), have been included as predictors to enhance the model's predictive capability. Although precipitation and temperature are components of LSM forcing, the LSM does not utilize all of the data in the forcing to their full potential (Sun et al., 2021). The amount of precipitation affects the recharge of groundwater and surface waters, while tem-

perature is an indicator of the energy available for evapotranspiration. Hence, precipitation and temperature may capture some specific aspect that the LSM models may not simulate accurately (Sun et al., 2019; Humphrey et al., 2017). These climate forcing data (precipitation and temperature) are obtained from GLDAS NOAH LSM output for the period of analysis. These products are selected because of their global coverage and successful usage in earlier investigations (Sun et al., 2019).

2.4 Climate indices

The Dipole Mode Index (DMI), North Atlantic Oscillation (NAO) and Oceanic Niño Index (ONI) have been widely utilized as optimal teleconnection predictors for the seasonality of surface temperature and precipitation (Harou et al., 2006; Brandimarte et al., 2011; Hafez, 2016). DMI is the anomalous SST gradient between the western and south-eastern equatorial Indian oceans associated with the Indian Ocean Dipole (Saji et al., 1999; Saji and Yamagata, 2003). The NAO characterizes the changes in strength between the subtropical high-pressure and subpolar low-pressure patterns in the atmosphere over the North Atlantic Ocean (Wallace and Gutzler, 1981; Barnston and Livezey, 1987). The difference between a rolling 3-month average SST in the eastern central tropical Pacific and the long-term average of the same 3 months is characterized as the ONI (Barnston et al., 1997). Recent studies have utilized some climate indices as predictors of TWSAs (Forootan et al., 2019; Phillips et al., 2012; Sun et al., 2021). The prediction skills are spatially impacted by specific climate phenomena or conditions, such as El Niño and La Niña events, represented by those climate indices (Sun et al., 2021).

3 Methodology

As mentioned earlier, TWS from CLSM is obtained directly as one of the outputs (Sun et al., 2021), whereas TWS from NOAH is obtained as the sum of the following components (Sun et al., 2019):

$$\text{TWS} = \text{SnWE} + \text{SMC} + \text{CWS}, \quad (1)$$

where SnWE represents snow depth water equivalent, SMC is soil moisture content and CWS is canopy water storage. For both of the aforementioned GLDAS products, the anomalies are computed by subtracting the long-term mean monthly TWS of the period from January 2004 to December 2009 from the monthly TWS values. Let the predictand variable GRACE TWSAs be denoted by t and the set of predictors be denoted by X . Then, the regression problem can be expressed as follows:

$$t = f(X, p), \quad (2)$$

where f represents the regression model and p denotes the model parameter to be solved using $\{X_i, t_i\}_{i=1}^N$ as training

Table 1. Information regarding the data products employed in this research. Web links for accessing these data products are also provided in Sect. 5.

Product	Source	Variables	Spatial resolution (lat × long), temporal resolution	Version (data period)
GRACE JPL mascon	Watkins et al. (2015), NASA/JPL (2023) and https://doi.org/10.5067/TEMSC-3JC63	Terrestrial water storage anomaly (TWSA)	0.50° × 0.50°, 1 month	RL06.1Mv03 (April 2002–December 2022)
GLDAS CLSM	Li et al. (2019), https://doi.org/10.5067/LYHA9088MFQ and https://doi.org/10.5067/TXBMLX370XX8	Terrestrial water storage (CTWSA)	0.25° × 0.25°, 1 d (aggregated to monthly)	GLDAS_CLSM025 v2.0 (Jan 1960–Jan 2003) GLDAS_CLSM025 v2.2 (Feb 2003–Dec 2022)
GLDAS NOAH	Rodell et al. (2004), https://doi.org/10.5067/9SQ1B3ZXP2C5 and https://doi.org/10.5067/SXAVCZFAQLNO	(Snow depth water equivalent + Soil moisture content + Canopy water storage) (NTWSA), precipitation (P), temperature (T)	0.25° × 0.25°, 1 month	GLDAS_NOAH025 v2.0 (Jan 1960–Mar 2002) GLDAS_NOAH025 v2.1 (Apr 2002–Dec 2022)
Dipole Mode Index	Saji and Yamagata (2003), Saji et al. (1999), and https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data	DMI	1 month	–
North Atlantic Oscillation	Wallace and Gutzler (1981), Barnston and Livezey (1987), and https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/norm.nao.monthly.b5001.current.ascii	NAO	1 month	–
Oceanic Niño Index	Barnston et al. (1997) and https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt	ONI	1 month	–

data; in the latter expression, $i = 1 \dots N$ is the index of training sample, and X includes CTWSA, NTWSA, P and T for the current month as well as for 1 and 2 months prior. Additionally, it includes three climate indices (DMI, NAO and ONI) for the current month. In this context, P , P_1 and P_2 represent precipitation for the current month, 1 month prior and 2 months prior, respectively. Four ML algorithms, namely, convolutional neural network (CNN), support vector regression (SVR), extra trees regressor (ETR) and stacking ensemble regression (SER), are trained to solve the regression problem described in Eq. (2). ML models built after training and validation can be used to simulate GRACE-like TWSAs using the inputs only. Figure 1 illustrates the overall workflow adopted in this study. An overview of the predictor selection technique and a brief description of the ML models used in this study are presented in the subsections below.

3.1 Predictor selection using Bayesian networks (BNs)

Among the multiple predictors mentioned in the previous section, the most relevant predictors for simulating TWSAs are selected utilizing the potential of Bayesian networks (BNs). The optimal predictors are a subset of all potential predictors of 15 variables. At each grid cell, out of the 15 variables, a subset is selected using BNs, which serves as the “optimum set of predictors” for that grid cell. The optimum set of predictors are selected based on probabilistic independence–dependence structure. BNs serve as compact representations of probabilistic relationships among a defined collection of random variables (Das and Chanda, 2022). These networks are characterized by a graph $G = (V, E)$, where each vertex (node) $v \in V$ corresponds to one of the aforementioned random variables in X . Through edges (arcs) $e \in E$, the network articulates the conditional independencies or dependencies that exist among the variables within X , collectively termed the graph’s dependence structure. This

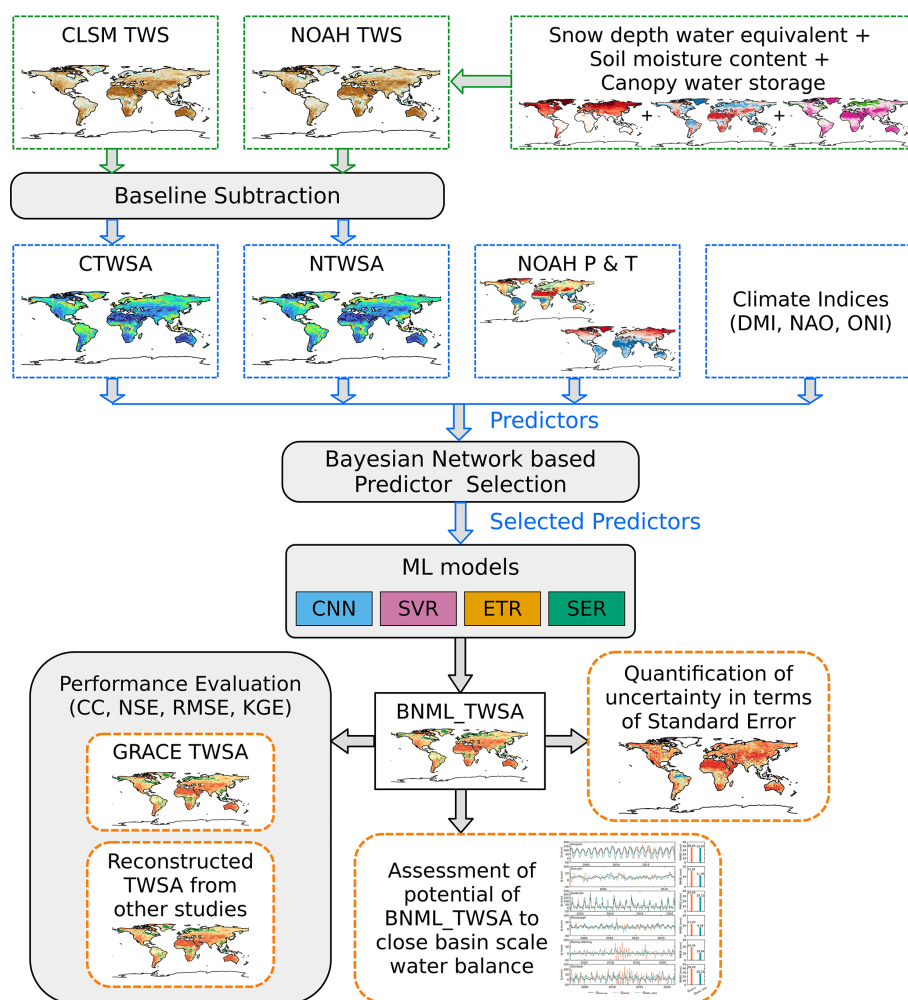


Figure 1. An overview of the methodology employed in this study.

framework employs directed acyclic graphs (DAGs) to concisely encapsulate probabilistic relations and effectively portray the joint distribution of the variables (Scutari and Nagarajan, 2011). In the context of a DAG, the inclusion of arcs may or may not indicate a causal relationship between variables where one variable can be understood as the cause and the other as the effect (Sevinc et al., 2020). When an edge connects two nodes in the graph, the node from which the edge originates is referred to as the parent node, while the node where the edge terminates is known as the child node. The process of determining the topology of the graph G is termed structure learning. This involves identifying the graph structure that effectively represents the conditional independencies observed within the data. Several algorithms have been presented in the literature to tackle this problem, and they are broadly classified into three categories: constraint-based algorithms (which are based on conditional independence tests), score-based algorithms (which are based on goodness-of-fit scores) and hybrid algorithms (which combine the previous two approaches). In the realm of contin-

uous variables, score-based algorithms have outperformed their constraint-based counterparts in terms of performance. The challenge with constraint-based algorithms lies in their tendency to yield partially directed acyclic graphs (PDAG), which can subsequently disrupt precise predictions for the target variable (Das and Chanda, 2020).

Score-based learning offers both precise and approximate solutions. Precise solutions ensure the identification of the graph that optimizes an objective function while adhering to a maximum in-degree constraint (Constantinou et al., 2021). These algorithms, also referred to as search-and-score algorithms, involve the utilization of heuristic optimization methods to tackle the task of acquiring the structure of a BN. In this process, each potential network configuration is assigned a score that indicates its goodness of fit, and the objective of the algorithm is to maximize this score. The hill-climbing (HC) algorithm, a classical heuristic search algorithm, has been the major choice for this purpose in most of the hydrology-related studies (Vitolo et al., 2018; Das and

Chanda, 2022, 2020; Dutta and Maity, 2021; Chanda and Das, 2022).

The HC algorithm commences its process with an initial empty graph and systematically explores potential DAGs to maximize the network's score, achieved through operations like adding, removing or reversing arcs, determined by the strength of edges (Scutari and Denis, 2021). The network's score is calculated using the Bayesian information criteria (BIC) to prevent overfitting, wherein the BIC score is determined by a formula incorporating the number of parameters in the global distribution (“ d ”) and the dataset length (“ N ”). A critical concept within the HC algorithm is edge strength, reflecting the disparity in the BIC score when a specific arc is included or excluded, indicating the degree of interdependency between connected nodes. A higher edge strength signifies a stronger correlation. In essence, the HC algorithm employs an iterative approach, optimizing the network score by manipulating the DAG's structure based on edge strengths determined through the BIC score, thereby striking a balance between model complexity and data fidelity while also emphasizing significant node interconnections. The BIC score formula is given by the following expression:

$$\text{BIC} = \sum_{i=1}^N \log(\text{Pr}(X_i | \text{MB}(X_i))) - \frac{d}{2} \log(N), \quad (3)$$

where MB represents the Markov blanket, which is the minimal set of nodes that can predict a target node. The `bnlearn` package in the R environment (Scutari, 2010) is used to develop the DAG networks.

3.2 Machine learning algorithms for TWSA modeling

This section presents a concise overview of ML algorithms used to model TWSAs. In this study, four types of ML algorithms have been used: neural network-based algorithms (CNN), kernel-based algorithms (SVR), tree-based algorithms (ETR) and an ensemble of these three (CNN, SVR and ETR) as stacking ensemble regression (SER).

3.2.1 Convolutional neural networks (CNNs)

CNNs belong to a category of neural networks particularly well suited to handling data with grid-like structures, such as images or time-series data. The primary advantage that they offer over conventional feed-forward neural networks is their utilization of mathematical linear operations, known as convolutions (Uz et al., 2022). These layers possess the capability to autonomously extract features, identifying crucial aspects within the input data that are essential for establishing the correlation between input and output variables. Hence, CNNs have the capacity to manage raw data and are devoid of the necessity to preprocess or manually extract features (Ferreira and da Cunha, 2020). CNNs find widespread application in image recognition tasks. For images, which possess two dimensions, convolutional filters of corresponding

dimensions are employed. Conversely, for tasks involving sequential data or time series, like the context of this study, CNNs with one-dimensional (1D) convolutional filters (1D-CNNs) are employed. In this study, a 1D-CNN comprises a single convolutional layer and two fully connected layers. The activation function “ReLU” is applied within the convolutional layer (Ferreira and da Cunha, 2020; Alibabaei et al., 2021; Ahmed et al., 2022). To mitigate overfitting, a dropout layer is introduced following each convolutional layer. The training algorithm employed for this model is Adam. The learning rate is established at 0.1, and the number of training epochs is determined using early stopping, with a maximum limit of 200 epochs. Additionally, the batch size is configured to 32.

3.2.2 Support vector regression (SVR)

Support vector regression (SVR) is a pivotal component of support vector machines (SVMs), an algorithm introduced by Cortes and Vapnik (1995), designed to handle nonlinear regression problems. SVR extends the fundamental concept of SVMs by effectively addressing regression tasks through a nonlinear mapping of input data into a higher-dimensional feature space. The underlying principle of SVR is based on the concept of the “kernel trick”. This technique employs a kernel function, such as the radial basis function (RBF) kernel, to transform the input data into a feature space where linear regression can be applied effectively. The kernel function aids in defining a hyperplane – a decision boundary – within this feature space. This hyperplane facilitates the prediction of target values by distinguishing between different types of data patterns. Moreover, SVR aims to establish a boundary layer at a certain distance from the hyperplane, enclosing the data points that lie proximate to the hyperplane, known as support vectors. Selecting an appropriate kernel function is a crucial step in SVR. Raghavendra and Deka (2014) highlight that polynomial and RBF kernels are commonly employed for nonlinear problems. However, Das and Chanda (2020) advocate for the superiority of the RBF kernel in nonlinear regression tasks, leading to its preference. For a more comprehensive understanding of the algorithms employed in this context, detailed insights and further elucidation can be gleaned from the work of Raghavendra and Deka (2014), Das and Chanda (2020), and Das et al. (2022).

3.2.3 Extra trees regressor (ETR)

In the realm of ensemble-based predictive modeling, the ETR shares a fundamental principle that is built on the foundation of decision trees and random forests, combining their strengths to create an ensemble of diverse decision trees, and is less likely to overfit a dataset (Ahmad et al., 2018). In the ETR framework, a random subset of features is employed to train each individual base estimator, akin to the approach used in random forest, where all of the predictors are

employed. This attribute ensures diversity among the constituent trees, contributing to the model's generalization capabilities and mitigating the risk of overfitting. While random forest selects the optimal feature from the random subset to split a node, ETR takes this a step further by introducing an additional layer of randomness (Kumar et al., 2022). ETR not only randomly selects a feature from the subset but also stochastically chooses the corresponding split value for the chosen feature. This distinctive feature selection strategy imparts an extra level of variance to the model, rendering it more robust and capable of capturing intricate relationships in the data (Ahmad et al., 2018). The predictions of the individual trees are combined to generate the ultimate prediction through a process of arithmetic averaging. The algorithm is influenced by two pivotal parameters: the count of predictors chosen randomly at each node and the minimum sample size required to initiate a node split (Sun et al., 2021).

3.2.4 Stacking ensemble regression (SER)

Stacking, a form of meta-learning, aims to enhance predictive performance by combining predictions from multiple base models through a higher-level integrated model (Zounemat-Kermani et al., 2021). The stacking generalization framework presents a couple of avenues for maximizing predictive gains. One approach involves employing diverse base learners, thereby fostering variability among the base models. Alternatively, enhancing the ensemble size while keeping the number of base learners constant can also provide the meta-learner with a broader range of insights. In this context, the term “meta-learner” refers to an aggregating model that learns the optimal way to combine outputs from the base learners. These “base learners” constitute the models whose individual predictions are assembled in the final step (Lee and Ahn, 2021). To mitigate overfitting, out-of-sample data are employed for training the meta-model. This entails utilizing predictions from the base learners on this external data. The overarching objective is for the meta-model to establish an optimal correlation between observed values and its own predictions. The process is aptly termed “stacking”, as it involves merging predictions from validation sets, thereby creating a fresh dataset for the meta-model to glean insights from. Additional valuable insights can be obtained from the comprehensive analysis of ensemble ML presented in the review of Martinez-Gil (2022), Zounemat-Kermani et al. (2021), and Lee and Ahn (2021). In the present study, CNN, SVR and ETR are used as the base models. During the training process, initial predictions are first generated by the base learners. Subsequently, these predictions from the base learners serve as inputs to the meta-learner, which produces the ultimate output. While all base models are utilized as potential meta-models to assess overall accuracy, preference is given to the generalized linear model (GLM) due to its superior effectiveness as a meta-learner compared to the alternative models.

3.3 Training and performance evaluation

The period of observed GRACE data used in this study is from April 2002 to December 2022 (216 months in total excluding the GRACE data gaps). Within this period, the available months from 2010 to 2016 (68 months) are used as the testing period, whereas the remaining portion of the dataset (148 months) is used to train the ML models. The 5-fold cross-validation technique is employed during the training phase to address the issue of insufficient data length. Additionally, all input parameters for the ML models have been normalized within the range of 0 to 1. The simulated TWSAs employing BNs as the optimal predictor selector in conjunction with various ML models are henceforth referred to as BNML_TWSA. The performance of the BNML_TWSA from each of the models is evaluated against GRACE/GRACE-FO TWSAs using several agreement metrics, including the Pearson correlation coefficient (CC), the Nash–Sutcliffe efficiency (NSE) coefficient, the root-mean-square error (RMSE) and the Kling–Gupta efficiency (KGE). The CC is calculated as follows:

$$CC = \frac{\sum_{i=1}^n (O_i - \bar{O})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (S_i - \bar{S})^2}}, \quad CC \in [-1, 1], \quad (4)$$

where O_i and S_i represent the TWSAs from GRACE/GRACE-FO and the simulated/reconstructed TWSAs, respectively, with \bar{O} and \bar{S} denoting their respective means.

The other three metrics are expressed as follows:

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - S_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2}, \quad NSE \in (-\infty, 1], \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - S_i)^2}, \quad RMSE \in [0, +\infty), \quad (6)$$

$$KGE = 1 - \sqrt{(CC - 1)^2 + \left(\frac{\bar{S}}{\bar{O}} - 1\right)^2 + \left(\frac{\sigma_S}{\sigma_O} - 1\right)^2}, \quad KGE \in (-\infty, 1]. \quad (7)$$

Here, σ_O is the standard deviation in observations and σ_S the standard deviation in simulations. The size of the test sample is denoted by n . A higher value for CC, NSE and KGE, closer to 1, and a lower value for RMSE, closer to 0, indicate superior performance (Nash and Sutcliffe, 1970; Gupta et al., 2009).

4 Results and discussion

4.1 Selected predictors using BNs

The optimal predictors for each grid cell across the globe (58 027 grid cells in total) are selected using BNs. For each

predictor, the spatial map showing the grid cells where they are selected by the BNs as optimal predictors is depicted in Fig. 2. Figure 2 also shows a bar plot with the number of grid cells where each predictor is retained by the BN. CTWSA is selected by the BN as an optimal predictor for the maximum number of grid cells (71.06 %), followed by ONI (56.03 %) and NTWSA (48.71 %). The selection of CTWSA as the best predictor of GRACE TWSAs is also suggested by other studies conducted in different regions around the globe (Sun et al., 2021). Among the 11 global river basins specifically investigated in this study, the number of grid cells with CTWSA as an optimal predictor is lower in the Amazon and Nile basins compared to the other basins. Only 3.55 % of the total grid cells hold the NAO as an optimal predictor as selected by the BNs. This is reflected at the river basin scale as well; for each of the river basins, a very limited percentage of grid cells retain the NAO as an optimal predictor. It is noteworthy that, in addition to TWSAs from LSMs, the ONI and DMI have been selected as optimal predictors for a substantial number of grid cells. Meteorological variables such as P and T , along with their observations from previous months, have been selected for fewer grid cells globally by the BN compared to ONI and DMI. This can be attributed to the fact that LSMs already incorporate these meteorological variables as forcing inputs. The inclusion of climate indices as potential predictors for a large number of grid cells can be seen as an effort to represent the climate change scenarios of that specific time period. In a limited number of grid cells (66), the BN did not select any predictors. For an additional 492 grid cells, the BN selected only one predictor, thereby limiting the application of certain ML algorithms to these grid cells. Consequently, for a total of 558 grid cells, which constitute less than 1 % of the grid cells considered in this study, the complete set of 15 predictors has been used as potential predictors.

4.2 Grid-cell-specific leader models

The predictors selected by the BNs at each grid cell are employed as input to predict the TWSAs using the four ML algorithms mentioned earlier: CNN, SVR, ETR and SER. The grid-cell-wise leader ML algorithm is identified based on the Pearson correlation coefficient (CC) between predicted TWSAs and GRACE TWSAs for the test period. The performance difference between the leading ML algorithm and the worst-performing ML model is depicted in Fig. 3. Grid cells that have negative CC values (~ 2.65 % of the total grid cells considered in this study) for the leading ML model have been excluded when calculating the difference. Although the improvement in terms of the CC value difference is not large for all grid cells globally, more than 14.4 % of grid cells show improvements greater than 0.2, while an additional 15.7 % of grid cells exhibit improvements of between 0.1 and 0.2. A grid cell exhibiting the maximum improvement within the basins considered in this study has been selected to demon-

strate this improvement. The time series and scatterplot are illustrated in Fig. 4. The estimated TWSAs by the best-performing model are in good agreement with the observed TWSAs during the testing period. This justifies the use of the best-performing (leading) model to predict the TWSAs. Figure 5 depicts the spatial distribution of the leader algorithms over the globe along with the frequency as a bar plot. ETR performs the best for the maximum number of grid cells, with a total of 25 703, followed by SVR, SER and CNN, which perform best for 11 609, 11 069 and 9646 grid cells, respectively. Thus, for most of the river basins (including the Krishna and Godavari in India; the Danube in Europe; the Nile, Zambezi and Limpopo on the African continent; the Mississippi in the USA; and the transboundary GBM and Indus), ETR emerges as the leader model in the maximum number of grid cells. The contribution of the leader algorithm as a percentage of the total grid cells for each river basin is shown in Fig. 5c. In the Limpopo Basin, it is observed that ETR performs best in 89.0 % of the grid cells, whereas CNN does not perform best in any of the grid cells in this basin. In the Murray–Darling Basin in Australia, the four ML algorithms show the best performance in an approximately equal number of grid cells (CNN: 25.9 %; SVR: 21.4 %; ETR: 26.1 %; SER: 26.6 %).

4.3 Performance evaluation of simulated global BNML_TWSA

For the leader ML models at each grid cell, the BNML_TWSA is evaluated against the GRACE TWSAs during the testing period (68 months). Performance measures, such as the CC, NSE, RMSE and KGE, are computed for the BNML_TWSA at both the basin-wide and grid cell levels. Similar performance measures are also computed for CTWSA and NTWSA. The spatial distribution of CC, NSE and KGE obtained from NTWSA, CTWSA and the BNML_TWSA using the identified leader ML models during the test period is shown in Fig. 6. According to Fig. 6a and d, it is evident that the agreement between CTWSA and GRACE TWSAs is better than that of NTWSA. However, the BNML_TWSA (as shown in Fig. 6g) performs better in most grid cells worldwide compared to the TWSAs obtained from the LSMs (CTWSA and NTWSA). The BNML_TWSA showed clear improvement in performance over NTWSA and even CTWSA in all of the basins, except for the western part of the Nile Basin and the southwestern part of the Mississippi Basin where CTWSA shows a closer match with the GRACE TWSAs in some grid cells. The cumulative distribution functions (CDFs) of the CC, NSE and KGE metrics are presented in Fig. 7. These CDFs demonstrate that BNML_TWSA exhibits significantly superior performance, characterized by substantially higher CC, NSE and KGE values. As shown by the BNML_TWSA results, the arid, semi-arid and certain wet regions (e.g., the arid desert part of the Nile Basin, the semiarid regions of the Mississippi Basin

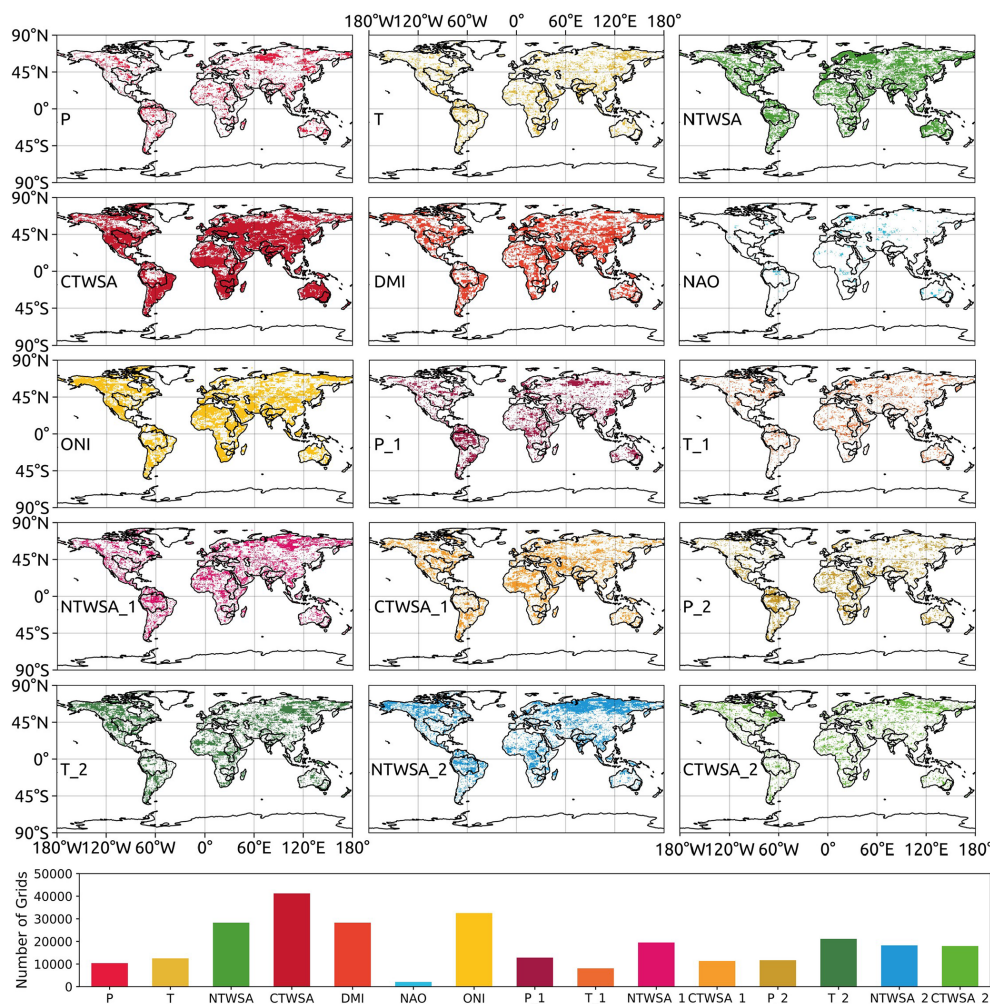


Figure 2. Spatial distribution and bar plot of selected predictors using a Bayesian network. P , P_1 and P_2 represent precipitation for the current month, 1 month prior and 2 months prior, respectively. Similarly, T , NTWSA and CTWSA, along with their observations 1 month prior and 2 months prior, are used as potential predictors. T denotes temperature, while TWSAs from NOAA and the Catchment Land Surface Model (CLSM) are denoted using NTWSA and CTWSA, respectively.

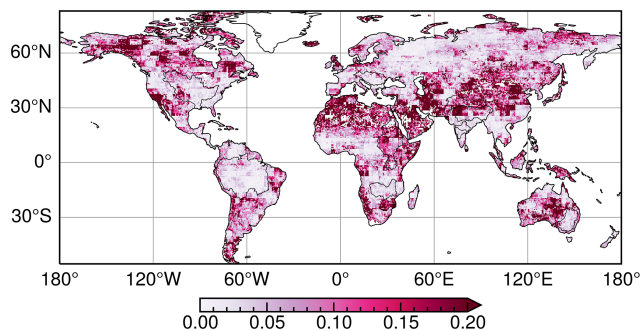


Figure 3. Difference between the correlation coefficient (CC) values obtained from the leader ML model (excluding negative CC values) and the worst-performing ML algorithm at each grid cell during the test period.

and parts of the Congo Basin) have poorer model performance, which is consistent with other global studies (Mo et al., 2022). A substantial improvement in the performance by the proposed model can be observed in most parts of India, eastern Europe and South America. Figure 6b, e and h depict the NSE values obtained from NTWSA, CTWSA and the BNML_TWSA, respectively. Similarly, Fig. 6c, f, and i present the spatial distribution of KGE values for NTWSA, CTWSA and BNML_TWSA across the globe. The results indicate the superior performance of BNML_TWSA compared to the other methods, which is evident from these plots at a global scale.

The grid-cell-wise CC, NSE, RMSE and KGE values are further compared for the three TWSA datasets (NTWSA, CTWSA and BNML_TWSA) in all river basins as box plots depicted in Fig. 8. The median values of each matrix are listed in Table 2. For most of the basins,

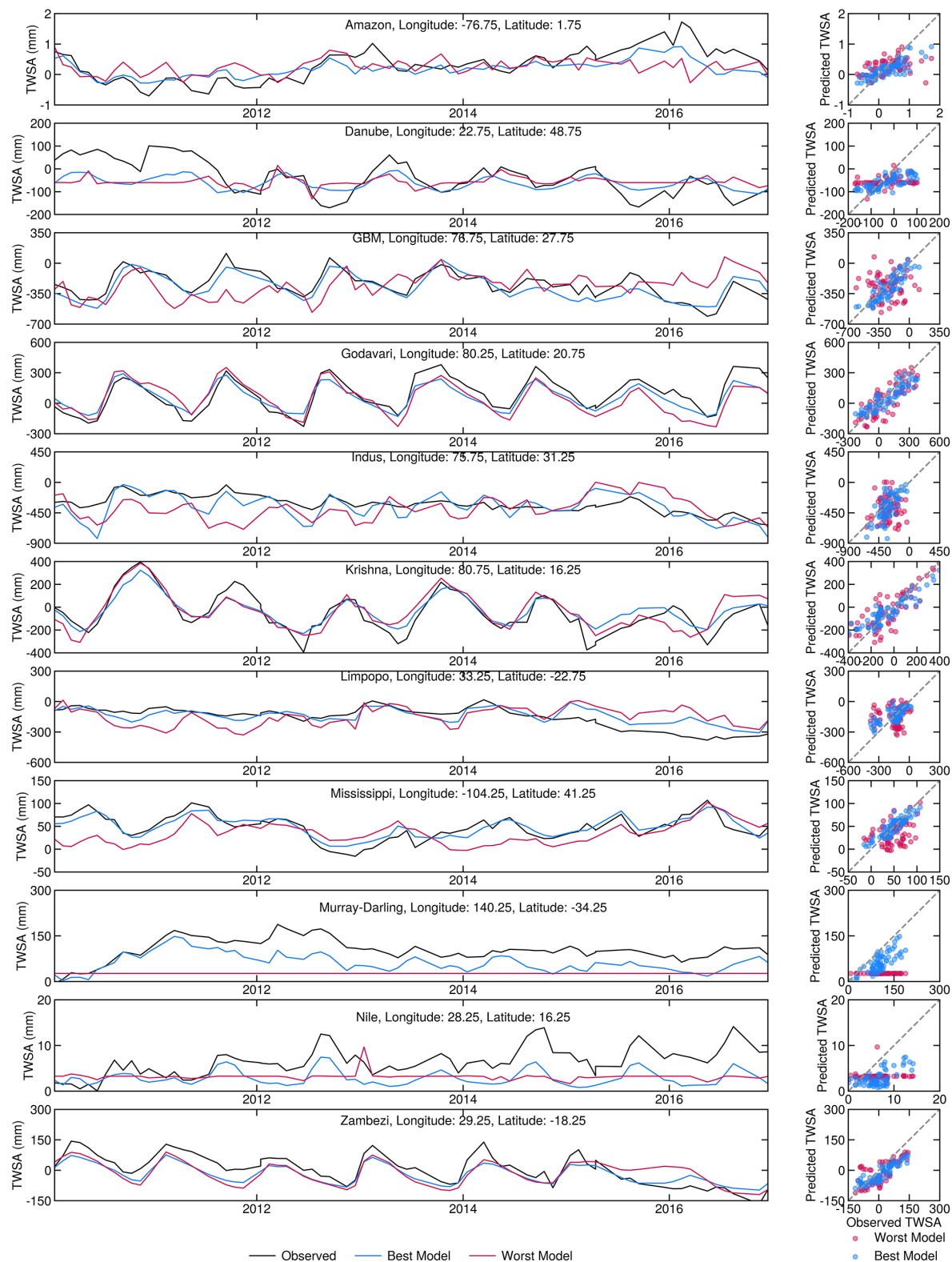


Figure 4. Time series (left columns) for grid cells showing the maximum improvement within the basins considered in this study, including observed TWSAs and TWSAs predicted by the best and worst models. Scatterplots (right columns) compare the TWSAs predicted by the best and worst models with the observed TWSAs.

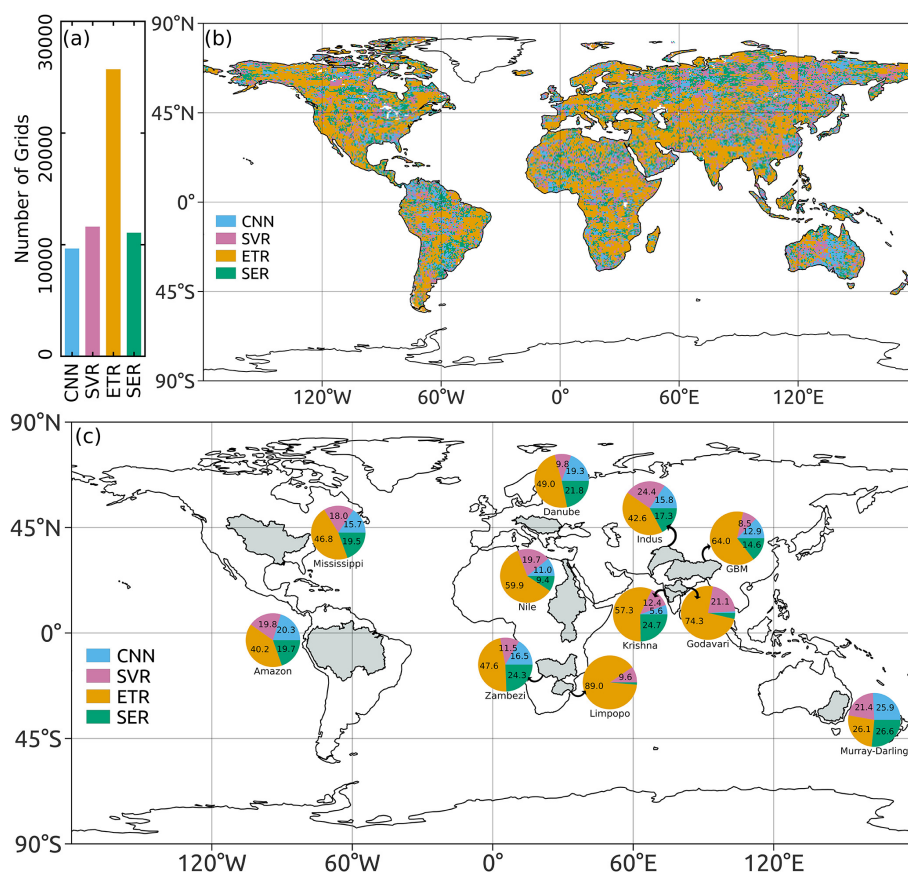


Figure 5. (a) Frequency, (b) spatial distribution of leader ML algorithms and (c) leader ML algorithms in terms of percentage for different river basins.

BNML_TWSA shows a higher median value of CC and a smaller range in the box plot compared to NTWSA and CTWSA, indicating a better performance for the proposed BNML_TWSA. However, for the Indus, Limpopo, Nile and Zambezi basins, CTWSA has a slightly better performance than BNML_TWSA. All of the grid cells in the Amazon, Danube, Godavari, Krishna, Mississippi and Zambezi basins have a CC value of ≥ 0.75 for BNML_TWSA. Similarly, for the GBM and Murray–Darling basins, the lowest CC value among all grid cells is above 0.5. The NSE, which compares the residual variance (the “noise”) with the variance in the measured data, shows poor results (< 0) for most of the basins for both NTWSA and CTWSA. On the other hand, the NSE values for BNML_TWSA are fairly high (> 0.68) in most basins, except for the Indus, Limpopo, Murray–Darling and Nile, and only the Indus Basin has a median NSE value of less than 0. Hence, based on the NSE value, BNML_TWSA has better performance than NTWSA and CTWSA for all of the river basins. The improved performance of BNML_TWSA can also be illustrated by the distribution of the RMSE values over each river basin, depicted in Fig. 8. When comparing the overall spread of the RMSE values for BNML_TWSA, CTWSA and NTWSA, it is ob-

served that the range of the RMSE values for BNML_TWSA is lower than that of CTWSA and NTWSA. This indicates that the interquartile range (IQR), the difference between the third quartile and the first quartile, of the RMSE values for BNML_TWSA is smaller than that of CTWSA and NTWSA. Similarly, the KGE values indicate that CTWSA exhibits the best performance in the Godavari Basin. For all other basins, BNML_TWSA shows superior performance.

4.4 Basin-scale quality assessment of gap-filled TWSAs

The results presented in the preceding sections demonstrate the superior ability of the proposed model to simulate GRACE TWSAs during the testing period. The leader model, constructed for each global grid cell, is utilized to generate a GRACE-like TWSA series from April 2002 to December 2022 using the input parameter set selected by the BNs for each grid cell. Figure 9 shows the time series of average reconstructed TWSAs for all grid cells within a river basin. TWSAs from GLDAS NOAH (NTWSA), GLDAS CLSM (CTWSA) and GRACE are also included in Fig. 9 for comparison. The seasonal variation in TWSAs is greatly captured by the proposed models and the other two LSM outputs. Sim-

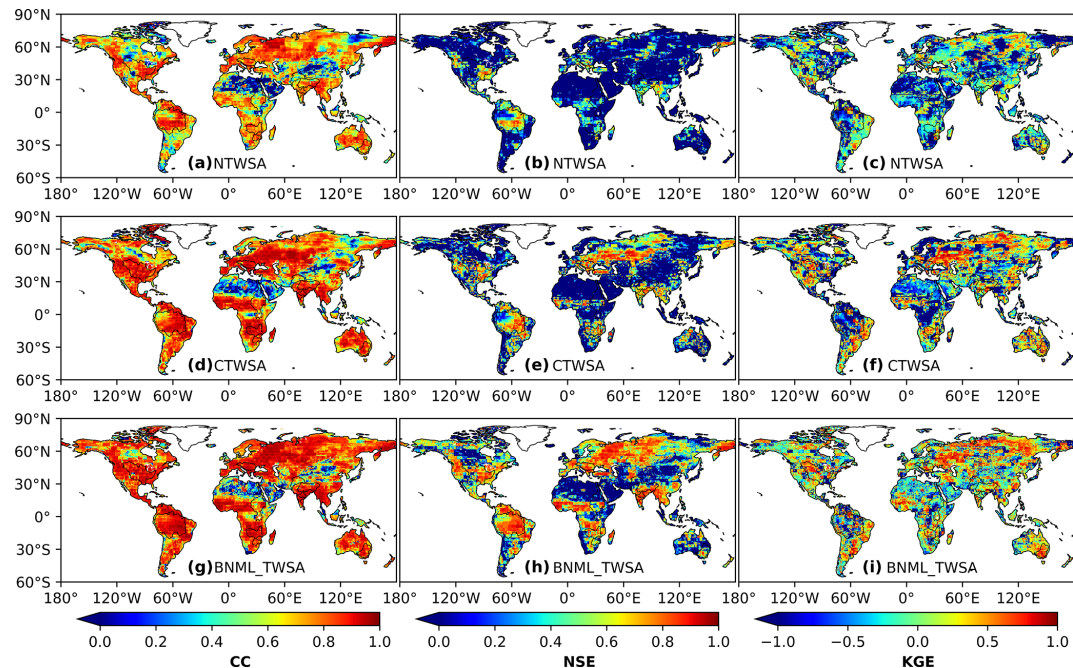


Figure 6. Correlation coefficient (CC), Nash–Sutcliffe efficiency (NSE) coefficient and Kling–Gupta efficiency (KGE) values between observed GRACE TWSAs and NTWSA, CTWSA and BNML_TWSA.

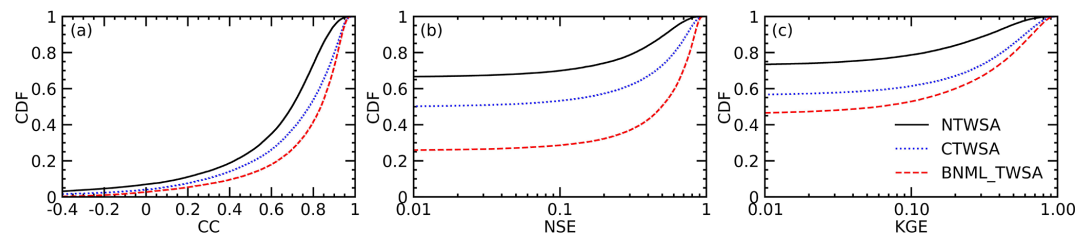


Figure 7. Cumulative distribution functions (CDFs) of the correlation coefficient (CC), Nash–Sutcliffe efficiency (NSE) coefficient and Kling–Gupta efficiency (KGE) values, as depicted in Fig. 6.

ilar to observations from previous sections, the CTWSA has better performance than the NTWSA. However, even better performance is achieved by the proposed BNML_TWSA. Compared to the observed GRACE TWSAs, both NTWSA and CTWSA underestimate the peak values of TWSAs for the Amazon Basin, whereas the BNML_TWSA time series matches the GRACE TWSAs most closely. For the Indus and Nile river basins, the mismatch between GRACE TWSAs and TWSAs from LSM outputs has been widening since 2010; however, the proposed model simulates BNML_TWSA, which is very close to the GRACE TWSAs.

The basin-wise performance of the mean BNML_TWSA and CTWSA is assessed against the GRACE TWSAs and depicted as a scatterplot in Fig. 10. NTWSA is not included in this plot, as it has already been found to be a poorer match with GRACE TWSAs than the CTWSA. The CC of BNML_TWSA in the Zambezi Basin is the highest among all of the basins (0.989); however, the RMSE (46.1 mm) and

NSE (0.833) values do not suggest that its performance is the best. The basin-wide mean BNML_TWSA has a higher CC (≥ 0.904) than CTWSA for the GBM, Godavari, Indus, Krishna, Limpopo, Mississippi, Murray–Darling and Zambezi basins (see Fig. 10). In the Nile Basin, a high CC value of 0.82 is obtained for BNML_TWSA; however, this is the minimum among all river basins. The Nile Basin is one of the few basins where CTWSA has a slightly higher CC (0.889) than BNML_TWSA (see Fig. 10), but the time-series plot (see Fig. 9) shows that the BNML_TWSA is in better agreement with GRACE TWSAs. This indicates that, in general, the proposed BNML_TWSA is more reliable than CTWSA. Considering the NSE and RMSE values, the Amazon, GBM, Godavari, Krishna and Nile basins perform better with BNML_TWSA than with CTWSA. The only exception is Danube Basin which has a marginally lower performance for BNML_TWSA than CTWSA according to the basin-wide mean. However, when considering

Table 2. Median of the CC, NSE, RMSE and KGE values at grid cells of each basin. A bold value signifies the best performance.

Model		Amazon	Danube	GBM	Godavari	Indus	Krishna	Limpopo	Mississippi	Murray–Darling	Nile	Zambezi
CC	NTWSA	0.822	0.806	0.766	0.779	0.566	0.740	0.800	0.786	0.788	0.538	0.770
	CTWSA	0.876	0.917	0.861	0.898	0.654	0.895	0.839	0.903	0.853	0.739	0.925
	BNML_TWSA	0.910	0.918	0.879	0.927	0.651	0.924	0.837	0.907	0.865	0.675	0.923
NSE	NTWSA	0.537	0.433	0.266	0.475	−1.140	0.424	−0.897	0.105	−0.897	−1.373	0.159
	CTWSA	0.608	0.700	0.452	0.671	−0.392	0.636	0.351	0.539	−0.060	−1.570	0.651
	BNML_TWSA	0.796	0.766	0.699	0.839	−0.003	0.810	0.581	0.704	0.455	0.200	0.688
RMSE	NTWSA	119.5	56.0	123.0	114.4	51.7	102.3	54.4	61.4	57.3	61.6	123.3
	CTWSA	107.8	42.0	100.7	86.0	46.4	81.3	34.0	50.1	43.6	55.1	75.1
	BNML_TWSA	83.3	35.9	74.5	63.7	37.1	63.0	30.5	39.9	35.0	31.7	70.0
KGE	NTWSA	−0.509	−0.119	−0.056	−0.168	−0.716	0.100	−1.351	−0.213	0.175	−0.648	−0.174
	CTWSA	−0.686	0.062	0.146	0.108	−0.355	−0.164	−0.277	0.223	0.402	−0.579	0.385
	BNML_TWSA	−0.194	0.224	0.182	0	−0.143	0.383	−0.238	0.346	0.592	−0.186	0.255

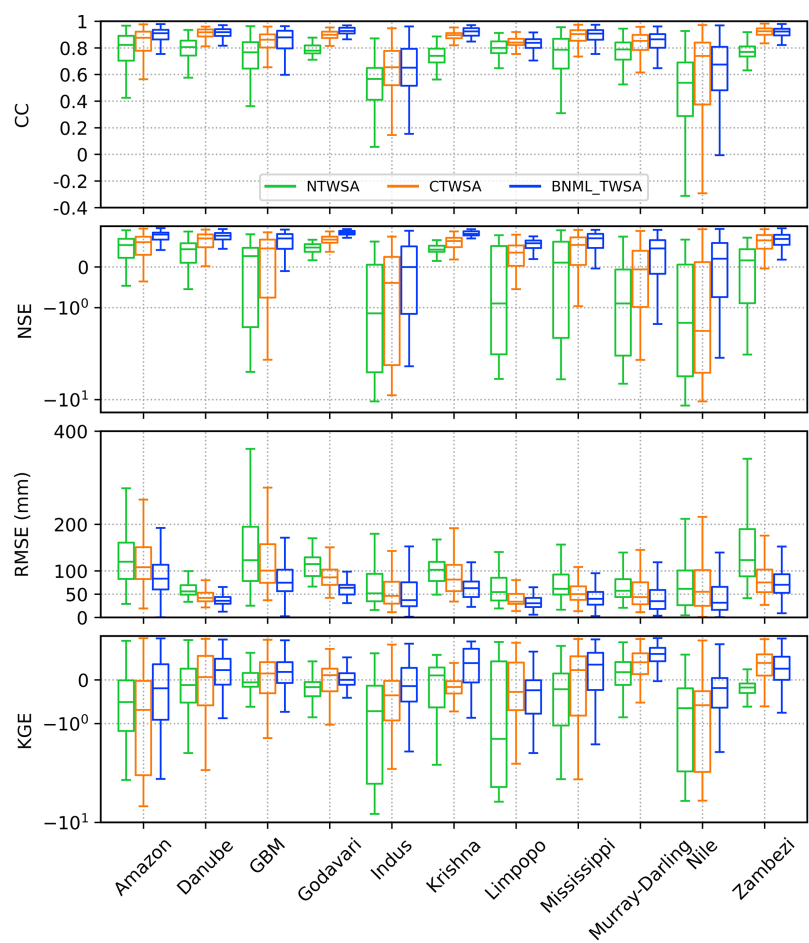


Figure 8. Box plot of CC, NSE, RMSE and KGE values for the grid cells of each basin, excluding the outliers.

the KGE values, BNML_TWSA outperforms CTWSA in most basins. With the exception of the Danube, Indus, Krishna, Murray–Darling and Zambezi basins, all other basins demonstrate superior performance for BNML_TWSA compared to CTWSA when evaluated using the KGE values.

4.5 Reflection of hydroclimatic extreme events in BNML_TWSA during the pre-GRACE (1960–2002) and intermediate gap periods

The leader ML model of each grid cell is used to hindcast BNML_TWSA during the pre-GRACE period (January 1960–March 2002) using the predictors selected by BNs. The basin-averaged BNML_TWSA series is shown in Fig. 11. Similar to the LSM outputs (CTWSA and NTWSA), increasing and decreasing trends are noted for BNML_TWSA series for some of the river basins. The variability in the TWSA hindcast is lowest for the Indus (−95.1 to 75.4 mm) and Limpopo (−61.6 to 83.7 mm) river basins, while the Amazon (−227.5 to 199.6 mm) and Krishna (−200.7 to 239.5 mm) basins have the highest variability. The minimum TWSA hindcast values for the Murray–Darling and Nile basins are around −20 mm – the lowest among all of the regions. During this hindcast period, a number of significant extreme climate events occurred; for example, two severe flood events in India are depicted in Fig. 12a–b. First, the Gomti River, a tributary of the Ganges River, overflowed and caused a severe flood that inundated half of the city of Lucknow in mid-October 1960. As shown in Fig. 12a, water storage increased around Lucknow in October 1960, which is marked by a green rectangle. Second, on 11 August 1979, a flood disaster occurred in Gujarat, India, when the Machchhu dam failed and submerged the town of Morbi, killing about 1500 people (Saharia et al., 2021). The deviation of BNML_TWSA from the long-term monthly mean for the Morbi region in Gujarat, which is marked by a green rectangle, is shown in Fig. 12b. Results clearly indicate the enhancement of TWS during this period. The proposed model simulated the TWSA series during the pre-GRACE period and identified recorded extreme climate events that occurred in the hindcast period.

During the GRACE gap period, several extreme climate events occurred around the world. In the continental USA, one of these extreme events was Hurricane Harvey, which made landfall on 25 August 2017 along Texas and Louisiana coast. This catastrophic flood event caused damage totaling USD 125 billion (Sun et al., 2021; United States National Hurricane Center, 2018). The flood event is expected to reveal an increase in the TWSA compared to the long-term mean TWSA of that region. This is well reflected in Fig. 12c, which depicts the difference between BNML_TWSA for September 2017 (as the event occurred towards end August 2017) and the long-term mean TWSA for September. Similarly, heavy rain on 15–16 July 2017 led to flooding in several districts of Gujarat, India, and the event reportedly

caused more than 200 deaths. This is depicted with a similar plot in Fig. 12d. The proposed model effectively reflects the impact of extreme climate events on TWS in the Texas and Louisiana coastal area and in Gujarat.

4.6 Comparison with previous studies

In this section, the reconstructed BNML_TWSA is compared with the global TWSA products developed recently by Humphrey and Gudmundsson (2019) and Sun et al. (2020). The reconstructed TWSA datasets developed using statistical models by Humphrey and Gudmundsson (2019) are named GRACE-REC. Two GRACE-REC datasets, which include monthly ensemble mean data from the JPL_MSWEF and JPL_ERA5 datasets, respectively, and the BNML_TWSA are each evaluated against the GRACE JPL mascon dataset. The grid intersection points and resolution ($0.50^\circ \times 0.50^\circ$) of BNML_TWSA, GRACE-REC and GRACE JPL mascon are uniform, which eliminates the requirement for re-gridding. The period of comparison is selected as the common available dataset duration of the above three products. Specifically, the period spans from April 2002 to July 2019 for JPL_ERA5 and from April 2002 to December 2016 for JPL_MSWEF. Figure 13 depicts the spatial distribution of the CC values obtained from the two GRACE-REC datasets of Humphrey and Gudmundsson (2019) along with corresponding CC values of the BNML_TWSA, each compared with the GRACE JPL mascon datasets. Figures 14 and 15 depict the NSE and KGE values, respectively, similar to Fig. 13. Based on Figs. 14 and 13, it is evident that the performance of BNML_TWSA surpasses that of GRACE-REC TWSAs. This superior performance of BNML_TWSA is even more apparent in Fig. 15. Notably, both of the GRACE-REC products exhibited suboptimal performance in the region near the Sahara Desert and Saudi Arabia.

Next, we compare the agreement of BNML_TWSA with the observed GRACE JPL mascon vs. the agreement of TWSAs derived from deep neural network (DNN) models (namely, DNN_JPL-M and DNN_CSR-M for the period from April 2002 to July 2018) by Sun et al. (2020) with the same GRACE JPL mascon. During the development of the DNN_JPL-M and DNN_CSR-M TWSA products in Sun et al. (2020), the JPL mascon and CSR mascon GRACE products are respectively used as targets. These two particular products of Sun et al. (2020) are selected for comparison as the study mentions that TWSAs derived using DNN models demonstrated superior performance compared to the other two learning-based models attempted in their study. The spatial resolution of DNN_JPL-M and DNN_CSR-M is $1.0^\circ \times 1.0^\circ$, whereas the spatial resolution of BNML_TWSA and JPL mascon is $0.50^\circ \times 0.50^\circ$. To ensure a uniform spatial resolution for all TWSA products, both BNML_TWSA and JPL mascon are regridded (upscaled) to $1.0^\circ \times 1.0^\circ$, similar to DNN_JPL-M and DNN_CSR-M. Figure 16a, b and c depict the respective CC, NSE and KGE values for DNN_JPL-

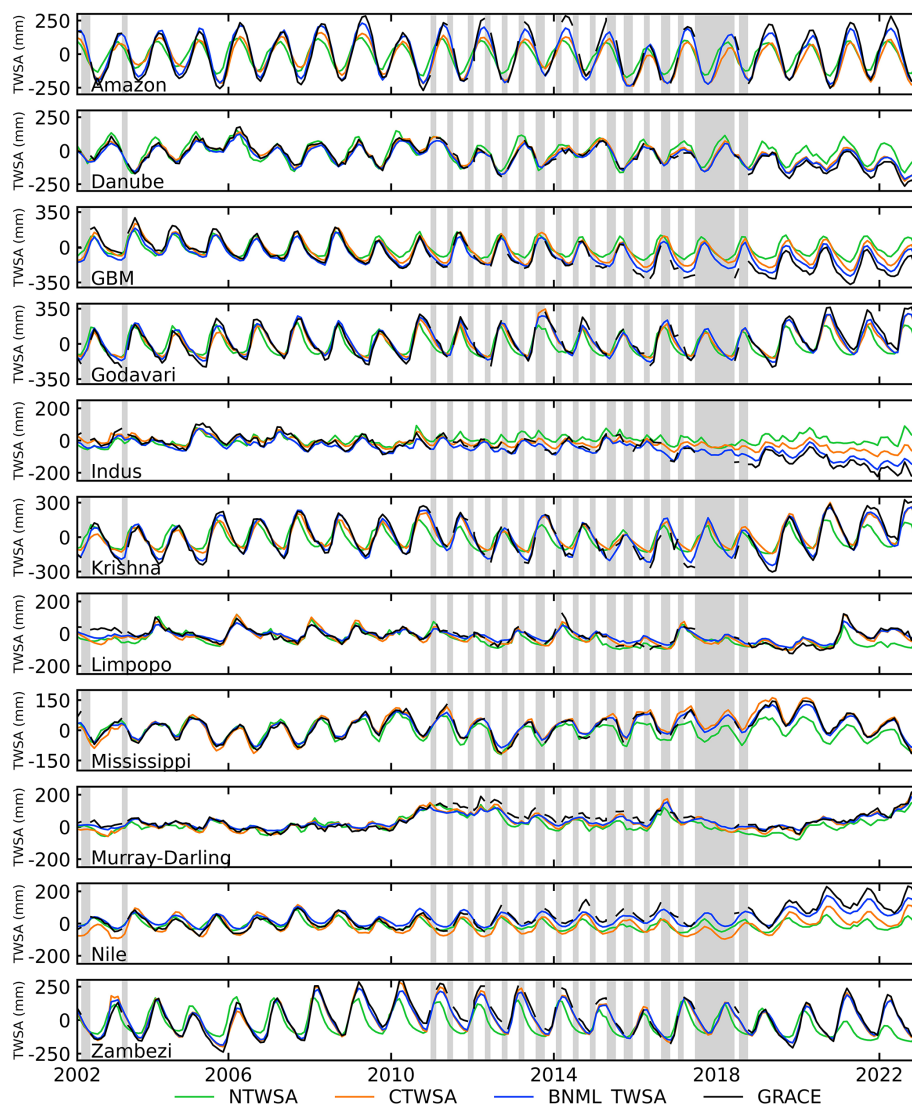


Figure 9. Comparison of TWSA time series from April 2002 to December 2022 (GRACE period). Vertical gray bars indicate missing GRACE observations.

M, and corresponding indices for DNN_CSR-M are shown in Fig. 16d, e and f, respectively. The CC, NSE and KGE values for BNML_TWSA are depicted in Fig. 16g, h and i, respectively. The prediction accuracy of BNML_TWSA is the best when compared to DNN_JPL-M and DNN_CSR-M TWSA.

The basin-wise median CC, NSE and KGE values of BNML_TWSA, the reconstructed TWSAs from Humphrey and Gudmundsson (2019), and TWSAs from Sun et al. (2020) are compared in Fig. 17 using a radar chart. Specifically, for Humphrey JPL_ERA5, the CC values are shown in Fig. 17a, the NSE values in Fig. 17d and the KGE values in Fig. 17g. Similarly, for Humphrey JPL_MSWEF, the CC values appear in Fig. 17b, the NSE values in Fig. 17e and the KGE values in Fig. 17h. The CC, NSE and KGE values of Sun DNN_JPL-M and Sun DNN_CSR-M, which have

a similar analysis period, are presented in Fig. 17c, f and i, respectively. Thus, Fig. 17a–c depict the basin-wise median CC values for the mentioned models, while Fig. 17d–f illustrate the basin-wise median NSE values after excluding those below 0. Similarly, Fig. 17g–i display the basin-wise median KGE values, excluding those below -1 . With the exception of the Danube, GBM, Indus, and Nile river basins, the median CC values for BNML_TWSA and Humphrey JPL_ERA5 (as shown in Fig. 17a) exhibit similarity. BNML_TWSA demonstrates higher median CC values compared to Humphrey JPL_ERA5 across the aforementioned basins. Specifically, the improvements in the CC values obtained with BNML_TWSA are as follows: from 0.86 to 0.90 for the Danube, from 0.80 to 0.90 for GBM, from 0.59 to 0.77 for the Indus and from 0.70 to 0.77 for the Nile. Similarly, compared to Humphrey JPL_MSWEF (Fig. 17b),

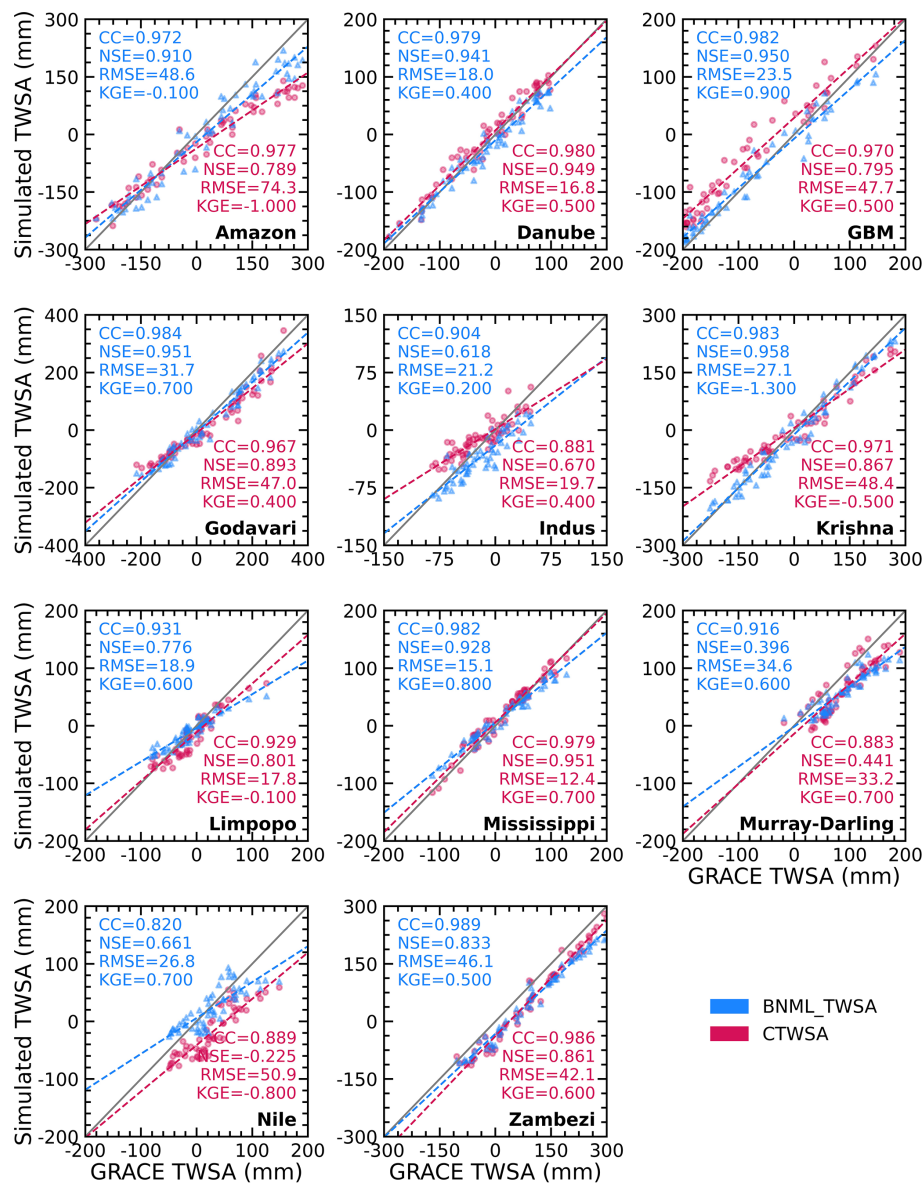


Figure 10. Scatterplot of the basin-wise mean of observed GRACE TWSA vs. BNML_TWSA and CTWSA for each river basin.

BNML_TWSA provides CC improvements, such as from 0.75 to 0.9 for the GBM region and from 0.47 to 0.74 for the Indus region. When compared with the reconstructed TWSA products by Sun et al. (2020), BNML_TWSA surpasses their accuracy, as the median CC values are high across all of the river basins (Fig. 17c). In terms of the median NSE values, BNML_TWSA outperforms Humphrey JPL_ERA5 in all of the river basins. Specifically, for BNML_TWSA, the median NSE value improves from -1.07 to 0.89 for the Murray–Darling Basin and from -0.17 to 0.77 for the Nile Basin. Similarly, BNML_TWSA exhibits improved performance compared to Humphrey JPL_MSWEF across all of the basins, particularly for the Murray–Darling Basin, where the median NSE improves from -0.24 to 0.72 . While

Sun DNN_JPL-M performs better than Sun DNN_CSR-M, BNML_TWSA consistently outperforms both of the aforementioned products of Sun et al. (2020) across all of the river basins. Notably, for the Nile Basin, the median NSE value improves from 0 to 0.53 with BNML_TWSA compared to Sun DNN_JPL-M. In terms of the median KGE values, BNML_TWSA also outperforms Humphrey JPL_ERA5 in all of the river basins. Similarly, except for the Limpopo and Indus basins, BNML_TWSA exhibits superior performance compared to Humphrey JPL_MSWEF in all other basins. Although Sun DNN_JPL-M performs better than Sun DNN_CSR-M, BNML_TWSA consistently outperforms both products, except in the Danube Basin, where

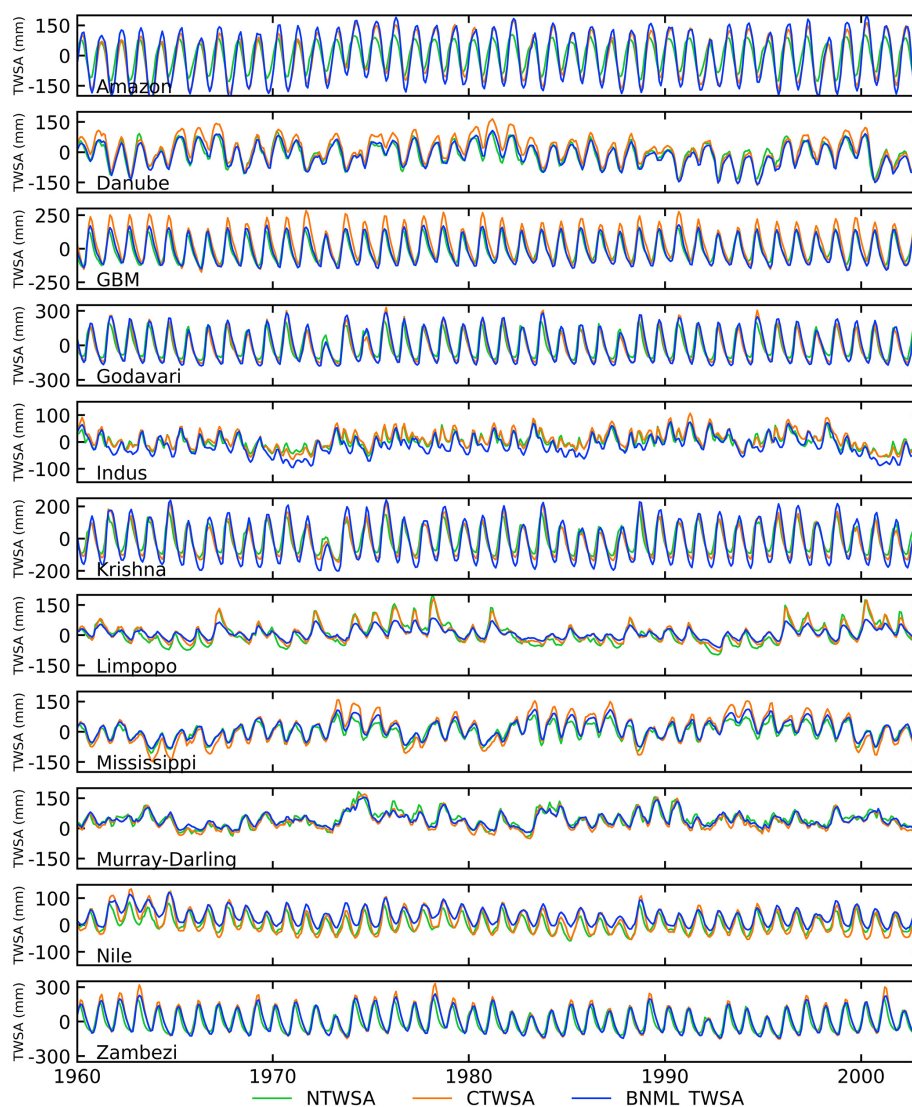


Figure 11. BNML_TWSA during the pre-GRACE period (January 1960–March 2002).

Sun DNN_JPL-M has a marginally higher KGE value than BNML_TWSA.

4.7 Comparison with streamflow measurements based on the basin-scale water balance

TWS change can be used to estimate streamflow measurement based on the water balance equation for moderately large ($> 100\,000\text{ km}^2$) river basins (Humphrey and Gudmundsson, 2019). The streamflow (Q) based on the water balance model over a watershed may be expressed as follows:

$$Q = P - ET - \Delta S, \quad (8)$$

where the water balance components P and ET are precipitation and evapotranspiration, respectively, and ΔS denotes

the TWS change over a time step. Comparison of evaluated streamflow using BNML_TWSA, GRACE TWSA, TWSA from Humphrey and Gudmundsson (2019) (JPL_MSWEF and JPL_ERA5), and Sun et al. (2020) (DNN_JPL-M and DNN_CSR-M) is discussed in this section. Out of the 11 river basins considered in this study, 6 basins – 1 from each continent – were selected based on the availability of streamflow data. More details on the six selected basins and their streamflow observation stations are given in Table 3. Streamflow observations are predominantly acquired from the Global Runoff Data Centre (GRDC), except for the Godavari River in India, for which the streamflow data are sourced from the Central Water Commission (CWC).

Observations of terrestrial water balance components for large river basins worldwide are limited, with sparsely distributed gauges for precipitation and even fewer observations

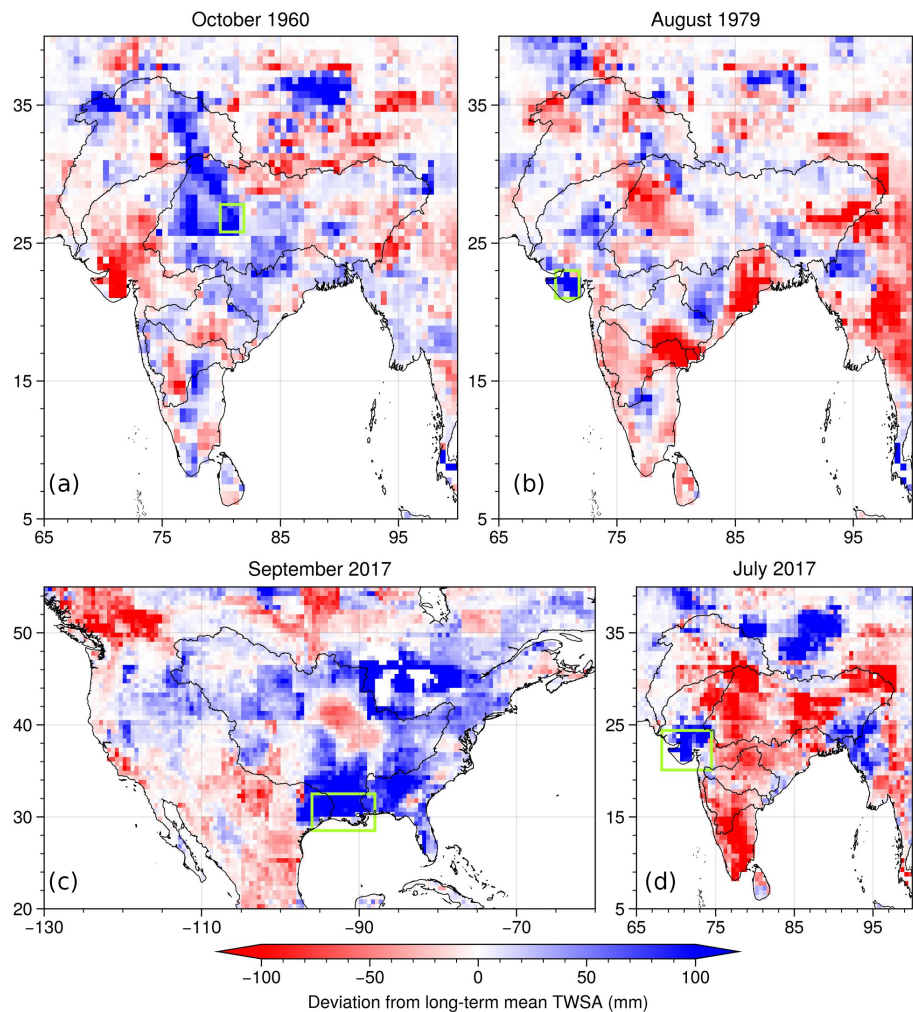


Figure 12. Difference between monthly and long-term mean monthly BNML_TWSA reflecting extreme hydroclimatic events. The zone of interest is marked by a green rectangle.

Table 3. Details of basin and streamflow observation locations for six global river basins. Data were sourced from the Global Runoff Data Centre (GRDC; <https://portal.grdc.bafg.de>, last access: 12 November 2024) and Central Water Commission (CWC; <https://indiawris.gov.in>, last access: 7 November 2023), India.

River basin	Source	Station for streamflow observation	Period of streamflow observation	Drainage area (km ²)
Amazon	GRDC	Óbidos	April 2002–December 2019	4 671 462
Danube	GRDC	Ceatal Izmail	April 2002–December 2010	779 812
Godavari	CWC	Polavaram	January 2003–December 2020	312 812
Mississippi	GRDC	Vicksburg	April 2002–October 2022	2 918 820
Murray–Darling	GRDC	Lock 1 downstream	April 2002–June 2023	770 171
Zambezi	GRDC	Katima Mulilo	April 2002–July 2021	334 883

for evapotranspiration. However, due to the availability of data from satellite sensors and outputs from global land surface models, it is possible to analyze the water balance of river basins with sparse observations. Details of the collected dataset and sources are presented in Table 4. Precipitation (P) data from five different sources were collected for each

grid cell within these river basins. The basin-scale average of all five precipitation products (GLDAS, GPCC, GPCP, IMERG and PERSIANN) is considered to be the “observed” precipitation for that particular basin. Similarly, for evapotranspiration (ET), the average of three products (GLDAS, FLDAS and GLEAM) is considered the observed ET for that

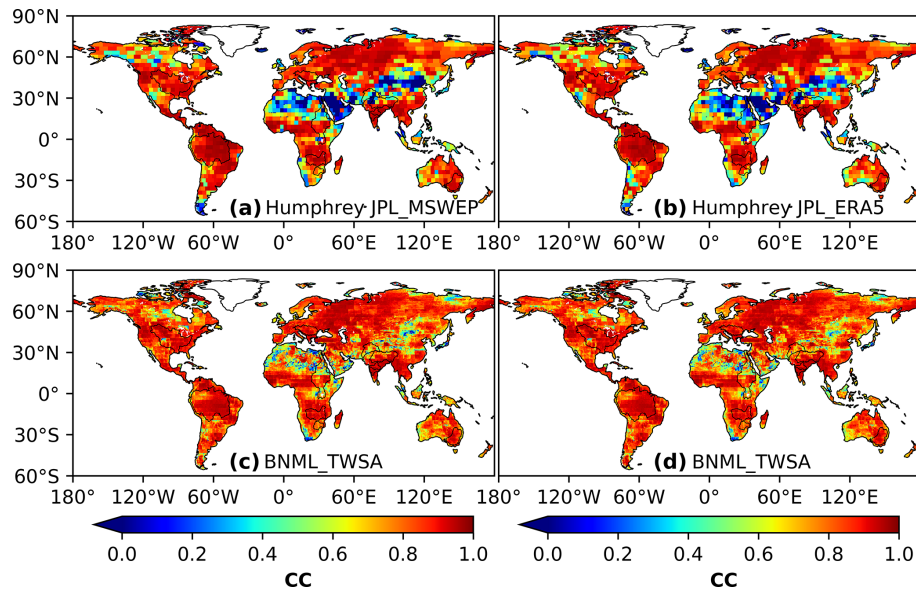


Figure 13. Comparison of correlation coefficient (CC) values obtained by two GRACE-REC products from Humphrey and Gudmundsson (2019) and BNML_TWSA, each evaluated against GRACE JPL mascon.

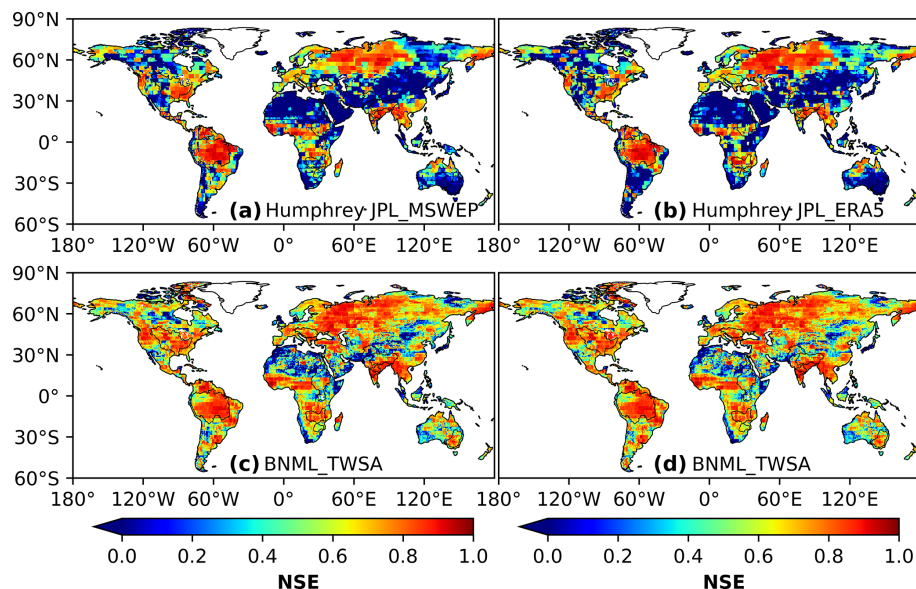


Figure 14. Comparison of Nash–Sutcliffe efficiency (NSE) values obtained by two GRACE-REC products from Humphrey and Gudmundsson (2019) and BNML_TWSA, each evaluated against GRACE JPL mascon.

basin. In this study, ΔS for the t th month is calculated as the central difference in the TWSAs, as shown below.

$$\Delta S = \frac{(TWSA_{t+1} - TWSA_{t-1})}{2} \quad (9)$$

Using the water balance components described in the previous section, the streamflow for each basin is calculated using various TWSA products, including GRACE, BNML_TWSA, JPL_MSWEp, JPL_ERA5, DNN_JPL-M

and DNN_CSR-M. This computation is performed based on the terrestrial water balance equation (Eq. 8). The computed Q values are compared with the observed Q values from the station, and the corresponding correlation coefficients (CC) are determined. Figure 18 presents the correlation coefficient (CC) values as a heatmap for all six river basins, highlighting the performance of BNML_TWSA. For the Amazon Basin, BNML_TWSA demonstrates strong performance, with a CC of 0.89, comparable to GRACE and JPL_MSWEp (CC: 0.9) and JPL_ERA5 (CC: 0.89). In the

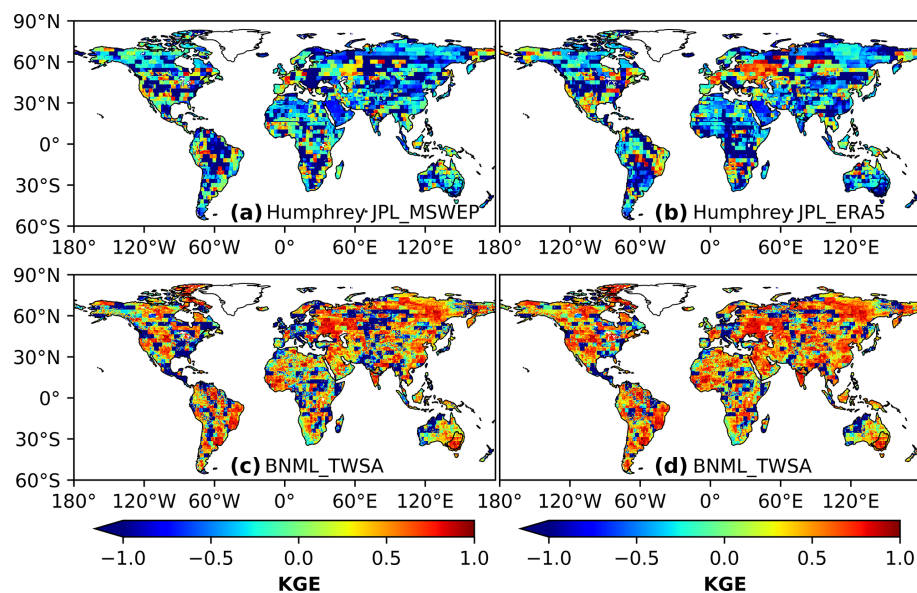


Figure 15. Comparison of Kling–Gupta efficiency (KGE) values obtained by two GRACE-REC products from Humphrey and Gudmundsson (2019) and BNML_TWSA, each evaluated against GRACE JPL mascon.

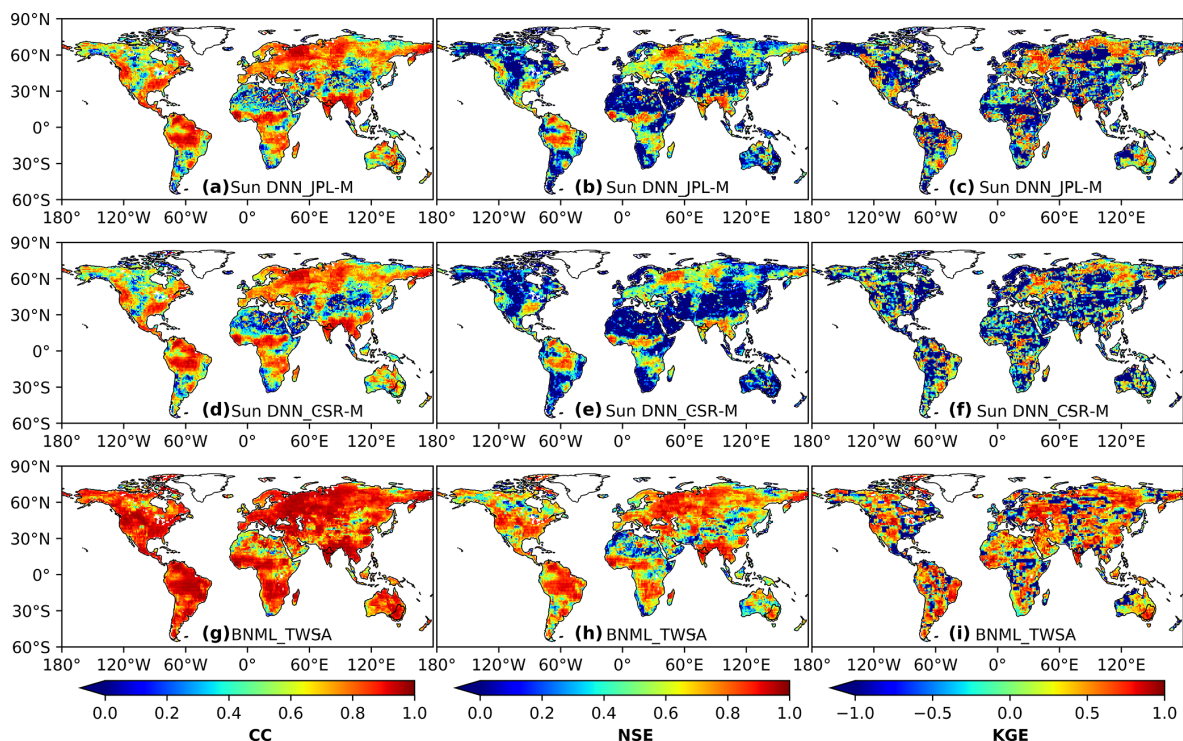


Figure 16. Comparison of correlation coefficient (CC), Nash–Sutcliffe efficiency (NSE) and Kling–Gupta efficiency (KGE) values obtained by two reconstructed TWSA products from Sun et al. (2020) and BNML_TWSA, each evaluated against GRACE JPL mascon.

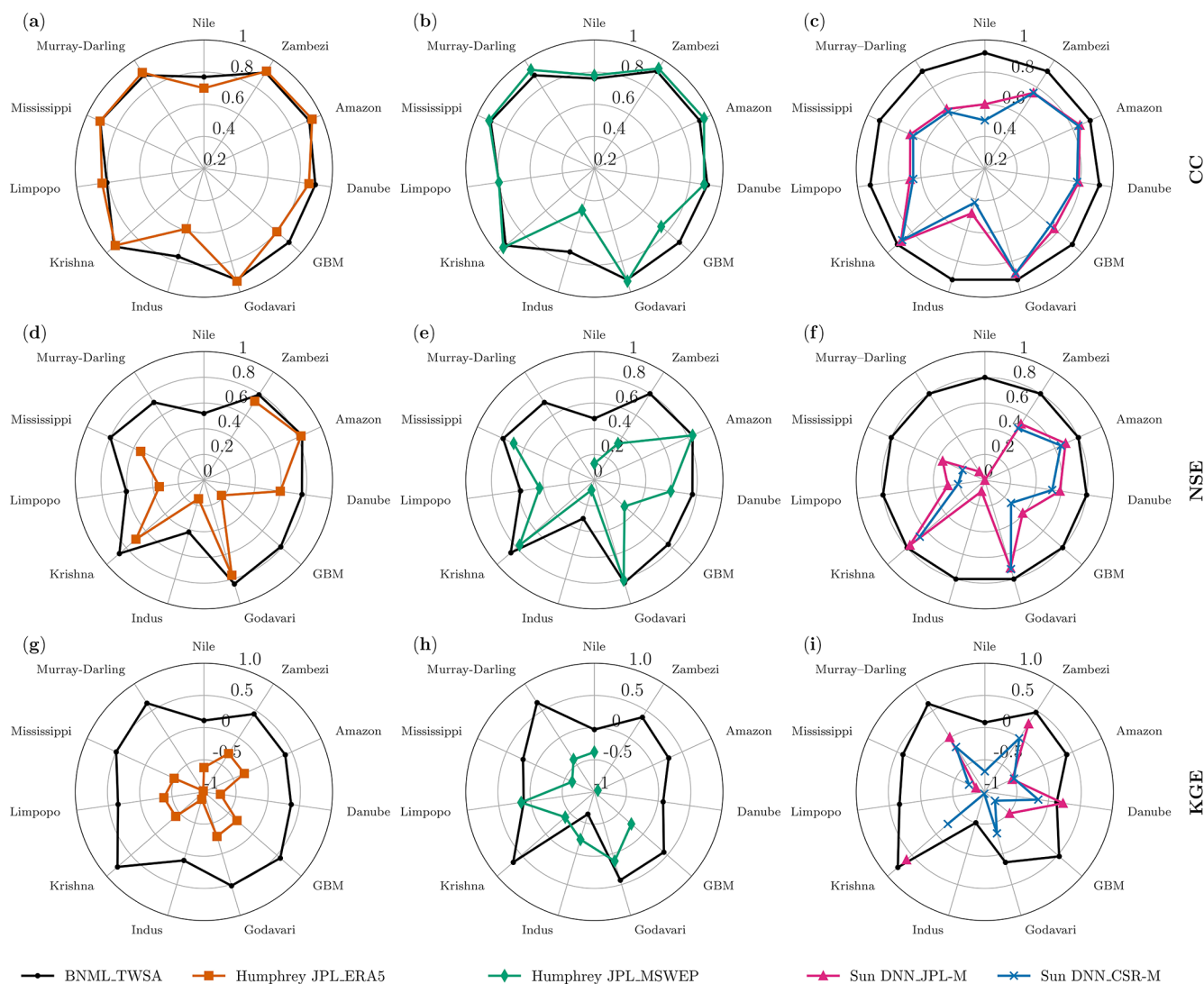


Figure 17. Comparison of the median values of the correlation coefficient (CC), Nash–Sutcliffe efficiency (NSE) and Kling–Gupta efficiency (KGE) across different basins. Note that median NSE values below 0 and KGE values below -1 are omitted during plotting.

Danube and Godavari basins, BNML_TWSA outperforms all other TWSA products, achieving the highest CC values, although other products also perform well. For the Mississippi Basin, BNML_TWSA, along with DNN_JPL-M and DNN_CSR-M, achieves the highest CC value of 0.7. For the Murray–Darling Basin, all TWSA products show minimal CC values due to the negligible magnitude of observed streamflow at the basin outlet. For the Zambezi Basin, JPL_MSWEF performs best, with a CC of 0.46, whereas BNML_TWSA achieves a CC value of 0.35. This evaluation highlights the superior and/or comparable performance of BNML_TWSA across most basins. The time series of the streamflow computed using BNML_TWSA ($Q_{\text{BNML_TWSA}}$) is presented alongside the observed streamflow (Q_{Observed}) and the streamflow computed using GRACE TWSA (Q_{GRACE}) in the left columns of Fig. 19. The time-

series plot (Fig. 19) clearly demonstrates that $Q_{\text{BNML_TWSA}}$ aligns more closely with Q_{Observed} compared to Q_{GRACE} . Additionally, the RMSE has been computed for Q_{GRACE} and $Q_{\text{BNML_TWSA}}$ against Q_{Observed} . These RMSE values clearly indicate the superior performance of $Q_{\text{BNML_TWSA}}$ for all six basins compared to Q_{GRACE} . For the Murray–Darling Basin, the magnitude of Q_{Observed} is negligible due to the large amount of water withdrawal for irrigation and consumption, in addition to heavy regulation (Candogan Yossef et al., 2012). All TWSA products struggle to capture the pattern of low streamflow in the Murray–Darling Basin.

4.8 Uncertainty, limitations and future scope

There are various sources that contribute to the uncertainties in reconstructed TWSAs. The primary source of uncertain-

Table 4. Overview of global precipitation, evapotranspiration and storage change data products utilized for streamflow calculations.

Dataset	Spatial resolution	Temporal resolution	Reference and data source
Precipitation (<i>P</i>)			
GLDAS	0.25°	1 month	Rodell et al. (2004), https://doi.org/10.5067/SXAVCZFAQLNO
GPCC	0.25°	1 month	Schneider et al. (2008); Rustemeier et al. (2022), https://doi.org/10.5676/DWD_GPCC/CLIM_M_V2022_025
GPCP	0.5°	1 month	https://doi.org/10.5067/MEASURES/GPCP/DATA304 (Huffman et al., 2022)
IMERG	0.1°	1 month	https://doi.org/10.5067/GPM/IMERG/3B-MONTH/07 (Huffman et al., 2023)
PERSIANN	0.25°	1 month	Ashouri et al. (2015), https://www.ncei.noaa.gov/data/precipitation-persiann/access/
Evapotranspiration (ET)			
GLDAS	0.25°	1 month	Rodell et al. (2004), https://doi.org/10.5067/SXAVCZFAQLNO
FLDAS	0.1°	1 month	https://doi.org/10.5067/5NHC22T9375G (McNally, 2018)
GLEAM	0.25°	1 month	Martens et al. (2017), Miralles et al. (2011), https://www.gleam.eu
Storage change (ΔS)			
GRACE (JPL mascon)	0.5°	1 month	Watkins et al. (2015), https://doi.org/10.5067/TEMSC-3JC63
BNML_TWSA	0.5°	1 month	Mandal et al. (2024), https://doi.org/10.6084/m9.figshare.25376695
JPL_MSWEP	0.5°	1 month	Humphrey and Gudmundsson (2019)
JPL_ERA5	0.5°	1 month	Humphrey and Gudmundsson (2019)
DNN_JPL-M	1°	1 month	Sun et al. (2020)
DNN_CSR-M	1°	1 month	Sun et al. (2020)



Figure 18. Basin-wise CC values obtained against observed *Q* and computed *Q* from water balance using TWSA data from GRACE, BNML_TWSA and other studies.

ties arises from the measurement errors, inherent processing errors, leakage errors and model assumptions associated with the original GRACE data, as documented by Boergens et al. (2022) and Gao et al. (2023). Nevertheless, this issue is effectively mitigated by utilizing the mascon solution, which demonstrates clear superiority over the spherical harmonics data (Kalu et al., 2024). The JPL mascon solution employs a coastline resolution improvement (CRI) filter to minimize leakage errors across land–ocean boundaries. Additionally,

gain factors are utilized to further mitigate these leakage errors. Moreover, a Bayesian framework is implemented to more effectively eliminate correlated errors compared to traditional empirical filters (Wiese et al., 2016). Another source of uncertainty stems from the ML models, which may be categorized into contributions from inadequacies and/or lack of knowledge regarding the model (epistemic) and data noise (aleatoric). In the present study, epistemic uncertainty has been reduced to some extent by training four different ML models at each grid cell and selecting the best-performing model to reconstruct the BNML_TWSA globally. On the other hand, aleatoric uncertainty may arise from the input dataset (i.e., the selected predictors). Analyzing the spatial distribution of selected predictors using BNs (Fig. 2), it becomes apparent that commonly employed forcing variables, such as precipitation (*P*) and temperature (*T*), do not rank among the top predictors in most grid cells. This observation suggests that these forcing variables are already accounted for in the LSMs, as indicated by Sun et al. (2019). However, these variables are still selected as optimal predictors in some of the grid cells, which implies that physics-based LSMs may not entirely capture the total information encapsulated in the raw data. Consequently, incorporating a diverse set of variables – including those already utilized in physics-based LSMs – as potential predictors could mitigate model structural errors and parameter uncertainties inherent in the LSMs (Sun et al., 2020, 2019). Furthermore, uncertainties may also depend on the actual source of the input variables.

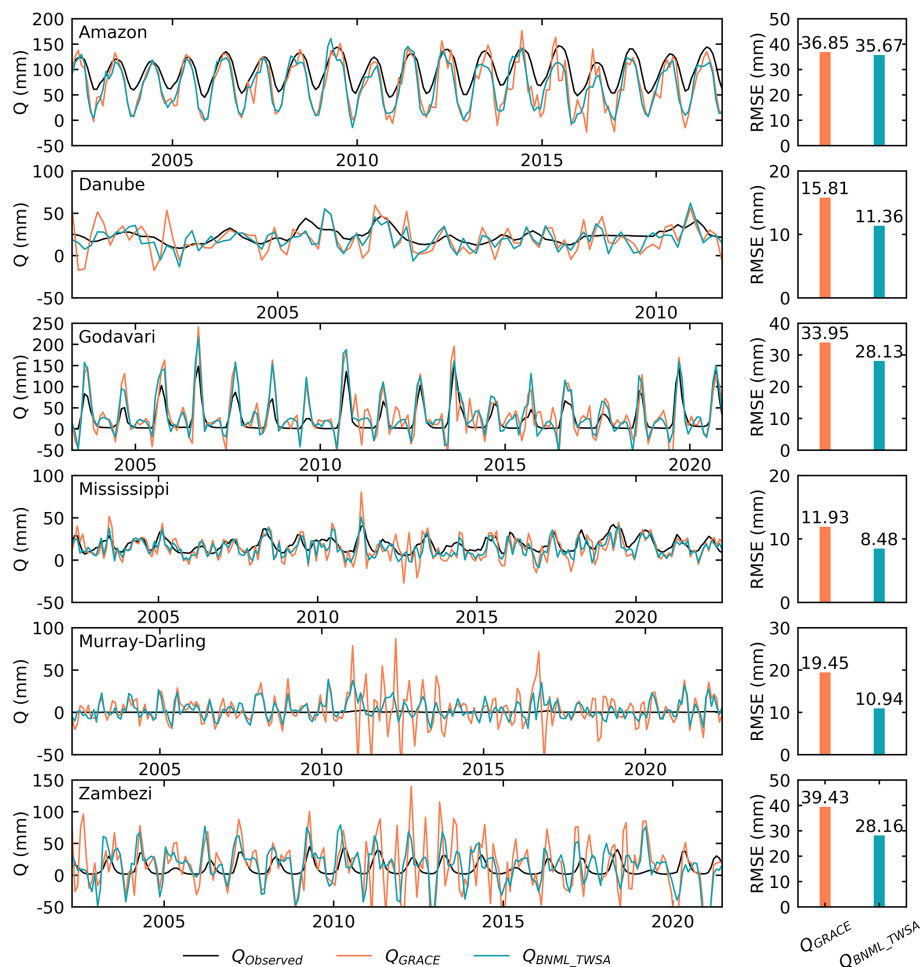


Figure 19. Comparison of observed streamflow ($Q_{Observed}$), Q obtained from water balance using GRACE TWS data (Q_{GRACE}) and Q obtained from water balance using BNML_TWSA TWS data (Q_{BNML_TWSA}). RMSE values (right columns) obtained for Q_{GRACE} and Q_{BNML_TWSA} against $Q_{Observed}$.

For example, precipitation from satellite sources will entail different uncertainties compared to LSM-based precipitation. In the present study, aleatoric uncertainty may arise from the absence of variables that capture the impact of anthropogenic activities. As we utilized variables from LSMs and climate indices as inputs to the ML models for reconstructing BNML_TWSA, the influence of anthropogenic activities is not adequately represented by these variables.

In this study, a model uncertainty assessment is performed for the reconstructed dataset during the model training phase, using the GRACE observations. The uncertainty in the model predictions is quantified by calculating confidence intervals (CIs) of the TWSA estimates. The CI is defined as the point estimate \pm the margin of error, where the margin of error is determined by the product of a confidence coefficient ($C_{confidence}$), derived from the standard normal curve, and the standard error in the point estimate. The standard error in the point estimate is computed using the residuals from the training set employed in the ML model. The residuals (ε) are cal-

culated as the difference between GRACE JPL mascon and the reconstructed BNML_TWSA during the training period, as outlined below:

$$GRACE_t = BNML_TWSA_t + \varepsilon. \quad (10)$$

These residuals capture errors arising from data noise and structural model inaccuracies, as discussed earlier. A classical approach to determining the standard error (σ_ε) of the residuals is given by the following:

$$\sigma_\varepsilon = \sqrt{\text{variance}(\varepsilon)}. \quad (11)$$

For most grid cells, the residuals follow a normal distribution (Fig. 20a). The normality of the residuals was verified using the Shapiro–Wilk test, with normality assumed when the p value exceeds 0.05. Consequently, it is appropriate to use the standard error to estimate the confidence interval (Humphrey and Gudmundsson, 2019). The confidence interval is calculated as follows:

$$95\% \text{ CI} = \text{Point estimate} \pm C_{confidence} \times \sigma_\varepsilon. \quad (12)$$

The spatial distribution of the standard error (σ_ε) is shown in Fig. 20b. The σ_ε values for grid cells in arid regions are significantly smaller compared to those in other regions, indicating improved accuracy in arid areas. This observation aligns with the findings of Humphrey and Gudmundsson (2019).

Climate change and anthropogenic activities are critical factors that can introduce additional uncertainties into the assessment of TWS. These uncertainties arise from factors such as land-use changes, irrigation practices and urbanization, which significantly influence regional water storage dynamics. In this study, variables derived from LSMs were utilized as potential predictors. However, future research could benefit from incorporating input variables from GHMs to better account for anthropogenic influences. GHMs are particularly well suited for modeling human interventions in water resources, offering a more realistic representation of these activities (Bibi et al., 2024). It is important to acknowledge that both LSMs and GHMs have inherent limitations when utilized as physically based sources of TWSAs (Bibi et al., 2024). The integration of ML models with physical models can help address these limitations, reducing errors in hydrological analyses (Xu et al., 2014). Numerous studies have demonstrated that ML models frequently outperform traditional hydrological models in various applications (Kim and Kim, 2021; Liang et al., 2023). This suggests that leveraging ML models, alongside advancements in physical modeling, holds great promise for improving the accuracy and reliability of hydrological assessments.

5 Data availability

The presented dataset is published at <https://doi.org/10.6084/m9.figshare.25376695> (Mandal et al., 2024), and updates will be published as and when needed. The BNML_TWSA dataset is available for all grid cells globally, with a spatial resolution of $0.50^\circ \times 0.50^\circ$, similar to the GRACE JPL mascon, and is provided in NetCDF format. The inputs to the ML models and the optimal predictors selected using BNs at each grid cell globally are also published. Additionally, the uncertainty associated with the BNML_TWSA is made available in terms of standard error, provided as a NetCDF file. All datasets utilized in this study are readily accessible; comprehensive dataset information, along with the respective links, is provided in this section. JPL GRACE mascon data are obtained from the Physical Oceanography Distributed Active Archive Center (NASA/JPL, 2023, <https://doi.org/10.5067/TEMSC-3JC63>; Watkins et al., 2015, https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC-GRFO_MASCON_CRI_GRID_RL06.1_V3). TWS data are retrieved from NASA GLDAS CLSM simulations (Li et al., 2019) (<https://doi.org/10.5067/LYHA9088MFWQ>, Li et al., 2018; <https://doi.org/10.5067/TXBMLX370XX8>, Li et al., 2020). Canopy surface water, soil moisture content, snow water, precipitation and tem-

perature data are taken from the GLDAS Noah Land Surface Model (Rodell et al., 2004) (<https://doi.org/10.5067/9SQ1B3ZXP2C5>, Beaudoin and Rodell, 2019; <https://doi.org/10.5067/SXAVCZFAQLNO>, Beaudoin and Rodell, 2020). The Dipole Mode Index is taken from the National Oceanic and Atmospheric Administration (NOAA) Physical Sciences Laboratory (Saji et al., 1999; Saji and Yamagata, 2003) (https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data). North Atlantic Oscillation data (Wallace and Gutzler, 1981; Barnston and Livezey, 1987) (<https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/norm.nao.monthly.b5001.current.ascii>) and the Oceanic Niño Index (Barnston et al., 1997) (<https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt>) are retrieved from the NOAA Climate Prediction Center.

6 Code availability

We have used standard Python packages such as tensorflow, keras, sklearn and mlxtend to build the ML models, while matplotlib was employed to generate the plots. No specific software tools for ML were used. For reproducibility, the Python codes used to build the ML models are available from the same DOI-based repository: <https://doi.org/10.6084/m9.figshare.25376695> (Mandal et al., 2024).

7 Conclusions

In this study, we utilized Bayesian networks (BNs), a novel predictor selection technique, and machine learning (ML) models to reconstruct the global terrestrial water storage anomaly (BNML_TWSA) product, which is a GRACE-like TWSA dataset, thereby filling data gaps in GRACE and generating hindcasts for the pre-GRACE period. The major conclusions from this study are outlined below.

For the target TWSAs, optimal inputs are selected among meaningful predictor variables, such as land surface model outputs (TWSA from the Catchment Land Surface Model, CTWSA, and the Noah Land Surface Model, NTWSA), meteorological variables (precipitation and temperature) and climate indices (the Dipole Mode Index, DMI; the North Atlantic Oscillation, NAO; and the Oceanic Niño Index, ONI). It is observed that the climate indices, ONI and DMI, are selected by BNs as optimal predictors for a large number of grid cells globally, along with TWSAs from LSM outputs. This establishes that, in addition to the available LSM-based TWSA products, large-scale climate indices are more important predictors of TWSAs than the local meteorological inputs.

At the global scale, convolutional neural network (CNN), support vector regression (SVR), extra trees regressor (ETR) and stacking ensemble regression (SER) models are employed following the selection of optimal predictors through BNs at each grid cell to finally obtain BNML_TWSA. It

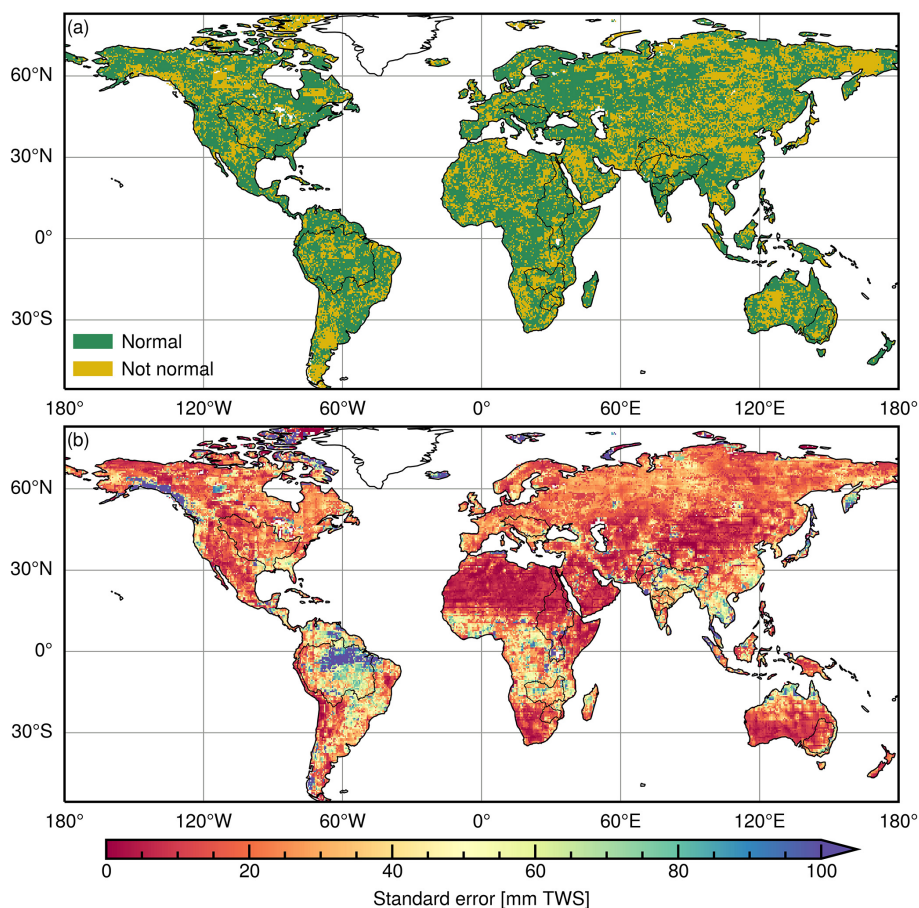


Figure 20. Characteristics of residuals of reconstructed BNML_TWSA computed against GRACE JPL mascon during the training period: (a) the Shapiro–Wilk normality test result on residuals; (b) the standard error of the residuals.

is noted that a single ML model cannot perform optimally across all grid cells worldwide, due to the significant spatial variability in the important predictors. However, the performance of ETR is found to be the best for most of the grid cells within the Ganga–Brahmaputra–Meghna, Godavari, Krishna, Limpopo and Nile river basins. ETR performs best in 44 % grid cells worldwide followed by SVR, SER and CNN.

The proposed approach yields a more reliable estimate of TWS compared to the outputs of global hydrological models (GHMs) and land surface models (LSMs), which have significant biases due to inherent uncertainty and a lack of representation of some of the physical processes. BNML_TWSA outperforms NTWSA and CTWSA for most grid cells worldwide. For river basins such as the Indus and Nile, BNML_TWSA matches GRACE TWSAs very closely, even during the period when the TWSAs from LSM outputs deviate substantially from GRACE TWSAs. With respect to evaluating the basin-wise average BNML_TWSA against GRACE TWSAs, the Zambezi Basin in Africa exhibited the highest correlation coefficient ($CC = 0.989$), followed by the Godavari ($CC = 0.984$) and Krishna ($CC = 0.983$) basins

in India. Further, the accurate reflection of historical extreme climate events, such as major floods, via the hind-casted BNML_TWSA supports the enhanced accuracy of the proposed model and the developed TWSA dataset. A comparative analysis with TWSA products developed in recent literature (Humphrey and Gudmundsson, 2019; Sun et al., 2020) indicates that BNML_TWSA surpasses these datasets when evaluated against the overlapping GRACE period. Additionally, streamflow, computed as residuals of the water balance derived using BNML_TWSA and the TWSAs developed in recent literature, has been evaluated against observed streamflow across six global river basins. The evaluation revealed a reasonably good correlation for BNML_TWSA. Furthermore, the uncertainty associated with BNML_TWSA is assessed for each grid cell in the form of standard error (σ_ε). The results showed that the standard error of the BNML_TWSA exhibits a smaller magnitude in grid cells located in arid regions compared to those in other regions. Hence, this study demonstrates that the proposed BN- and ML-based approach can effectively learn complex relationships between various inputs and GRACE TWSAs, enabling

global reconstruction and hindcasting of TWSAs, which is essential for several hydroclimatological studies.

Author contributions. NM: conceptualization, data curation, formal analysis, methodology and original draft preparation; PD: methodology and formal analysis; KC: conceptualization, supervision, and review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Acknowledgements. The authors gratefully acknowledge the editor, Benjamin Fersch and the anonymous reviewer for their invaluable suggestions, which significantly contributed to the substantial revision and improvement of this paper.

Financial support. This research has been supported by the DST-SERB, Government of India (grant no. CRG/2022/007728).

Review statement. This paper was edited by Kirsten Elger and reviewed by Benjamin Fersch and one anonymous referee.

References

- Ahmad, M. W., Reynolds, J., and Rezugui, Y.: Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees, *J. Clean. Prod.*, 203, 810–821, <https://doi.org/10.1016/j.jclepro.2018.08.207>, 2018.
- Ahmed, A. A., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., and Yang, L.: Hybrid deep learning method for a week-ahead evapotranspiration forecasting, *Stoch. Env. Res. Risk A.*, 36, 831–849, <https://doi.org/10.1007/s00477-021-02078-x>, 2022.
- Ahmed, M., Sultan, M., Elbayoumi, T., and Tissot, P.: Forecasting GRACE Data over the African Watersheds Using Artificial Neural Networks, *Remote Sensing*, 11, 1769, <https://doi.org/10.3390/rs11151769>, 2019.
- Alibabaei, K., Gaspar, P. D., and Lima, T. M.: Modeling soil water content and reference evapotranspiration from climate data using deep learning method, *Appl. Sci.*, 11, 5029, <https://doi.org/10.3390/app11115029>, 2021.
- Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D. K., Knapp, K. R., Cecil, L. D., Nelson, B. R., and Prat, O. P.: PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies, *B. Am. Meteorol. Soc.*, 96, 69–83, 2015.
- Bamston, A. G., Chelliah, M., and Goldenberg, S. B.: Documentation of a highly ENSO-related sst region in the equatorial pacific: Research note, *Atmos.–Ocean*, 35, 367–383, <https://doi.org/10.1080/07055900.1997.9649597>, 1997 (data available at: <https://www.cpc.ncep.noaa.gov/data/indices/oni.ascii.txt>, last access: 4 June 2023).
- Barnston, A. G. and Livezey, R. E.: Classification, Seasonality and Persistence of Low-Frequency Atmospheric Circulation Patterns, *Mon. Weather Rev.*, 115, 1083–1126, [https://doi.org/10.1175/1520-0493\(1987\)115<1083:CSAPOL>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1083:CSAPOL>2.0.CO;2), 1987 (data available at: <https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/norm.nao.monthly.b5001.current.ascii>, last access: 4 June 2023).
- Beaudoing, H. and Rodell, M.: NASA/GSFC/HSL, GLDAS Noah Land Surface Model L4 monthly 0.25×0.25 degree V2.0, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/9SQ1B3ZXP2C5>, 2019.
- Beaudoing, H. and Rodell, M.: NASA/GSFC/HSL, GLDAS Noah Land Surface Model L4 monthly 0.25×0.25 degree V2.1, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/SXAVCZFAQLNO>, 2020.
- Becker, M., Meyssignac, B., Xavier, L., Cazenave, A., Alkama, R., and Decharme, B.: Past terrestrial water storage (1980–2008) in the Amazon Basin reconstructed from GRACE and in situ river gauging data, *Hydrol. Earth Syst. Sci.*, 15, 533–546, <https://doi.org/10.5194/hess-15-533-2011>, 2011.
- Bibi, S., Zhu, T., Rateb, A., Scanlon, B. R., Kamran, M. A., El-nashar, A., Bennour, A., and Li, C.: Benchmarking multimodel terrestrial water storage seasonal cycle against Gravity Recovery and Climate Experiment (GRACE) observations over major global river basins, *Hydrol. Earth Syst. Sci.*, 28, 1725–1750, <https://doi.org/10.5194/hess-28-1725-2024>, 2024.
- Boergens, E., Kvas, A., Eicker, A., Dobslaw, H., Schawohl, L., Dahle, C., Murböck, M., and Flechtner, F.: Uncertainties of GRACE-Based Terrestrial Water Storage Anomalies for Arbitrary Averaging Regions, *J. Geophys. Res.-Sol. Ea.*, 127, e2021JB022081, <https://doi.org/10.1029/2021JB022081>, 2022.
- Brandimarte, L., Di Baldassarre, G., Bruni, G., D'Odorico, P., and Montanari, A.: Relation between the North-Atlantic Oscillation and hydroclimatic conditions in Mediterranean areas, *Water Resour. Manage.*, 25, 1269–1279, <https://doi.org/10.1007/s11269-010-9742-5>, 2011.
- Candogan Yossef, N., van Beek, L. P. H., Kwadijk, J. C. J., and Bierkens, M. F. P.: Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes, *Hydrol. Earth Syst. Sci.*, 16, 4233–4246, <https://doi.org/10.5194/hess-16-4233-2012>, 2012.
- Chanda, K. and Das, P.: Dimensionality reduction of correlated meteorological variables by Bayesian network-based graphical modeling, in: *Handbook of HydroInformatics Volume II: Advanced Machine Learning Techniques*, vol. 2, 69–76, Elsevier, <https://doi.org/10.1016/B978-0-12-821961-4.00021-X>, 2022.
- Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K.: Large-scale empirical validation of Bayesian Network structure learning algorithms with

- noisy data, *Int. J. Approx. Reason.*, 131, 151–188, <https://doi.org/10.1016/j.ijar.2021.01.001>, 2021.
- Cortes, C. and Vapnik, V.: Support-vector networks, *Mach. Learn.*, 20, 273–297, <https://doi.org/10.1007/BF00994018>, 1995.
- Das, P. and Chanda, K.: Bayesian Network based modeling of regional rainfall from multiple local meteorological drivers, *J. Hydrol.*, 591, 125563, <https://doi.org/10.1016/j.jhydrol.2020.125563>, 2020.
- Das, P. and Chanda, K.: A Bayesian network approach for understanding the role of large-scale and local hydro-meteorological variables as drivers of basin-scale rainfall and streamflow, *Stoch. Env. Res. Risk A.*, 37, 1535–1556, <https://doi.org/10.1007/s00477-022-02356-2>, 2022.
- Das, P. and Chanda, K.: Selection of optimum GCMs through Bayesian networks for developing improved machine learning based multi-model ensembles of precipitation and temperature, *Stoch. Env. Res. Risk A.*, 39, 155–179, <https://doi.org/10.1007/s00477-024-02856-3>, 2024.
- Das, P., Sachindra, D. A., and Chanda, K.: Machine Learning-Based Rainfall Forecasting with Multiple Non-Linear Feature Selection Algorithms, *Water Resour. Manage.*, 36, 6043–6071, <https://doi.org/10.1007/s11269-022-03341-8>, 2022.
- Dutta, R. and Maity, R.: Time-varying network-based approach for capturing hydrological extremes under climate change with application on drought, *J. Hydrol.*, 603, 126958, <https://doi.org/10.1016/j.jhydrol.2021.126958>, 2021.
- Felfelani, F., Wada, Y., Longuevergne, L., and Pokhrel, Y. N.: Natural and human-induced terrestrial water storage change: A global analysis using hydrological models and GRACE, *J. Hydrol.*, 553, 105–118, <https://doi.org/10.1016/j.jhydrol.2017.07.048>, 2017.
- Ferreira, L. B. and da Cunha, F. F.: Multi-step ahead forecasting of daily reference evapotranspiration using deep learning, *Comput. Electron. Agr.*, 178, 105728, <https://doi.org/10.1016/j.compag.2020.105728>, 2020.
- Forootan, E., Kusche, J., Loth, I., Schuh, W.-D., Eicker, A., Awange, J., Longuevergne, L., Diekkrüger, B., Schmidt, M., and Shum, C.: Multivariate prediction of total water storage changes over West Africa from multi-satellite data, *Surv. Geophys.*, 35, 913–940, <https://doi.org/10.1007/s10712-014-9292-0>, 2014.
- Forootan, E., Khaki, M., Schumacher, M., Wulfmeyer, V., Mehrnegar, N., van Dijk, A. I., Brocca, L., Farzaneh, S., Akinluyi, F., Ramillien, G., Shum, C. K., Awange, J., and Mostafaie, A.: Understanding the global hydrological droughts of 2003–2016 and their relationships with teleconnections, *Sci. Total Environ.*, 650, 2587–2604, <https://doi.org/10.1016/j.scitotenv.2018.09.231>, 2019.
- Gao, S., Hao, W., Fan, Y., Li, F., and Wang, J.: A Multi-Source GRACE Fusion Solution via Uncertainty Quantification of GRACE-Derived Terrestrial Water Storage (TWS) Change, *J. Geophys. Res.-Sol. Ea.*, 128, e2023JB026908, <https://doi.org/10.1029/2023JB026908>, 2023.
- Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, 377, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>, 2009.
- Hafez, Y.: Study on the relationship between the oceanic nino index and surface air temperature and precipitation rate over the Kingdom of Saudi Arabia, *Journal of Geoscience and Environment Protection*, 4, 146, <https://doi.org/10.4236/gep.2016.45015>, 2016.
- Harou, A. P., Lajoie, R. F., Kniveton, D. R., and Frogley, M. R.: The influence of the Indian Ocean dipole mode on precipitation over the Seychelles, *Int. J. Climatol.*, 26, 45–54, <https://doi.org/10.1002/joc.1239>, 2006.
- Huffman, G. J., Behrangi, A., Bolvin, D. T., and Nelkin, E. J.: GPCP Version 3.2 Satellite-Gauge (SG) Combined Precipitation Data Set, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/MEASURES/GPCP/DATA304>, 2022.
- Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., and Tan, J.: GPM IMERG Final Precipitation L3 1 month 0.1 degree x 0.1 degree V07, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/GPM/IMERG/3B-MONTH/07>, 2023.
- Humphrey, V. and Gudmundsson, L.: GRACE-REC: a reconstruction of climate-driven water storage changes over the last century, *Earth Syst. Sci. Data*, 11, 1153–1170, <https://doi.org/10.5194/essd-11-1153-2019>, 2019.
- Humphrey, V., Gudmundsson, L., and Seneviratne, S. I.: A global reconstruction of climate-driven subdecadal water storage variability, *Geophys. Res. Lett.*, 44, 2300–2309, <https://doi.org/10.1002/2017GL072564>, 2017.
- Jing, W., Di, L., Zhao, X., Yao, L., Xia, X., Liu, Y., Yang, J., Li, Y., and Zhou, C.: A data-driven approach to generate past GRACE-like terrestrial water storage solution by calibrating the land surface model simulations, *Adv. Water Resour.*, 143, 103683, <https://doi.org/10.1016/j.advwatres.2020.103683>, 2020.
- Kalu, I., Ndehedehe, C. E., Okwuashi, O., Eyoh, A. E., and Ferreira, V. G.: Reconstructing terrestrial water storage anomalies using convolution-based support vector machine, *J. Hydrol.-Reg. Stud.*, 46, 101326, <https://doi.org/10.1016/j.ejrh.2023.101326>, 2023.
- Kalu, I., Ndehedehe, C. E., Ferreira, V. G., and Kennard, M. J.: Machine learning assessment of hydrological model performance under localized water storage changes through downscaling, *J. Hydrol.*, 628, 130597, <https://doi.org/10.1016/j.jhydrol.2023.130597>, 2024.
- Khan, M. I. and Maity, R.: Hybrid deep learning approach for multi-step-ahead daily rainfall prediction using GCM simulations, *IEEE Access*, 8, 52774–52784, <https://doi.org/10.1109/ACCESS.2020.2980977>, 2020.
- Kim, C. and Kim, C.-S.: Comparison of the performance of a hydrologic model and a deep learning technique for rainfall-runoff analysis, *Tropical Cyclone Research and Review*, 10, 215–222, <https://doi.org/10.1016/j.tccr.2021.12.001>, 2021.
- Kumar, A. S., Anirudh, A., Jeet Golecha, M., Karthik Raja, A., Jayaraman, B., and Mirmalinee, T. T.: Multi Regressor Based User Rating Predictor for ImageCLEF Aware 2022, *CEUR Workshop Proceedings*, 3180, 1590–1595, 2022.
- Lee, D.-G. and Ahn, K.-H.: A stacking ensemble model for hydrological post-processing to improve streamflow forecasts at medium-range timescales over South Korea, *J. Hydrol.*, 600, 126681, <https://doi.org/10.1016/j.jhydrol.2021.126681>, 2021.
- Li, B., Beaudoin, H., and Rodell, M.: NASA/GSFC/HSL, GLDAS Catchment Land Surface Model L4 daily 0.25 × 0.25 degree V2.0, Greenbelt, Maryland, USA, Goddard Earth Sciences

- Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/LYHA9088MFWQ>, 2018.
- Li, B., Rodell, M., Kumar, S., Beaudoin, 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., and Bettadpur, S.: Global GRACE Data Assimilation for Groundwater and Drought Monitoring: Advances and Challenges, *Water Resour. Res.*, 55, 7564–7586, <https://doi.org/10.1029/2018WR024618>, 2019.
- Li, B., Beaudoin, H., and Rodell, M.: NASA/GSFC/HSL, GLDAS Catchment Land Surface Model L4 daily 0.25×0.25 degree GRACE-DA1 V2.2, Greenbelt, Maryland, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/TXBMLX370XX8>, 2020.
- Li, F., Kusche, J., Rietbroek, R., Wang, Z., Forootan, E., Schulze, K., and Lück, C.: Comparison of data-driven techniques to reconstruct (1992–2002) and predict (2017–2018) GRACE-like gridded total water storage changes using climate inputs, *Water Resour. Res.*, 56, e2019WR026551, <https://doi.org/10.1029/2019WR026551>, 2020.
- Liang, W., Chen, Y., Fang, G., and Kaldybayev, A.: Machine learning method is an alternative for the hydrological model in an alpine catchment in the Tianshan region, *Central Asia, J. Hydrol.-Reg. Stud.*, 49, 101492, <https://doi.org/10.1016/j.ejrh.2023.101492>, 2023.
- Long, D., Shen, Y., Sun, A., Hong, Y., Longuevergne, L., Yang, Y., Li, B., and Chen, L.: Drought and flood monitoring for a large karst plateau in Southwest China using extended GRACE data, *Remote Sens. Environ.*, 155, 145–160, <https://doi.org/10.1016/j.rse.2014.08.006>, 2014.
- Mandal, N. and Chanda, K.: Performance of machine learning algorithms for multi-step ahead prediction of reference evapotranspiration across various agro-climatic zones and cropping seasons, *J. Hydrol.*, 620, 129418, <https://doi.org/10.1016/j.jhydrol.2023.129418>, 2023.
- Mandal, N., Das, P., and Chanda, K.: Optimal feature selection for improved ML based reconstruction of Global Terrestrial Water Storage Anomalies, *Figshare*, <https://doi.org/10.6084/M9.FIGSHARE.25376695>, 2024.
- Martens, B., Miralles, D. G., Lievens, H., van der Schalie, R., de Jeu, R. A. M., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E. C.: GLEAM v3: satellite-based land evaporation and root-zone soil moisture, *Geosci. Model Dev.*, 10, 1903–1925, <https://doi.org/10.5194/gmd-10-1903-2017>, 2017.
- Martinez-Gil, J.: A comprehensive review of stacking methods for semantic similarity measurement, *Machine Learning with Applications*, 10, 100423, <https://doi.org/10.1016/j.mlwa.2022.100423>, 2022.
- McNally, A.: NASA/GSFC/HSL, FLDAS Noah Land Surface Model L4 Global Monthly 0.1×0.1 degree (MERRA-2 and CHIRPS), Greenbelt, MD, USA, Goddard Earth Sciences Data and Information Services Center (GES DISC) [data set], <https://doi.org/10.5067/5NHC22T9375G>, 2018.
- Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, *Hydrol. Earth Syst. Sci.*, 15, 453–469, <https://doi.org/10.5194/hess-15-453-2011>, 2011.
- Mo, S., Zhong, Y., Forootan, E., Mehrnegar, N., Yin, X., Wu, J., Feng, W., and Shi, X.: Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during GRACE and GRACE-FO gap, *J. Hydrol.*, 604, 127244, <https://doi.org/10.1016/j.jhydrol.2021.127244>, 2022.
- NASA/JPL: JPL GRACE and GRACE-FO Mascon Ocean, Ice, and Hydrology Equivalent Water Height Coastal Resolution Improvement (CRI) Filtered Release 06.1 Version 03, NASA Physical Oceanography Distributed Active Archive Center [data set], <https://doi.org/10.5067/TEMSC-3JC63>, 2023.
- Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, *J. Hydrol.*, 10, 282–290, [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6), 1970.
- Nie, N., Zhang, W., Zhang, Z., Guo, H., and Ishwaran, N.: Reconstructed terrestrial water storage change (Δ TWS) from 1948 to 2012 over the Amazon Basin with the latest GRACE and GLDAS products, *Water Resour. Manage.*, 30, 279–294, <https://doi.org/10.1007/s11269-015-1161-1>, 2016.
- Phillips, T., Nerem, R., Fox-Kemper, B., Famiglietti, J., and Rajagopalan, B.: The influence of ENSO on global terrestrial water storage using GRACE, *Geophys. Res. Lett.*, 39, L16705, <https://doi.org/10.1029/2012GL052495>, 2012.
- Raghavendra, S. and Deka, P. C.: Support vector machine applications in the field of hydrology: A review, *Applied Soft Computing Journal*, 19, 372–386, <https://doi.org/10.1016/j.asoc.2014.02.002>, 2014.
- Rodell, M., Houser, P. R., Jambor, U., Gottschalk, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, *B. Am. Meteorol. Soc.*, 85, 381–394, <https://doi.org/10.1175/BAMS-85-3-381>, 2004.
- Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J., Beaudoin, H. K., Landerer, F. W., and Lo, M.-H.: Emerging trends in global freshwater availability, *Nature*, 557, 651–659, <https://doi.org/10.1038/s41586-023-06879-8>, 2018.
- Rustemeier, E., Hänsel, S., Finger, P., Schneider, U., and Ziese, M.: GPCC Climatology Version 2022 at 0.25° : Monthly Land-Surface Precipitation Climatology for Every Month and the Total Year from Rain-Gauges built on GTS-based and Historical Data, Global Precipitation Climatology Centre [data set], https://doi.org/10.5676/DWD_GPCC/CLIM_M_V2022_025, 2022.
- Saharia, M., Jain, A., Baishya, R. R., Haobam, S., Sreejith, O., Pai, D., and Rafieeiniasab, A.: India flood inventory: creation of a multi-source national geospatial database to facilitate comprehensive flood research, *Nat. Hazards*, 108, 619–633, <https://doi.org/10.1007/s11069-021-04698-6>, 2021.
- Saji, N. and Yamagata, T.: Possible impacts of Indian Ocean dipole mode events on global climate, *Clim. Res.*, 25, 151–169, <https://doi.org/10.3354/cr025151>, 2003 (data available at: https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data, last access: 4 June 2023).
- Saji, N., Goswami, B. N., Vinayachandran, P., and Yamagata, T.: A dipole mode in the tropical Indian Ocean, *Nature*, 401, 360–363, <https://doi.org/10.1038/43854>, 1999 (data

- available at: https://psl.noaa.gov/gcos_wgsp/Timeseries/Data/dmi.had.long.data, last access: 4 June 2023).
- Satish Kumar, K., AnandRaj, P., Sreelatha, K., and Sridhar, V.: Reconstruction of GRACE terrestrial water storage anomalies using Multi-Layer Perceptrons for South Indian River basins, *Sci. Total Environ.*, 857, 159289, <https://doi.org/10.1016/j.scitotenv.2022.159289>, 2023.
- Scanlon, B. R., Rateb, A., Pool, D. R., Sanford, W., Save, H., Sun, A., Long, D., and Fuchs, B.: 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>, 2021.
- Schneider, U., Fuchs, T., Meyer-Christoffer, A., and Rudolf, B.: Global precipitation analysis products of the GPCC, Global Precipitation Climatology Centre (GPCC), DWD, Internet Publikation, 112, 2008.
- Scutari, M.: Learning Bayesian networks with the bnlearn R Package, *J. Stat. Softw.*, 35, 1–22, <https://doi.org/10.18637/jss.v035.i03>, 2010.
- Scutari, M. and Denis, J.-B.: Bayesian networks: with examples in R, Chapman and Hall/CRC, ISBN 042934743X, <https://doi.org/10.1201/9780429347436>, 2021.
- Scutari, M. and Nagarajan, R.: Identifying significant edges in graphical models of molecular networks, *Artif. Intell. Med.*, 57, 207–217, 2011.
- Sevinc, V., Kucuk, O., and Goltas, M.: A Bayesian network model for prediction and analysis of possible forest fire causes, *Forest Ecol. Manage.*, 457, 117723, <https://doi.org/10.1016/j.foreco.2019.117723>, 2020.
- Sun, A. Y., Wang, D., and Xu, X.: Monthly streamflow forecasting using Gaussian process regression, *J. Hydrol.*, 511, 72–81, <https://doi.org/10.1016/j.jhydrol.2014.01.023>, 2014.
- Sun, A. Y., Scanlon, B. R., Zhang, Z., Walling, D., Bhanja, S. N., Mukherjee, A., and Zhong, Z.: Combining Physically Based Modeling and Deep Learning for Fusing GRACE Satellite Data: Can We Learn From Mismatch?, *Water Resour. Res.*, 55, 1179–1195, <https://doi.org/10.1029/2018WR023333>, 2019.
- Sun, A. Y., Scanlon, B. R., Save, H., and Rateb, A.: Reconstruction of GRACE Total Water Storage Through Automated Machine Learning, *Water Resour. Res.*, 57, 1–20, <https://doi.org/10.1029/2020WR028666>, 2021.
- Sun, Z., Long, D., Yang, W., Li, X., and Pan, Y.: Reconstruction of GRACE data on changes in total water storage over the global land surface and 60 basins, *Water Resour. Res.*, 56, e2019WR026250, <https://doi.org/10.1029/2019WR026250>, 2020.
- United States National Hurricane Center: Costliest U.S. tropical cyclones tables update (Report), United States National Hurricane Center, 12 January 2018, <https://www.nhc.noaa.gov/news/UpdatedCostliest.pdf> (last access: 30 May 2025), 2018.
- Uz, M., Atman, K. G., Akyilmaz, O., Shum, C. K., Keleş, M., Ay, T., Tandoğdu, B., Zhang, Y., and Mercan, H.: Bridging the gap between GRACE and GRACE-FO missions with deep learning aided water storage simulations, *Sci. Total Environ.*, 830, 154701, <https://doi.org/10.1016/j.scitotenv.2022.154701>, 2022.
- Vitolo, C., Scutari, M., Ghalaieny, M., Tucker, A., and Russell, A.: Modeling Air Pollution, Climate, and Health Data Using Bayesian Networks: A Case Study of the English Regions, *Earth Space Sci.*, 5, 76–88, <https://doi.org/10.1002/2017EA000326>, 2018.
- Wallace, J. M. and Gutzler, D. S.: Teleconnections in the geopotential height field during the Northern Hemisphere winter, *Mon. Weather Rev.*, 109, 784–812, [https://doi.org/10.1175/1520-0493\(1981\)109<0784:TITGHF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2), 1981 (data available at: <https://www.cpc.ncep.noaa.gov/products/precip/CWlink/pna/norm.nao.monthly.b5001.current.ascii>, last access: 4 June 2023).
- Watkins, M. M., Wiese, D. N., Yuan, D.-N., Boening, C., and Landerer, F. W.: Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons, *J. Geophys. Res.-Sol. Ea.*, 120, 2648–2671, <https://doi.org/10.1002/2014JB011547>, 2015 (data available at: https://podaac.jpl.nasa.gov/dataset/TELLUS_GRAC-GRFO_MASCON_CRI_GRID_RL06.1_V3, last access: 27 May 2023).
- Wiese, D. N., Landerer, F. W., and Watkins, M. M.: Quantifying and reducing leakage errors in the JPL RL05M GRACE mascon solution, *Water Resour. Res.*, 52, 7490–7502, <https://doi.org/10.1002/2016WR019344>, 2016.
- Xu, T., Valocchi, A. J., Choi, J., and Amir, E.: Use of Machine Learning Methods to Reduce Predictive Error of Groundwater Models, *Groundwater*, 52, 448–460, <https://doi.org/10.1111/gwat.12061>, 2014.
- Yang, X., Tian, S., You, W., and Jiang, Z.: Reconstruction of continuous GRACE/GRACE-FO terrestrial water storage anomalies based on time series decomposition, *J. Hydrol.*, 603, 127018, <https://doi.org/10.1016/j.jhydrol.2021.127018>, 2021.
- Yu, Q., Wang, S., He, H., Yang, K., Ma, L., and Li, J.: Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model, *Int. J. Appl. Earth Obs.*, 102, 102404, <https://doi.org/10.1016/j.jag.2021.102404>, 2021.
- Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, *J. Hydrol.*, 598, 126266, <https://doi.org/10.1016/j.jhydrol.2021.126266>, 2021.