scientific data

OPEN

DATA DESCRIPTOR

GRAiCE: reconstructing terrestrial water storage anomalies with recurrent neural networks

Irene Palazzoli[1]✉, Serena Ceola[1] & Pierre Gentine[2,3]

The Gravity Recovery and Climate Experiment (GRACE) and its follow-on (GRACE-FO) missions have provided estimates of Terrestrial Water Storage Anomalies (TWSA) since 2002, enabling the monitoring of global hydrological changes. However, temporal gaps within these datasets and the lack of TWSA observations prior to 2002 limit our understanding of long-term freshwater variability. In this study, we develop GRAiCE, a set of four global monthly TWSA reconstructions from 1984 to 2021 at 0.5° spatial resolution, using Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) neural networks. Our models accurately reproduce GRACE/GRACE-FO observations at the global scale and effectively capture the impacts of climate extremes. Overall, GRAiCE outperforms a previous reference TWSA reconstruction in predicting observed TWSA and provides reliable water budget estimates at the river basin scale. By generating long-term continuous TWSA time series, GRAiCE will offer valuable insights into the impacts of climate variability and change on freshwater resources.

Background & Summary

Terrestrial Water Storage (TWS) is the sum of all freshwater reservoirs on and below Earth's continental land surface, i.e., surface water, groundwater, soil moisture, snow, and ice, representing a central component of the global water cycle[1–3]. Changes in TWS reflect the alteration of hydrological fluxes, which may be caused by natural variability, climate change, and human pressure[4,5]. Therefore, the observed variations of TWS with respect to a long-term mean, namely TWS Anomalies (TWSA), allow tracking global freshwater availability, detecting hydrological extremes, such as droughts and floods, and estimating the impacts of climate change and anthropogenic activities on the water cycle[6–8].

Since 2002, the Gravity Recovery and Climate Experiment (GRACE) satellite mission has been measuring the Earth's gravity field variations, providing unprecedented continuous and rapid TWSA estimates at the global scale[9–11]. GRACE overcomes the shortcomings of hydrological modeling (and its assumptions), the limitations of sparse *in-situ* measurements, and alternative sensors, resulting in the best available data for the global monitoring of TWSA[8,12]. However, the GRACE mission was interrupted in 2017 and it was resumed only in 2018 under the name of GRACE Follow-On (GRACE-FO) mission, generating a data gap of about one year (11 months) between July 2017 and May 2018. Besides this long-term temporal gap between the two missions, other gaps are found within the GRACE time series, in which 20 months are missing overall, and in the GRACE-FO time series, in which there is a total gap of 4 months of observations between 2018 and 2019. These data gaps hinder a complete understanding of long-term changes in freshwater storage, and thus need to be filled in to ensure data continuity[13,14]. To address this issue, numerous studies have focused on the reconstruction of GRACE time series based on data-driven approaches and statistical techniques[15,16]. For instance, Humphrey and Gudmundsson[2] developed a climate-driven regression model to reconstruct the detrended and deseasonalized GRACE TWSA using precipitation and air temperature data, whereas Rateb *et al.*[17] applied a Bayesian inference approach to decompose and model GRACE/GRACE-FO time series and generate reconstructed TWSA solutions along with an uncertainty analysis. Recently, there has been a significant increase in the application of machine learning (ML) algorithms to produce GRACE-like TWSA reconstructions[5,13,16,18,19]. These ML models are usually designed to learn the complex relationship between GRACE TWSA and its climatic and hydrological drivers[20].

[1]Department of Civil, Chemical, Environmental, and Materials Engineering, Alma Mater Studiorum – Università di Bologna, Bologna, Italy. [2]Department of Earth and Environmental Engineering, Columbia University, New York, USA. [3]Columbia Climate School, Columbia University, New York, NY, USA. ✉e-mail: irene.palazzoli@unibo.it

Long Short-Term Memory (LSTM) networks are a powerful extension of Recurrent Neural Network (RNN) that have the capability of learning long-term dependencies between input and output data of the model, becoming an effective tool for time series forecasting[21–23]. LSTM networks have been extensively applied for time series prediction in hydrological problems to overcome the limited predictive power of existing physically-based hydrological models and generate estimates in ungauged basins[24–27]. However, the use of LSTM algorithms for TWSA reconstruction, especially at the global scale, remains largely unexplored, with only one regional application over the Tarim River Basin in China[28].

Here, we apply the (Unidirectional) Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) neural networks to develop GRA*i*CE, a collection of four climate-driven models that reproduce GRACE/GRACE-FO TWSA observations and reconstruct TWSA time series before the GRACE era (i.e., prior to 2002). We identify five meteorological predictors that play a key role in controlling the water cycle – monthly data of total precipitation, snow depth water equivalent, surface net solar radiation, surface air temperature, and surface air relative humidity – along with data of solar-induced fluorescence to approximate vegetation dynamics. We use the Optuna tuning tool[29] to determine the optimal hyperparameters of the ML models within different climatic regions and run the model fit at the grid cell level. The resulting GRA*i*CE dataset produces continuous monthly TWSA time series between 1984 and 2021 with a 0.5° spatial resolution at the global scale. By providing reliable long-term TWSA reconstructions globally, GRA*i*CE serves as a valuable resource for the analysis of climate- and human-induced alterations of the terrestrial freshwater storage that occurred over the last decades.

Methods

The TWSA reconstructions presented here consist of four monthly TWSA time series from January 1984 to December 2021 on a grid with a 0.5° resolution. Each reconstruction of the target variable (i.e., GRACE/GRACE-FO TWSA observations) is produced by either an LSTM or BiLSTM model, using two distinct sets of predictors. The first set of predictors combines five fundamental meteorological variables with vegetation dynamics data, whereas the second set relies only on the meteorological variables (Fig. 1).

Datasets. *GRACE and GRACE-FO observations.* We employ the latest version (version 03) of the GRACE and GRACE-FO monthly retrievals produced by the Jet Propulsion Laboratory (JPL) using the mass concentration (*mascon*) approach[30]. The *mascon*-based solution presents significant improvements compared to the spherical harmonic products, both in terms of leakage errors and postprocessing, generating a dataset that is more correlated to *in-situ* data and more user friendly. The JPL *mascon* dataset provides gridded monthly TWSA data as water height anomalies in cm of equivalent water thickness with a 0.5° spatial resolution[31]. These anomalies of freshwater storage are relative to the baseline period between 2004 and 2009. We adopt GRACE/GRACE-FO measurements from April 2002 to December 2021 as the target variable used to train, validate, and test our ML models.

Meteorological forcings and vegetation dynamics data. We collect data on relevant meteorological variables covering the period from January 1982 to December 2021 to train the ML models and reconstruct TWSA time series. Specifically, we identify five key meteorological predictors, including monthly averages of total precipitation (PCP), estimates of snow depth water equivalent (SWE), surface net solar radiation (SRAD), surface air temperature (TMP), and surface air relative humidity (RH). These climatic variables are crucial in influencing TWSA by regulating water supply – through PCP and SWE, which represent water sources and storage – and water demand – through SRAD, TMP, and RH, which drive evapotranspiration. Specifically, precipitation is the primary source of water for surface and groundwater, it directly impacts soil moisture and overall water availability, while snow acts as a seasonal reservoir which gradually releases water in spring. Air temperature and solar radiation regulate evaporation, transpiration, and snow/ice melt. Finally, humidity affects both evaporation and transpiration and is associated with precipitation and soil moisture levels.

To account for the influence of terrestrial ecosystems' photosynthetic activity on TWSA through transpiration, we include a long-term reconstruction of solar-induced fluorescence (SIF)[32] – a well-established proxy for photosynthesis – as a predictor of vegetation dynamics from January 1982 to December 2021. SIF has been found to effectively constrain transpiration in multiple studies[33–35]. Beyond capturing transpiration rates, SIF also reflects reductions in soil moisture and groundwater levels, correlating with the water uptake that supports vegetation growth. By including SIF in our models, we gain critical insights into how vegetation interacts with the water cycle, thereby improving the accuracy of our TWSA reconstructions.

All predictors are resampled to align with the resolution of the JPL *mascon* solutions (0.5°). To guarantee consistency in time coverage with the SIF dataset, meteorological forcings are collected only from 1982 to 2021, despite the availability of prior and subsequent data. For the same reason, we used GRACE/GRACE-FO observations only up to 2021 for model training, although GRACE-FO measurements get constantly updated till present. The meteorological and vegetation dynamics features spanning the time period between April 2002 and December 2021 are employed to train, validate, and test our models in reproducing the target variable, whereas data from January 1982 to March 2002 are used to extend the TWSA time series backward, reconstructing periods before the launch of the GRACE mission (Fig. 1). Table 1 shows a detailed description of the target variable and the meteorological and vegetation dynamics predictors used in our models.

Modeling approach. The LSTM and BiLSTM networks are employed to reconstruct and predict TWSA time series, leveraging their capabilities to dynamically learn which past information is essential and should be retained. This inherent ability to manage lagged relationships makes them particularly robust to varying lag lengths[36], enabling the models to effectively capture the temporal dynamics between the input predictors and the target variable. In particular, BiLSTM is an extension of the Unidirectional LSTM, which has the additional

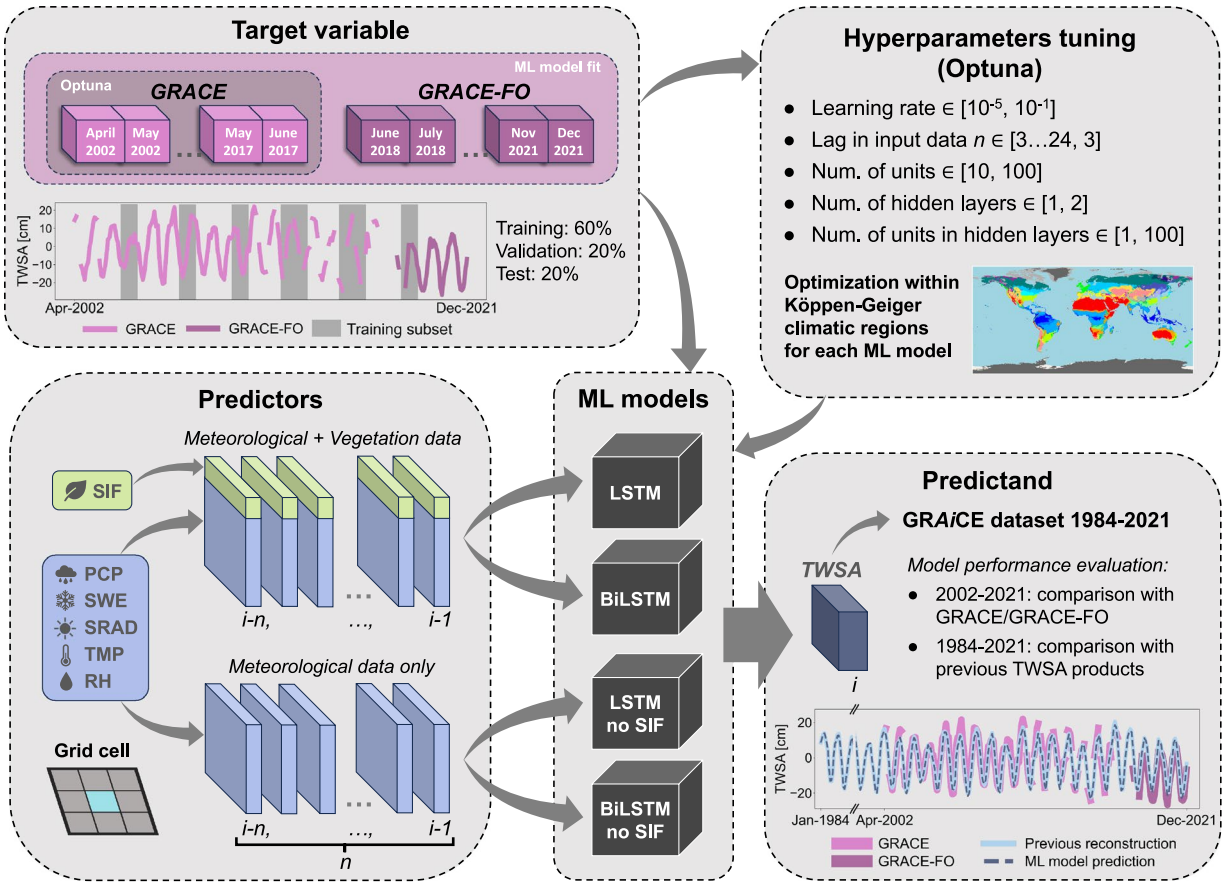

Fig. 1 Schematic representation of the overall modeling approach. For each model, the GRACE dataset (2002–2017) is used to optimize five key hyperparameters through the Optuna framework across Köppen-Geiger climatic regions, whereas both GRACE and GRACE-FO time series (2002–2021) are employed to train (60% of data), validate (20% of data), and test (20% of data) the models at the grid cell level. Each type of ML model – LSTM and BiLSTM – is fed with two sets of predictors: one including both meteorological and vegetation data, and another including meteorological data only. In the *Predictors* box, SIF is solar-induced fluorescence, PCP is total precipitation, SWE is snow depth water equivalent, SRAD is surface net solar radiation, TMP is surface air temperature, and RH is surface air relative humidity. TWSA prediction at time step *i* is obtained with predictors data from the previous *n* time steps.

Target/input data	Products	Time period	Spatial resolution
GRACE/GRACE-FO JPL-RL06.1Mv03 mascon solutions[30]	Terrestrial Water Storage Anomalies TWSA [cm eqH_2O]	April 2002–December 2021	0.5°
Meteorological forcings (ERA5-Land)[55]	Total precipitation PCP [m]	January 1982–December 2021	0.1°
	Snow depth water equivalent SWE [m eqH_2O]		
	Surface net solar radiation SRAD [J m^{-2}]		
Meteorological forcings (Essential Climate Variables from ERA5)[56]	Surface air temperature TMP [K]	January 1982–December 2021	0.25°
	Surface air relative humidity RH [%]		
Vegetation dynamics (LCSIF)[57,58]	Solar-induced fluorescence SIF [mW m^{-2} nm^{-1} sr^{-1}]	January 1982–December 2021	0.05°

Table 1. Target variable and input data of LSTM and BiLSTM models.

advantage of processing the input sequence both in forward and backward directions. As a result, BiLSTM networks use the input data (predictor) twice during their training: once in the original sequence (forward layer) and once in its reverse form (backward layer). By processing both past and future context, BiLSTM offers a more comprehensive understanding of the data and can be thought of as performing a "reanalysis" based on available observations. Thanks to its bidirectional training capability, the BiLSTM approach tends to provide more accurate time series predictions than Unidirectional LSTM[37,38].

We design two versions of the LSTM and BiLSTM networks, differentiated by the adopted predictors, leading to a total of four TWSA models (Fig. 1). The first version of the models is fed with a combination of all

meteorological forcings and the vegetation dynamics predictor (6 features). The second version includes only the meteorological forcings (5 features), and we refer to these networks as LSTMnoSIF and BiLSTMnoSIF. We adopt the Köppen-Geiger classification system[39] to group grid cells within the same climatic region and we run model fits at the grid cell level. For each grid cell, we perform a cross-validation process by running five separate model attempts, retaining the model fit with the lowest Mean Absolute Error (MAE). The fitting procedure is performed using the Adam optimization algorithm and the mean squared error loss function for all models.

As already mentioned, meteorological and vegetation dynamics data are available across different temporal periods. As a result, LSTMnoSIF and BiLSTMnoSIF models (using meteorological data alone) could potentially estimate TWSA till present, whereas LSTM and BiLSTM models (incorporating also SIF data) can estimate TWSA till 2021. The inclusion of SIF data constrains spatial coverage as well, as they allow TWSA reconstruction over 61,215 grid cells, rather than the 91,528 grid cells covered by meteorological forcings. Table S1 in the Supplementary Information shows in detail the total number of grid cells included in each climatic region, with and without the spatial constraints due to the use of SIF data among the predictors.

Model design and hyperparameters optimization. All models share the following parameters:

- *epochs*: a maximum of 1,000 complete passes through the training dataset;
- *early stopping*: training halts if no improvement is observed for 30 consecutive epochs (*patience* = 30);
- *batch size*: model's internal parameters are updated after every 32 training samples.

In particular, the *early stopping* mechanism is crucial for monitoring performance on the validation set, effectively minimizing the risk of overfitting. By halting training when further improvements are no longer observed, it prevents the model from becoming overly tailored to the training data, ensuring a balance between model accuracy and its ability to generalize to unseen data.

To further mitigate model overfitting, we conduct an extensive hyperparameters optimization. We apply the Optuna framework[29] to optimize critical hyperparameters, such as learning rate, the number of LSTM/BiLSTM units, hidden layers, and lag in input data, over different Köppen-Geiger climatic regions (see also Fig. 1). As a result, in each model, grid cells within a specific climatic region share the same optimized values for these hyperparameters. Specifically, we optimize the input variable lag (n), which represents the number of past months of predictor data used to predict the target variable at the current time step. This lag is selected from a range of 3 to 24 months, in increments of 3. Since n can extend up to 24 months (2 years) starting from 1982, our TWSA reconstructions begin in January 1984. This approach ensures that predictions provided by all models start from the same date, regardless of the chosen n value associated with each climatic region. The best combination of hyperparameters is determined using GRACE TWSA data from 2002 to 2017 as the target variable and is extracted from a total number of 100 optimization trials. A detailed list of the final hyperparameter values obtained through Optuna can be found in Table S2 of the Supplementary Information. Finally, to further improve our models' accuracy, we retune the learning rate values with Optuna in pixels presenting a low performance, i.e., a Pearson's Correlation Coefficient (PCC) lower than 0.60 between predicted and observed TWSA in the training and validation subsets. We explore a range of values around the previous learning rate, LR, i.e., between $LR \times 10^{-3}$ and $LR \times 10$ on logarithmic scale. We retain the new LR value only in those grid cells where it generates an increase in PCC for both the training and validation subsets. The percentage of grid cells with $PCC < 0.60$ found in each climatic region, along with the new learning rate values and the percentage of grid cells where they improve PCC, are listed in Table S3 (LSTM and BiLSTM models) and Table S4 (LSTMnoSIF and BiLSTMnoSIF models) of the Supplementary Information.

Data preprocessing. During models training, validation, and test, we first subtract the initial GRACE TWSA measurement (in April 2002) from all TWSA time steps at each grid cell. This value is then added back into the TWSA time series predicted by the models to maintain consistency with the original data. To minimize the effects of autocorrelation during model training, the TWSA data are divided into chunks containing a number of TWSA observations equal to or greater than the number of prior time steps used as predictors (n). These chunks are then randomly shuffled, ensuring that the model is not biased by sequential dependencies in the data. We use a constant seed value (seed = 42) to initialize the random number generator, ensuring reproducibility of the shuffling process. We allocate 60% of the chunks for training, 20% for validation, and 20% for testing, allowing the model's performance to be evaluated on unseen data while minimizing the risk of overfitting. To further standardize the input datasets, both predictors and the target variable are scaled using a *robust scaler*, which removes the median and scales the data based on their interquartile range. This method is particularly effective at mitigating the influence of outliers as it guarantees that the data are appropriately normalized before being fed into the model.

Model performance evaluation. We assess the models' performance by comparing TWSA predictions against GRACE/GRACE-FO observations, quantifying how effectively our LSTM and BiLSTM models capture observed TWSA trends.

In addition, we validate GR*Ai*CE by comparing its predictions with both previous TWSA reconstructions and independent river basin water budgets. Several reconstructions leveraging GRACE measurements have been previously developed to reproduce TWSA time series before the GRACE era and fill the gaps in GRACE/GRACE-FO datasets. In this study, we compare the performance of our models to the reference GRACE-REC reconstruction, which is generated by a statistical model trained on GRACE observations and meteorological

data spanning the period 1901–2019[2]. We select GRACE-REC because it overall outperforms other global hydrological models and accurately reflects variations in water storage driven by climate change over the past century[18,40].

We also evaluate the accuracy of our TWSA reconstructions in the period preceding GRACE observations (i.e., prior to 2002) and their capability to reproduce water budgets at the river basin scale. To achieve this, we use estimates of TWS variations from three independent products:

1. the Basin-Scale Water-Balance (BSWB) dataset by Hirschi and Seneviratne[41]. The BSWB dataset provides monthly variations of water budgets from 1979 to 2015 for 341 large catchments worldwide (areas larger than 100,000 km^2). It is obtained by integrating streamflow measurements from the Global Runoff Data Centre (GRDC) with ERA-Interim reanalysis of atmospheric moisture flux convergence. The temporal coverage of the BSWB estimates depends on the availability of runoff data and varies across different river basins. Specifically, out of the 338 river basins for which drift-corrected TWSA values are available, we select the 300 basins having at least 60 monthly estimates since 1984 to enhance the reliability of our comparison;
2. the long-term Total Water Storage Change reconstruction (TWSC-rec) by Li *et al.*[16]. The TWSC reconstruction uses a unified methodological framework combining different methods – multiple linear regression (MLR), artificial neural network (ANN), and autoregressive exogenous (ARX) approaches – applied to climatic variables and GRACE observations to estimate TWSC in 26 large river basins from 1992 to 2018;
3. the total water storage change from the SAtellite Water Cycle (SAWC) dataset by Pellet *et al.*[42]. The SAWC dataset provides total water storage changes based on river discharge data, closing the water budget over 10 sub-basins within the Amazon River basin from 1980 to 2015.

In our analysis, we use data starting from 1984 to ensure alignment with the GRA*i*CE dataset, except for the TWSC-rec product, which starts in 1992.

We adopt four performance metrics to quantify model accuracy: the Pearson's Correlation Coefficient (PCC), the Nash-Sutcliffe Efficiency (NSE), the Normalized Root Mean Squared Error (NRMSE), and the Mean Bias Error (MBE). PCC measures the linear correlation between observations and predictions and has a range of $[-1, 1]$. $PCC = \pm 1$ indicates a perfect linear correlation between observations and predictions, with the sign representing the direction of the correlation (i.e., a positive or negative correlation), while $PCC = 0$ implies that there is no linear correlation. NSE quantifies the predictive power of a model in reproducing observed data and it can assume values in the range of $(-\infty, 1]$. When $NSE = 1$ the model predictions perfectly match observed data, whereas $NSE \leq 0$ indicates that model predictions are at most as accurate as the mean value of the observations. NRMSE is the root mean squared error normalized by the range of observed data– the difference between the maximum and minimum values of observed data – and it can assume values in the range of $[0, \infty)$. The closer NRMSE is to 0, the better the model accuracy. Finally, MBE quantifies the tendency of a model to over- or underestimate observed values and is calculated as the difference between the monthly predicted and GRACE/GRACE-FO TWSA averaged along the whole time series. As MBE gets closer to 0 the model performance increases, whereas larger positive (or negative) MBE values indicate the model's overestimation (or underestimation).

Data Records

Our global TWSA reconstructions GRA*i*CE[43] are available on the Zenodo data repository and can be accessed at https://doi.org/10.5281/zenodo.10953658. The repository includes four netCDF files containing monthly TWSA estimates produced by each model from 1984 to 2021 with a 0.5° spatial resolution. An overview of the model's setup and the values of key hyperparameters tuned across the Köppen-Geiger climatic regions can be found in the Methods section and in the Supplementary Information tables. Updates of the dataset will be regularly released.

Technical Validation

Results and accuracy assessment. We assess the accuracy of our TWSA reconstructions by comparing them against GRACE/GRACE-FO observations over the 61,215 grid cells for which all models can provide TWSA estimates, i.e., grid cells covered by both meteorological and vegetation dynamics data. Overall, predictions of the global mean monthly TWSA from 2002 to 2021 – including training and test subsets – yield a PCC greater than 0.9 (Fig. 2a), with BiLSTM exhibiting the best performance ($PCC = 0.97$). Moreover, the largest differences in performance are found between models using different sets of predictors, i.e., when comparing LSTM against LSTMnoSIF and BiLSTM against BiLSTMnoSIF. This suggests the importance of including data that encode vegetation dynamics. The comparison between PCC values for the training and test subsets as well as for the whole time series (Fig. 2b) confirms that the BiLSTM model generally produces more accurate TWSA predictions.

When examining PCC distribution across the main climatic regions (Fig. 2c), we find that BiLSTMnoSIF generates slightly more accurate TWSA estimates within the continental and tropical regions, where median PCC is 0.84 and 0.91, respectively, while within the same regions BiLSTM has a median PCC of 0.81 and 0.89, respectively. However, BiLSTM outperforms the other models in the remaining climatic areas, especially in the polar region where it performs considerably better than the LSTM model.

Despite the generally high correlation, we note that PCC values in the test subset are somewhat lower than those in the training subset, which may indicate potential model overfitting during the training procedure at

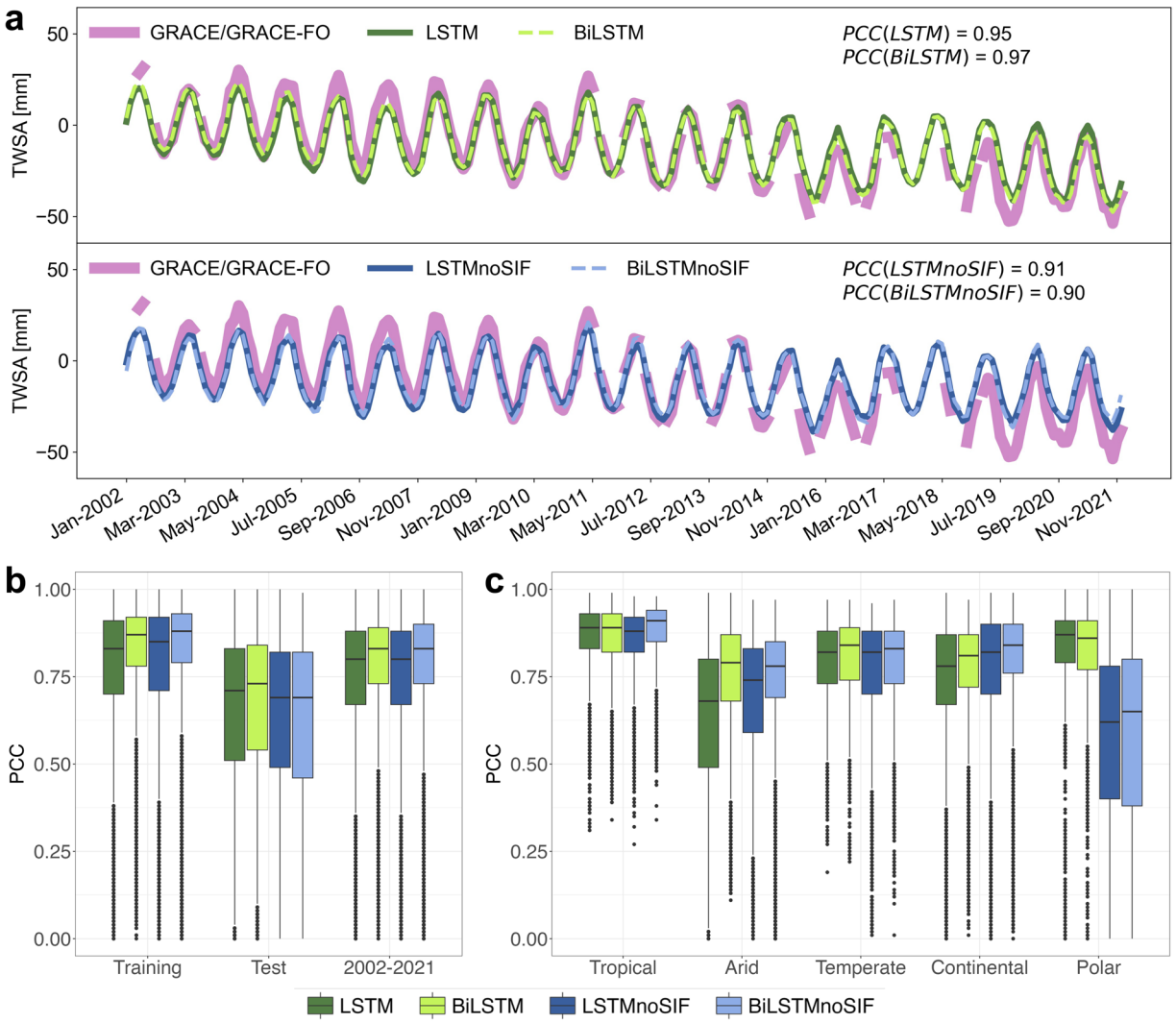

Fig. 2 Comparison of models' performance across the whole study areas and climatic regions. **(a)** Global mean monthly TWSA from 2002 to 2021 derived from GRACE/GRACE-FO observations and models predictions. On the top panel, comparison between observed values and predictions from LSTM and BiLSTM using both meteorological and vegetation predictors. On the bottom panel, comparison between observed values and predictions from LSTM and BiLSTM using only meteorological predictors (LSTMnoSIF and BiLSTMnoSIF). Pearson's Correlation Coefficient (PCC) between predicted and observed TWSA are shown in the top right corner of each panel. **(b)** Distribution of PCC values within the training and validation subsets (referred to as *Training*), the test subset, and the whole GRACE/GRACE-FO time series (2002–2021). **(c)** Distribution of PCC values across main climatic regions, computed considering the whole GRACE/GRACE-FO dataset (2002–2021). All models are evaluated over the 61,215 grid cells covered by SIF data. In each boxplot, the bottom, middle, and top of the box represent the 25th, 50th, and 75th percentiles, respectively.

the grid cell level. To investigate this, we retrain our models across climatic regions, assigning a unique model fit to all grid cells of a specific climatic zone. In this case, we include latitude and longitude values as additional predictors to distinguish the geographical location of grid cells within the same climatic region. However, we find that this approach produces generally less accurate predictions over both training and test subsets compared to the models trained at the grid cell level (Figures S1 and S2 of the Supplementary Information). Moreover, the training procedure for the LSTMnoSIF and BiLSTMnoSIF models across climatic regions fails to adequately fit certain areas, most notably the polar zones, as evidenced by the blank areas in Figure S2 of the Supplementary Information.

Given the overall similarity in accuracy across models, the results discussed hereafter focus on the BiLSTM model only, which generally performs better than other models, has greater training capabilities, and uses the full set of predictors. For reference, the performance of LSTM, LSTMnoSIF, and BiLSTMnoSIF models is reported in Figure S3, Figure S4, and Figure S5 of the Supplementary Information.

The spatial distribution of PCC, NSE, NRMSE, and MBE reveals the general capability of BiLSTM to successfully reproduce GRACE/GRACE-FO data at the grid cell level. However, it also highlights hotspot regions

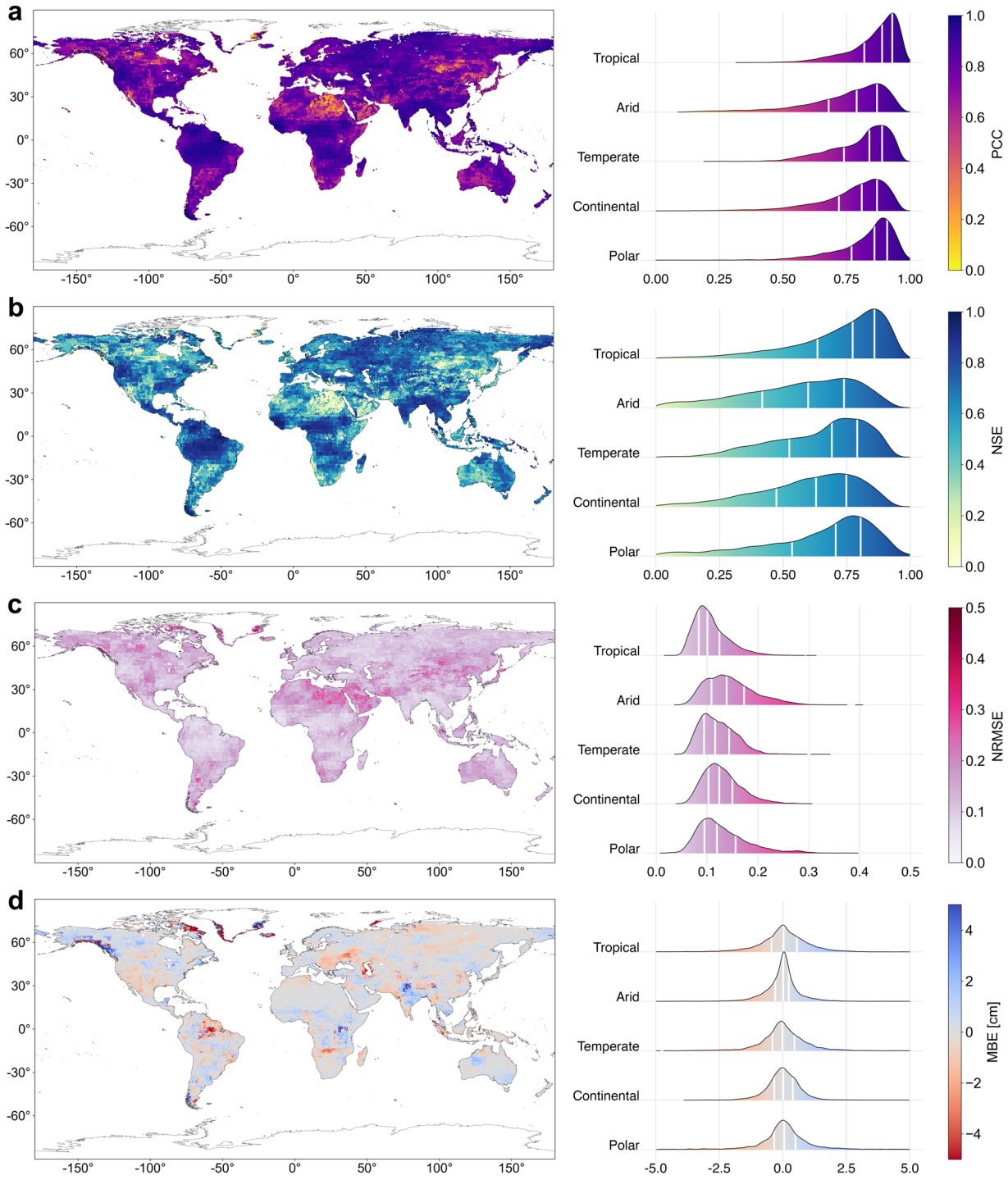

Fig. 3 Performance metrics comparing BiLSTM predictions against GRACE/GRACE-FO observations in 2002–2021 (i.e., training plus test subsets). On the left column, spatial distribution of the metrics. On the right column, distribution across main climatic regions with summary statistic lines (vertical white lines) indicating the first, second, and third quartile. **(a)** Pearson's Correlation Coefficient (PCC). **(b)** Nash-Sutcliffe Efficiency (NSE). **(c)** Normalized Root Mean Squared Error (NRMSE). **(d)** Mean Bias Error (MBE).

where our climate-driven model is not able to fully capture the observed trend in TWSA (see maps in the left column of Fig. 3). Discrepancies between GRACE/GRACE-FO measurements and BiLSTM estimates in these areas may be related to rapid variations in water availability driven by anthropogenic water withdrawals[7,44,45] or may be linked to the exclusion from the model of other sources of gravity changes, such as oil extraction. Maps of PCC, NSE, and NRMSE (left column of Fig. 3a, b, and c) consistently depict the lowest accuracy of the model in southwestern US, northeastern Africa, North China Plain, eastern and central China, northern and central India, southwestern Russia, and the Middle East. These areas have experienced significant surface water

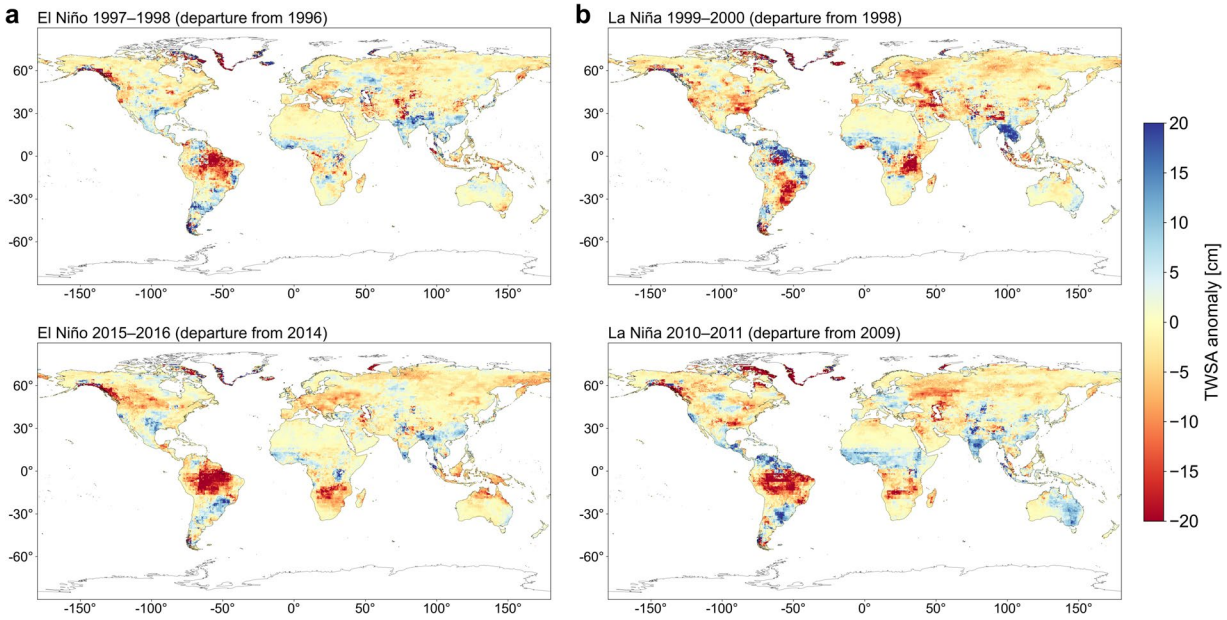

Fig. 4 Anomalies in mean TWSA during El Niño and La Niña events. TWSA values from the BiLSTM model are averaged over the June–February period, with anomalies calculated relative to the mean TWSA from January to December of the preceding year. Blue (red) shades denote excess (deficit) in TWSA. **(a)** El Niño events in 1997–1998 (top panel) and 2015–2016 (bottom panel). **(b)** La Niña events in 1999–2000 (top panel) and 2010–2011 (bottom panel).

and groundwater depletion as a consequence of anthropogenic withdrawals, which in some cases amplified the impacts of droughts, as in Southern California, northern Middle East, northwestern China, and southwestern Russia[7]. Beyond the combination of droughts and unregulated withdrawals, in the Middle East and Central Asia losses in surface water have been also attributed to river diversion/damming projects[46]. The map of MBE values (left column of Fig. 3d) identifies areas of over- and underestimation of GRACE/GRACE-FO measurements by BiLSTM, with positive and negative errors generally balancing each other. One of the most prominent TWSA overestimations are found over the Indo-Gangetic Basin. In this area, as well as in the US High Plains Aquifer and North China Plain, groundwater level has substantially declined from 1990 to 2014 because of excessive water withdrawals mainly for irrigation purposes[47–51]. Moreover, in northern and central India as well as in the North China Plain the wide extent of irrigated land, covering more than half of these regions, further suggests that the observed decrease in TWS is heavily influenced by human withdrawals[7,51]. In addition, the use of groundwater in hydraulic fracturing for unconventional oil and gas production along with the exploitation of geothermal energy could further modify groundwater flow regimes[49]. All these human-induced changes in TWS contribute to the model's overestimation, as BiLSTM does not account for such anthropogenic influences among its predictors. Finally, the largest TWSA underestimations are located in the humid regions of northern South America and in the Absheron Peninsula, where MBE reaches −8.06 and −13.58 cm, respectively.

When considering climatic conditions, the weakest model performance – reflected by lower PCC and NSE and higher NRMSE and MBE – occurs in arid regions, whereas the best results are observed in areas with a tropical climate (see right column of Fig. 3). Less accurate model predictions in arid regions might reflect the presence of anthropogenic groundwater depletion, especially for aquifers located in cultivated dryland areas where rapid and accelerating declines in groundwater level are found to be particularly evident[51].

Even though this study primarily focuses on model development and validation, we also examine the models' ability to simulate TWSA surpluses and deficits in response to major El Niño and La Niña events. Specifically, we consider the strongest events on record covered by our reconstructions: the 1997–1998 and 2015–2016 El Niño events, and the 1999–2000 and 2010–2011 La Niña events. For each event, we average TWSA predictions from June to February and compare them against a 1-year baseline period, corresponding to the year preceding each event (e.g., from January to December 1996 for the 1997–1998 El Niño). Figure 4 illustrates TWSA surpluses and deficits predicted by the BiLSTM model during these events, proving that our reconstruction effectively captures the main hydrological impacts associated with these climate extremes. Specifically, during El Niño phases, our model captures the typical increase in precipitation over the tropical Pacific, along with precipitation deficits across tropical continental regions[52,53]. Notably, South America experiences severe droughts in the northern region and intense flooding in the southern regions[12,54], whereas the 2015–2016 El Niño resulted in a remarkable TWSA decline in Southern Africa due to extreme drought conditions[12]. Conversely, La Niña impacts on TWSA typically produce opposite patterns, although they may affect different continental areas. In this case, particularly evident is the substantial increase in TWSA emerging in eastern Australia during the 2010–2011 La Niña[52].

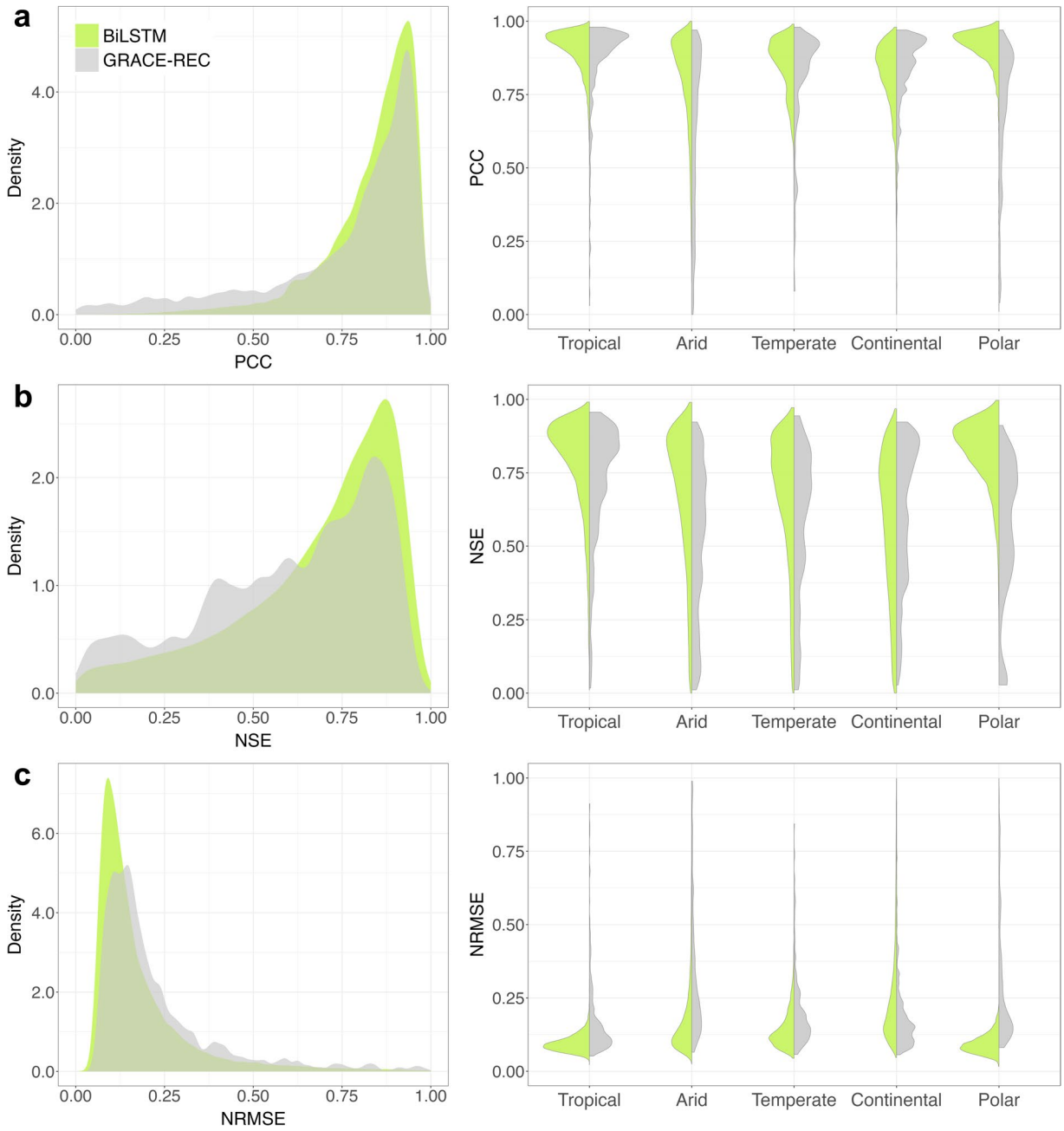

Fig. 5 Comparison between PCC, NSE, and NRMSE for BiLSTM and GRACE-REC models over their training subsets. The left column shows density plots of each metric across the whole study area, whereas on the right column, violin plots describe the distribution of the metrics values within the main climatic regions. **(a)** PCC. **(b)** NSE. **(c)** NRMSE.

Comparison with existing TWSA reconstruction. We compare TWSA predictions from the GRA*i*CE dataset with those from the GRACE-REC reconstruction to evaluate which dataset better reproduces GRACE/GRACE-FO measurements. To guarantee consistency between the datasets, we use the GRACE-REC monthly reconstruction produced with ERA5 forcing data and trained on JPL *mascon* solutions. Since GRACE-REC provides de-seasonalized TWSA, we reintroduce the GRACE-based TWS seasonal cycle as provided by Humphrey and Gudmundsson[2]. The datasets comparison is carried out over their respective training subsets, i.e., over GRACE time series from 2002 to 2017 for GRACE-REC data and over 60% of shuffled months between 2002 and 2021 for the BiLSTM model. We evaluate the overall performance of the TWSA reconstructions using the PCC, NSE, and NRMSE metrics. As illustrated in Fig. 5, the BiLSTM model outperforms GRACE-REC reconstruction both at the global scale (left column of Fig. 5a,b,c) and across climatic regions (right column of Fig. 5a,b,c).

A key distinction between the models should be highlighted: GRACE-REC has been calibrated specifically to reproduce de-seasonalized TWSA time series, whereas our GRA*i*CE dataset is trained on the original GRACE

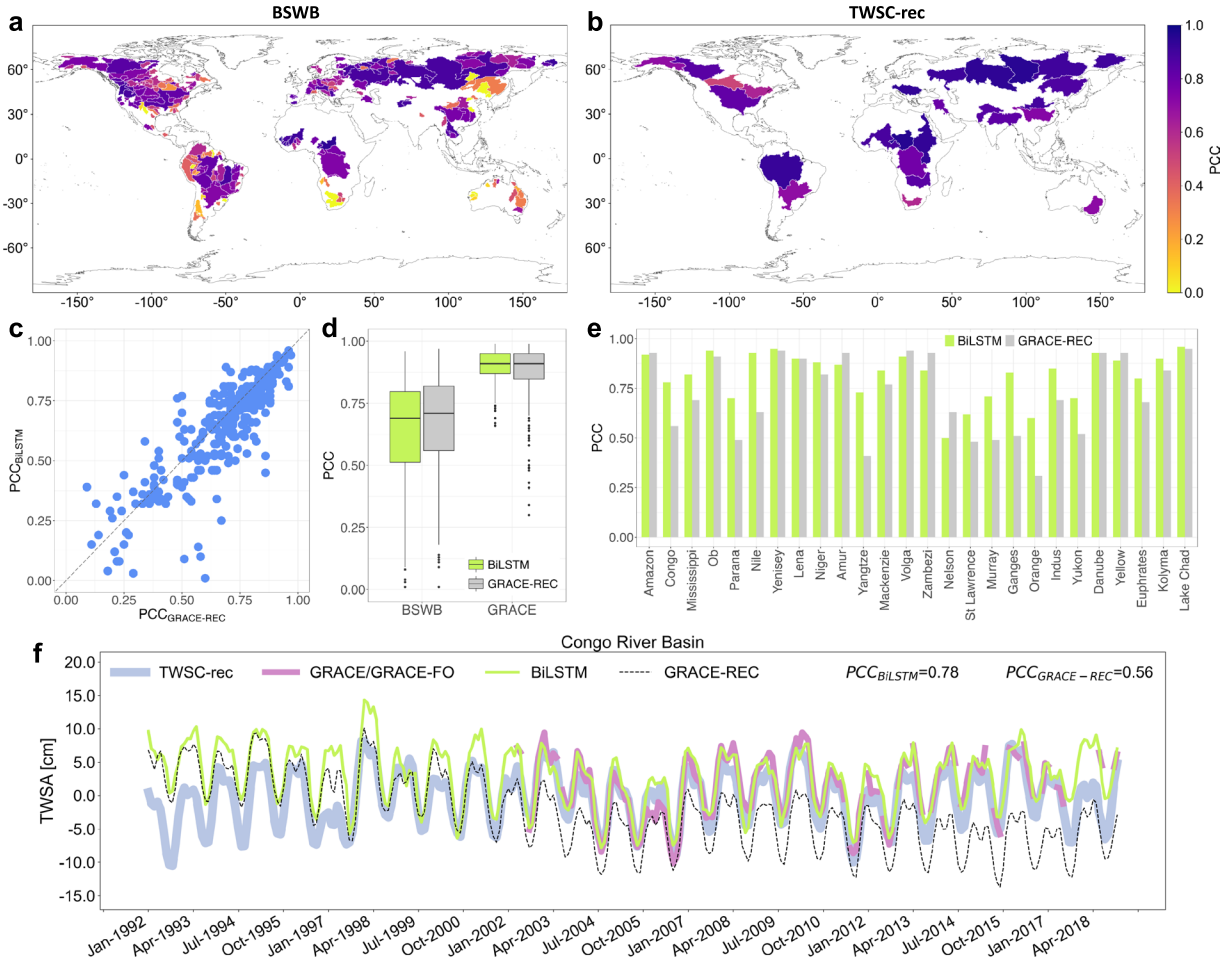

Fig. 6 Comparison of river basin water budgets from BiLSTM and GRACE-REC models against the BSWB (drift-corrected data) and the TWSC-rec datasets. **(a)** Spatial distribution of Pearson's Correlation Coefficient (PCC) between BSWB-based TWSA and BiLSTM TWSA across 300 large watersheds. **(b)** As in panel a, but showing PCC values between the TWSC-rec product and the BiLSTM model across 26 major river basins. **(c)** Scatterplot comparing the correlation of BSWB data with GRACE-REC (x-axis) and BiLSTM (y-axis) across the basins shown in panel a. **(d)** Distribution of PCC values for BiLSTM and GRACE-REC models against the BSWB dataset and GRACE observations across the basins shown in panel a. **(e)** PCC comparison between BiLSTM and GRACE-REC models against the TWSC-rec product across 26 major river basins. **(f)** Time series of water budgets for the Congo River Basin (1984–2018) as provided by the TWSC-rec product, BiLSTM and GRACE-REC reconstructions, and GRACE/GRACE-FO observations. PCC values between TWSC-rec and the BiLSTM and GRACE-REC reconstructions are reported in the top right corner of the panel.

observations, which include both seasonal and non-seasonal components. Moreover, it is worth noting that the seasonal component reintroduced into GRACE-REC is derived directly from GRACE observations. As Humphrey and Gudmundsson[2] highlighted, this approach assumes that the seasonal cycle remains unchanged throughout the reconstruction period, which may be unrealistic. Moreover, reintroducing the observed seasonal cycle could artificially improve GRACE-REC performance. Therefore, we also compared deseasonalized TWSA from both datasets. Even in this case, the BiLSTM model yields TWSA predictions that are closer to GRACE/GRACE-FO observations.

Comparison with basin-scale water budgets. We evaluate the performance of the GRA*i*CE dataset in reconstructing TWSA prior to the GRACE era (before 2002) by comparing it against several water budget datasets at the river basin scale. The first comparison uses the BSWB dataset[41]. Since the temporal coverage of the BSWB dataset depends on the availability of runoff data, we focus on 300 basins where drift-corrected TWSA is available for at least 60 months since 1984. For these basins, we calculate TWS variations from both the GRA*i*CE (BiLSTM model) and GRACE-REC reconstructions. As shown in Fig. 6a, the water budgets derived from the BiLSTM model are generally well correlated with BSWB data. In the 76 river basins exhibiting weak correlations (PCC < 0.5), we compare GRACE observations with the BSWB dataset as well. This comparison is feasible in 49 basins where the BSWB time series overlaps with GRACE data, and we find that about half of these basins also show weak correlations between GRACE measurements and BSWB. GRACE-REC data exhibit correlation levels comparable to those of the BiLSTM model in the 300 basins (Fig. 6c).

To further examine the accuracy of TWSA estimates, we compare BiLSTM and GRACE-REC water budgets with GRACE observations across the same 300 river basins (Fig. 6d). As expected, correlations are significantly higher in this comparison, as both models were trained on GRACE data. However, this comparison still demonstrates that BiLSTM can accurately reproduce post-2002 TWSA at the river basin scale.

The comparison with the TWSC-rec[16] product yields different results. This analysis focuses on 26 larger river basins compared to those in the BSWB dataset. As illustrated in Fig. 6b, the BiLSTM model consistently captures TWSC-rec data, with PCC reaching or exceeding 0.5 in nearly all basins. Moreover, the BiLSTM model generally shows stronger correlations with TWSC-rec data than GRACE-REC (Fig. 6e). For example, in the Congo River basin (Fig. 6f), the BiLSTM model aligns closely with TWSC-rec estimates, achieving a PCC of 0.78, while GRACE-REC reaches a PCC of 0.56.

These comparisons with BSWB and TWSC-rec products suggest that while both the BiLSTM and GRACE-REC models generally perform well at the river basin scale, their accuracy may vary depending on the region and the level of spatial aggregation.

Finally, we also calculate the TWS flux for the Amazon River basin using a second-order central finite difference scheme applied to spatially averaged TWSA values, comparing the results against the SAWC dataset[42]. The high level of agreement between the BiLSTM predictions and the water storage change estimates from SAWC (Figure S6), as evidenced by a high PCC value, further reinforces the robustness of the GR*Ai*CE dataset in reconstructing TWSA prior to GRACE era and its effectiveness for basin-scale applications.

The GR*Ai*CE dataset provides four global long-term TWSA reconstructions from 1984 to 2021, with a 0.5° spatial resolution, generated using climate-driven ML models. These reconstructions deliver accurate estimates of TWS changes, as proved by comparisons with GRACE/GRACE-FO measurements and previous TWSA products. However, certain limitations should be acknowledged. The reliability of GR*Ai*CE prior to the GRACE era (i.e., before 2002) may be lower due to the lack of direct observational data for that period. Furthermore, the accuracy of the models heavily depends on the quality of input data, such as climate variables (e.g., precipitation and temperature) and vegetation data (i.e., solar-induced fluorescence). Although GR*Ai*CE effectively captures large-scale TWSA patterns, it may struggle to account for the complexity of local hydrological extremes, particularly in regions with high variability. The 0.5° spatial resolution may also limit its ability to model highly localized processes or extreme events in smaller basins. Despite these challenges, GR*Ai*CE serves as a valuable tool for detecting climate-driven changes in TWSA, offering critical insights into long-term shifts in the global hydrological cycle.

Usage Notes

GR*Ai*CE reconstructions from 1984 to 2021 with a 0.5° spatial resolution are available for download on Zenodo at https://doi.org/10.5281/zenodo.10953658. Data are stored as 4 separate netCDF files, each containing the monthly reconstructions of TWSA data produced by LSTM, BiLSTM, LSTMnoSIF, and BiLSTMnoSIF models.

Code availability

All codes for preparing input data, developing models, and analyzing their performance are written in Python and are accessible at: https://github.com/Pallre/GRAiCE.git. Given the large size of datasets and the extensive computational time required, the codes for hyperparameter tuning with Optuna and model training were executed on the Ginsburg High Performance Computing cluster at Columbia University.

Received: 24 April 2024; Accepted: 3 January 2025;
Published online: 25 January 2025

References

1. Pokhrel, Y. *et al*. Global terrestrial water storage and drought severity under climate change. *Nat. Clim. Chang.* **11**, 226–233 (2021).
2. Humphrey, V. & Gudmundsson, L. GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. *Earth Syst. Sci. Data* **11**, 1153–1170 (2019).
3. Eicker, A., Forootan, E., Springer, A., Longuevergne, L. & Kusche, J. Does GRACE see the terrestrial water cycle "intensifying"? *J. Geophys. Res. Atmos.* **121**, 733–745 (2016).
4. Rodell, M. & Reager, J. T. Water cycle science enabled by the GRACE and GRACE-FO satellite missions. *Nat. Water* **1**, 47–59 (2023).
5. Sun, A. Y., Scanlon, B. R., Save, H. & Rateb, A. Reconstruction of GRACE total water storage through automated machine learning. *Water Resour. Res.* **57**, e2020WR028666 (2021).
6. Hosseini-Moghari, S. M. *et al*. Using GRACE satellite observations for separating meteorological variability from anthropogenic impacts on water availability. *Sci. Rep.* **10**, 15098 (2020).
7. Rodell, M. *et al*. Emerging trends in global freshwater availability. *Nature* **557**, 651–659 (2018).
8. Scanlon, B. R. *et al*. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. *P. Natl. Acad. Sci. USA* **115**, 1080–1089 (2018).
9. Sun, A. Y. *et al*. Combining physically based modeling and deep learning for fusing GRACE satellite data: Can we learn from mismatch? *Water Resour. Res.* **55**, 1179–1195 (2019).
10. Ferreira, V. *et al*. Estimating GRACE terrestrial water storage anomaly using an improved point mass solution. *Sci. Data* **10**, 234 (2023).
11. Gou, J. & Soja, B. Global high-resolution total water storage anomalies from self-supervised data assimilation using deep learning algorithms. *Nat. Water* **2**, 139–150 (2024).
12. Li, F., Kusche, J., Chao, N., Wang, Z. & Löcher, A. Long-term (1979-present) total water storage anomalies over the global land derived by reconstructing GRACE data. *Geophys. Res. Lett.* **48**, e2021GL093492 (2021).
13. Mo, S. *et al*. Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during GRACE and GRACE-FO gap. *J. Hydro.* **604**, 127244 (2022).

14. Yang, X., You, W., Tian, S., Jiang, Z. & Wan, X. A two-step linear model to fill the data gap between GRACE and GRACE-FO terrestrial water storage anomalies. *Water Resour. Res.* **59**, e2022WR034139 (2023).
15. Yang, X., Tian, S., You, W. & Jiang, Z. Reconstruction of continuous GRACE/GRACE-FO terrestrial water storage anomalies based on time series decomposition. *J. Hydro.* **603**, Part B, 127018 (2021).
16. Li, F. *et al*. 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 (2020).
17. Rateb, A., Sun, A., Scanlon, B. R., Save, H. & Hasan, E. Reconstruction of GRACE mass change time series using a Bayesian framework. *Earth Space Sci.* **9**, e2021EA002162 (2022).
18. Yin, J. *et al*. GTWS-MLrec: global terrestrial water storage reconstruction by machine learning from 1940 to present. *Earth Syst. Sci. Data* **15**, 5597–5615 (2023).
19. Xiong, J., Guo, S., Abhishek, J., Li & Yin, J. A novel standardized drought and flood potential index based on reconstructed daily GRACE data. *J. Hydrometeor.* **23**, 1419–1438 (2022).
20. Zhang, Q., Pan, Y., Zhang, C. & Gong, H. Deriving the terrestrial water storage anomaly from GRACE spherical harmonic coefficients using a convolutional neural network. *Earth Space Sci.* **10**, e2023EA003023 (2023).
21. Van Houdt, G., Mosquera, C. & Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **53**, 5929–5955 (2020).
22. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. LSTM: A search space odyssey. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2222–2232 (2017).
23. Hochreiter, S. & Schmidhuberm, J. Long Short-Term Memory. *Neural Comput.* **9**, 1735–1780 (1997).
24. Li, B. *et al*. Improving LSTM hydrological modeling with spatiotemporal deep learning and multi-task learning: A case study of three mountainous areas on the Tibetan Plateau. *J. Hydro.* **620**, Part A, 129401 (2023).
25. Wilbrand, K. *et al*. Predicting streamflow with LSTM networks using global datasets. *Front. Water* **5**, 1166124 (2023).
26. Yin, H. *et al*. Rainfall-runoff modeling using long short-term memory based step-sequence framework. *J. Hydro.* **610**, 127901 (2022).
27. Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **22**, 6005–6022 (2018).
28. Zhang, S., Wang, Y., Wang, Y., Li, Z. & Hou, Y. Spatiotemporal evolution and influencing mechanisms of ecosystem service value in the Tarim River Basin, Northwest China. *Remote Sens.* **15**, 591 (2023).
29. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*, 2623–2631, https://doi.org/10.1145/3292500.3330701 (2019).
30. Wiese, D. N., Yuan, D.-N., Boening, C., Landerer, F. W. & Watkins, M. M. JPL GRACE and GRACE-FO mascon ocean, ice, and hydrology equivalent water height CRI filtered. Ver. RL06.1Mv03. *PO.DAAC, CA, USA* https://doi.org/10.5067/TEMSC-3JC63 (2023).
31. Watkins, M. M., Wiese, D. N., Yuan, D. N., Boening, C. & Landerer, F. W. Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons: improved gravity observations from GRACE. *J. Geophys. Res. Solid Earth* **120**, 4 (2015).
32. Zhang, Y., Joiner, J., Alemohammad, S. H., Zhou, S. & Gentine, P. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* **15**, 5779–5800 (2018).
33. Zhang, Q. *et al*. Solar-induced chlorophyll fluorescence sheds light on global evapotranspiration. *Remote Sens. Environ.* **305**, 114061 (2024).
34. Jonard, F. *et al*. Value of sun-induced chlorophyll fluorescence for quantifying hydrological states and fluxes: current status and challenges. *Agric. For. Meteorol.* **291**, 108088 (2020).
35. Maes, W. H. *et al*. Sun-induced fluorescence closely linked to ecosystem transpiration as evidenced by satellite data and radiative transfer models. *Remote Sens. Environ.* **249**, 112030 (2020).
36. Hochreiter, S. & Schmidhuber, J. LSTM can solve hard long time lag problems. *Adv. Neur. Inf. Process. Syst.* **9** (1996).
37. Abduljabbar, R. L., Dia, H. & Tsai, P. W. Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data. *Sci. Rep.* **11**, 23899 (2021).
38. Siami-Namini, S., Tavakoli, N. & Namin, A. S. The performance of LSTM and BiLSTM in forecasting time series. *2019 IEEE International Conference on Big Data (Big Data)*, 3285–3292, https://doi.org/10.1109/BigData47090.2019.9005997 (2019).
39. Beck, H. *et al*. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Sci. Data* **5**, 180214 (2018).
40. Deng, S., Liu, Y. & Zhang, W. A comprehensive evaluation of GRACE-like terrestrial water storage (TWS) reconstruction products at an interannual scale during 1981–2019. *Water Resour. Res.* **59**, e2022WR034381 (2023).
41. Hirschi, M. & Seneviratne, S. I. Basin-scale water-balance dataset (BSWB): an update. *Earth Syst. Sci. Data* **9**, 251–258 (2017).
42. Pellet, V., Aires, F. & Yamazaki, D. Coherent satellite monitoring of the water cycle over the Amazon. Part 2: Total water storage change and river discharge estimation. *Water Resour. Res.* **57**, e2020WR028648 (2021).
43. Palazzoli, I., Ceola, S. & Gentine, P. GRAiCE: terrestrial water storage anomalies reconstructions. *Zenodo* https://doi.org/10.5281/zenodo.10953658 (2024).
44. Jing, W. *et al*. Extending GRACE terrestrial water storage anomalies by combining the random forest regression and a spatially moving window structure. *J. Hydro.* **590**, 125239 (2020).
45. Zhang, Y. *et al*. Southern Hemisphere dominates recent decline in global water availability. *Science* **382**, 579–584 (2023).
46. Pekel, J. F. *et al*. High-resolution mapping of global surface water and its long-term changes. *Nature* **540**, 418–422 (2016).
47. Ambika, A., Wardlow, B. & Mishra, V. Remotely sensed high resolution irrigated area mapping in India for 2000 to 2015. *Sci. Data* **3**, 160118 (2016).
48. Chen, C. *et al*. China and India lead in greening of the world through land-use management. *Nat. Sustain.* **2**, 122–129 (2019).
49. Kuang, X. *et al*. The changing nature of groundwater in the global water cycle. *Science* **383**, eadf0630 (2024).
50. McDermid, S. *et al*. Irrigation in the Earth system. *Nat. Rev. Earth Environ.* **4**, 435–453 (2023).
51. Jasechko, S. *et al*. Rapid groundwater decline and some cases of recovery in aquifers globally. *Nature* **625**, 715–721 (2024).
52. Chen, J. *et al*. Applications and challenges of GRACE and GRACE Follow-On satellite gravimetry. *Surv. Geophys.* **43**, 305–345 (2022).
53. Phillips, T., Nerem, R. S., Fox-Kemper, B., Famiglietti, J. S. & Rajagopalan, B. The influence of ENSO on global terrestrial water storage using GRACE. *Geophys. Res. Lett.* **39**, L16705 (2012).
54. Moura, M. M. *et al*. Relation of El Niño and La Niña phenomena to precipitation, evapotranspiration and temperature in the Amazon basin. *Sci. Total Environ.* **651**, 1639–1651 (2019).
55. Hersbach, H. *et al*. Essential climate variables for assessment of climate variability from 1979 to present. *Copernicus Climate Change Service (C3S) Data Store (CDS)*. https://cds.climate.copernicus.eu/cdsapp#!/dataset/ecv-for-climate-change?tab=overview (2018).
56. Muñoz Sabater, J. ERA5-Land monthly averaged data from 1950 to present. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.68d2bb30 (2019).
57. Fang, J., & Lian, X. Reconstructed global long-term contiguous solar-induced fluorescence (LCSIF), AVHRR period (1982-2000) (2.0), *Zenodo*, https://doi.org/10.5281/zenodo.7916851 (2023).
58. Fang, J. & Lian, X. Reconstructed global long-term contiguous solar-induced fluorescence (LCSIF), MODIS period (2001-2021) (2.0). *Zenodo* https://doi.org/10.5281/zenodo.7916879 (2023).

Acknowledgements
This research was supported by the Horizon 2020 Marie Skłodkowska-Curie Research and Innovation Staff Exchange (H2020-MSCA-RISE-2017) grant in the framework of the GHAIA Project (grant number 777822). P.G. would like to acknowledge support is by the National Science Foundation (NSF) Science and Technology Center (STC) Learning the Earth with Artificial Intelligence and Physics (LEAP), Award # 2019625-STC. I.P. and S.C. acknowledge that this study was carried out within the RETURN Extended Partnership funded by the European Union Next-GenerationEU (National Recovery and Resilience Plan—NRRP, Mission 4, Component 2, Investment 1.3—D.D. 1243 2/8/2022, PE00000005).

Author contributions
I.P. and P.G. designed the study. I.P. collected and processed data for analysis, developed and tested ML models, performed the analysis, and wrote the manuscript. P.G. and S.C. provided guidance and supervision. All authors reviewed and edited the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04403-3.

Correspondence and requests for materials should be addressed to I.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.