Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during GRACE and GRACE-FO gap

Mo, Shaoxing; Zhong, Yulong; Forootan, Ehsan; Mehrnegar, Nooshin; Yin, Xin; Wu, Jichun; Feng, Wei; Shi, Xiaoqing

Published in:
Journal of Hydrology

DOI (link to publication from Publisher):
10.1016/j.jhydrol.2021.127244

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Journal of Hydrology 604 (2022) 127244

Contents lists available at ScienceDirect

Journal of Hydrology

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

Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during GRACE and GRACE-FO gap

Shaoxing Mo [a], Yulong Zhong [b], Ehsan Forootan [c], Nooshin Mehrnegar [c], Xin Yin [d], Jichun Wu [a,*], Wei Feng [e,f,*], Xiaoqing Shi [a]

[a] Key Laboratory of Surficial Geochemistry of Ministry of Education, School of Earth Sciences and Engineering, Nanjing University, Nanjing, China
[b] School of Geography and Information Engineering, China University of Geosciences (Wuhan), Wuhan, China
[c] Geodesy Group, Department of Planning, Aalborg University, Aalborg, Denmark
[d] State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Nanjing Hydraulic Research Institute, Nanjing, China
[e] School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai, China
[f] State Key Laboratory of Geodesy and Earth's Dynamics, Institute of Geodesy and Geophysics, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan, China

ARTICLE INFO

This manuscript was handled by Marco Borga, Editor-in-Chief, with the assistance of Di Long, Associate Editor

Keywords:
GRACE
Bayesian convolutional neural network
Gap filling
ERA5
Deep learning

ABSTRACT

The monthly terrestrial water storage anomaly (TWSA) observations during the gap period between the Gravity Recovery and Climate Experiment (GRACE) satellite and its Follow-On (GRACE-FO) are missing, leading to discontinuity in the time series, and thus, impeding full utilization and analysis of the data. Despite previous efforts undertaken to tackle this issue, a gap-filling TWSA product with desirable accuracy at a global scale is still lacking. In this study, a straightforward and hydroclimatic data-driven Bayesian convolutional neural network (BCNN) is proposed to bridge this gap. Benefiting from the excellent capability of BCNN in handling image data and the integration of recent deep learning advances (including residual-skip connections and spatial-channel attentions), the proposed method can automatically extract informative features for TWSA predictions from multiple predictor data. The BCNN predictions are compared with reanalyzed/simulated TWSA, Swarm solution, and the TWSA prediction products generated by three recent studies, using commonly used accuracy metrics. Results demonstrate BCNN's superior performance to obtain higher-quality TWSA predictions, particularly in relatively arid regions. Additionally, a comparison with two independent datasets at the basin scale further suggests that the BCNN-infilled TWSA is reliable to bridge the gap and enhance data consistency. Our gap-filling product can ultimately contribute to correcting the bias in long-term trend estimates, maintaining the continuity of TWSA time series and thus benefiting subsequent applications desiring continuous data records.

1. Introduction

The Gravity Recovery and Climate Experiment (GRACE) satellite and its successor GRACE Follow-On (GRACE-FO) provide unprecedentedly accurate observations of the spatiotemporal dynamics of terrestrial water storage anomaly (TWSA). These TWSA observations have been widely utilized, often together with hydrological models, to assess water cycle, droughts and floods, and impacts of changing climate on terrestrial water storage (e.g., AghaKouchak et al., 2015; Famiglietti et al., 2011; Feng et al., 2018; Gentine et al., 2019; Long et al., 2013; Rateb et al., 2020; Richey et al., 2015; Rodell et al., 2018; Soltani et al., 2021; Tapley et al., 2019; Yan et al., 2021; Yin et al., 2021; Zhong et al., 2018). These studies have substantially augmented our knowledge toward the complex hydrological systems, consequently informing restricted water resources management.

Initially, GRACE was targeted to cover a 5-year period, which was exceeded by 10 years to October 2017. Its follow-on GRACE-FO was then launched in May 2018. This has led to approximately one year of data gap (July 2017–May 2018) (Li et al., 2020), leading to discontinuity in the time series and thus impeding full utilization and analysis of the data (Sun et al., 2020; Yi and Sneeuw, 2021). Particularly, considering that the TWSA observations are usually assimilated into hydrological models

* Corresponding authors at: Key Laboratory of Surficial Geochemistry of Ministry of Education, School of Earth Sciences and Engineering, Nanjing University, Nanjing, China (Jichun Wu). School of Geospatial Engineering and Science, Sun Yat-Sen University, Zhuhai, China (Wei Feng).
E-mail addresses: smo@nju.edu.cn (S. Mo), zhongyl@cug.edu.cn (Y. Zhong), efo@plan.aau.dk (E. Forootan), MehrnegarN@cardiff.ac.uk (N. Mehrnegar), xiny@smail.nju.edu.cn (X. Yin), jcwu@nju.edu.cn (J. Wu), fengwei@sysu.edu.cn (W. Feng), shixq@nju.edu.cn (X. Shi).

https://doi.org/10.1016/j.jhydrol.2021.127244
Received 22 September 2021; Received in revised form 14 November 2021; Accepted 21 November 2021
Available online 28 November 2021

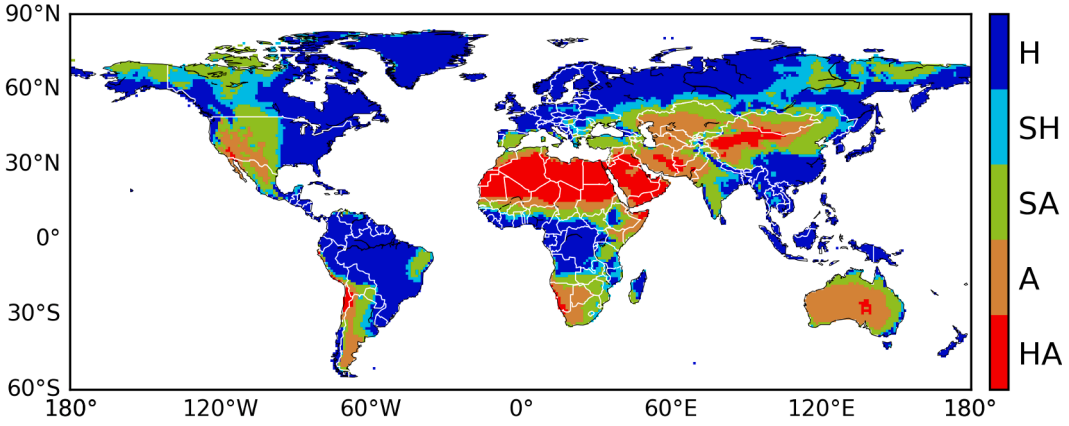

Fig. 1. The hyper-arid (HA; AI (aridity index) < 0.05), arid (A; 0.05 ⩽ AI < 0.2), semi-arid (SA; 0.2⩽AI< 0.5), semi-humid (SH; 0.5 ⩽ AI < 0.65), and humid (H; AI ⩾ 0.65) regions in the globe. The data are downloaded from https://doi.org/10.6084/m9.figshare.7504448.v3.

for higher reliability (Li et al., 2019; Mehrnegar et al., 2020; Mehrnegar et al., 2021; Nie et al., 2019; Soltani et al., 2021; Yin et al., 2020; Zaitchik et al., 2008), discontinuity in the time series observations may introduce significant biases and uncertainties in the model predictions and consequently mislead decision making (Sun et al., 2020). This is especially the case when there existed climate extremes during the gap as they usually cause abnormal changes in the TWSA signals. Two alternatives to the GRACE satellites that provide measurements of TWSA are the Geodetic Satellite Laser Ranging (SLR) and the European Space Agency (ESA)'s Swarm Earth explorer mission (Friis-Christensen et al., 2008). While there have been studies that bridged the gap between GRACE and GRACE-FO missions based on the SLR and Swarm data (e.g., Forootan et al., 2020; Meyer et al., 2019; Richter et al., 2021), their inherent low resolution relative to GRACE (-FO) limits the gap-filling quality. Bridging this gap with comparably accurate TWSA predictions is thus of crucial importance for practical applications.

There have been many efforts undertaken to reconstruct the missing GRACE TWSA signals at regional or global scales using data-driven methods (e.g., Ahmed et al., 2019; Forootan et al., 2014; Forootan et al., 2020; Humphrey et al., 2017; Humphrey and Gudmundsson, 2019; Jing et al., 2020; Li et al., 2020; Li et al., 2021; Long et al., 2014; Sun et al., 2019; Sun et al., 2020; Sun et al., 2020; Wang et al., 2021; Yi and Sneeuw, 2021). For example, Long et al. (2014) utilized artificial neural network (ANN) to learn the relationship between GRACE TWSA and hydroclimatic variables to reconstruct the basin-averaged TWSA. A similar method was then proposed in Sun et al. (2020) to reconstruct the gridded TWSA at the global scale. More recently, Li et al. (2020, 2021) reconstructed the gridded TWSA by integrating machine learning and spatiotemporal decomposition techniques to extract and leverage the spatiotemporally correlated features of data for higher performance. While these studies have generally obtained desired performances in specific humid regions, a gap-filling product with generally high accuracy over the global scale (especially in the relatively arid regions; see Fig. 1 for the climate regions) is still lacking, calling for innovative solutions.

Filling the gap between GRACE and GRACE-FO at a global scale is challenging due to: (1) the difficulty in capturing the long-term TWSA trends caused by anthropogenic activities and/or climate change, which accounts for the decreased performance of existing methods in the relatively arid regions (Humphrey and Gudmundsson, 2019; Li et al., 2020; Sun et al., 2020), and (2) the lack of efficient algorithms to extract informative features from multi-source predictor data and suppress unnecessary ones for TWSA predictions, so that the prediction models can achieve higher accuracy (Li et al., 2020; Sun et al., 2020). Recent years, have witnessed a rapid development of deep learning and its impressive performance in a variety of applications (Gu et al., 2018; LeCun et al., 2015). The advent of deep learning provides new opportunities for addressing many long-standing challenges facing research in hydrology and Earth sciences (Reichstein et al., 2019; Shen, 2018; Sun and Scanlon, 2019). Thus, in this study, we aim to develop a new Bayesian convolutional neural network (BCNN), driven by hydroclimatic inputs, to bridge the GRACE and GRACE-FO gap. The two mentioned challenges, regarding to filling the gap between GRACE and GRACE-FO will be addressed by (1) using the long-term trends retrieved from the available GRACE (-FO) data in the pre- and post-gap periods, and (2) by developing a deep learning-based prediction model.

One superior advantage of convolutional neural network (CNN) over the traditional statistical and machine learning methods (including the deep fully-connected ANN with multiple layers) employed in previous GRACE/GRACE-FO gap-filling studies is its ability to directly take raw data fields (images) as inputs without requiring additional prepossessing (Gu et al., 2018; Mo et al., 2019; Mo et al., 2019; Mo et al., 2020; Shen, 2018; Sun et al., 2019). This property makes CNN very suitable for handling computer vision tasks involving image data (Gu et al., 2018; LeCun et al., 2015). CNN can fully extract and utilize the spatially correlated features associated with images for predictions. Sun et al. (2019) applied CNN for prediction of TWSA fields in India and it outperformed the hydrological models in providing more accurate TWSA estimates. To the best of our knowledge, we present the first attempt to employ CNN for filling the GRACE and GRACE-FO gap at the global scale. In BCNN, the global-scale hydroclimatic predictor fields and target GRACE TWSA fields are treated as images to leverage CNN's superior capability in image processing. To obtain improved gap-filling results, the development of our BCNN model integrates the recent advances in deep learning, including the channel and spatial attention mechanisms (Woo et al., 2018), residual (He et al., 2016) and skip (Ronneberger et al., 2015) connection modules, and Bayesian training strategy (Liu and Wang, 2016; Zhu and Zabaras, 2018). Particularly, the Bayesian training strategy enables BCNN to quantify the predictive uncertainties.

To evaluate BCNN's gap-filling results, we conduct comparisons with ERA5-land-reanalyzed TWSA (Muñoz Sabater, 2019), Noah-simulated TWSA (Rodell et al., 2004), and the prediction products generated by three recent studies (Humphrey and Gudmundsson, 2019; Li et al., 2021; Sun et al., 2020) at a grid cell scale, and with the Swarm solution (Bezděk et al., 2016) at a basin scale for 15 world's major river basins. It will be shown that the combination of residual-skip connections and spatial-channel attentions enables BCNN to automatically and efficiently extract informative features from multi-source data and, consequently, achieve a clearly improved performance in filling the gap. The gap-filling quality is further validated through comparison with two independent standardized datasets, namely the CPC (Climate Prediction Center) soil moisture (van den Dool et al., 2003) and Noah TWSA (Rodell et al., 2004) at the basin scale (The standardization here is to exclude the influence of amplitude and magnitude differences (Scanlon

et al., 2019)).

The rest of the paper is organized as follows. The data used are described in Section 2. In Section 3, the BCNN model, including its architecture design and training, is introduced. In Section 4, we evaluate BCNN's predictions by comparing with multiple TWSA products. Finally, the conclusions are summarized in Section 5.

2. Data and processing

2.1. GRACE TWSA data

The GRACE mascon product released by the Jet Propulsion Laboratory (JPL), which has a spatial resolution of $0.5^\circ \times 0.5^\circ$ (Watkins et al., 2015), is used in this study. The JPL GRACE TWSA data are provided as anomalies with respect to the 2004 to 2009 mean. The observations cover two periods, that is, April 2002–June 2017 (GRACE mission) and June 2018-present (GRACE-FO mission), with a 11-month gap in between. In addition, there are some one- or two-month gaps within each mission, these gaps are interpolated using the data of neighboring months. Our aim is to fill the 11-month gap (i.e., July 2017–May 2018) for the land areas with the BCNN method. To facilitate the comparison with previous GRACE prediction studies, we resampled averagely the data to $1^\circ \times 1^\circ$ grids.

2.2. ERA5-land driving data

The driving data used to predict the GRACE TWSA are extracted from the ERA5-land (ERA5L) climate reanalysis dataset released by the European Centre for Medium-Range Weather Forecasts (Muñoz Sabater, 2019). The data are provided at a spatial resolution of $0.1^\circ \times 0.1^\circ$. Four predictors are considered, including the monthly precipitation, temperature, cumulative water storage change (CWSC), and ERA5L-derived TWSA. The spatial resolution of these data is averagely resampled to $1^\circ \times 1^\circ$ to be consistent with GRACE TWSA. CWSC is calculated as the cumulative difference between the inflow (i.e., precipitation P) and outflow (i.e., evapotranspiration ET and runoff RO) of a grid cell:

$$\mathrm{CWSC}_t = \sum_{i=1}^{t}\left(P_i - ET_i - RO_i\right), \tag{1}$$

where t denotes the month index. Anthropogenic activities (e.g., groundwater extraction) can also influence the water storage, but they are difficult to quantify and thus not considered here.

The ERA5L dataset includes water storage in soil moisture, snow, and canopy. Thus, the ERA5L TWSA is calculated by summing these components and then subtracting the long-term mean between 2004 and 2009 to be consistent with GRACE TWSA, as represented by:

$$\mathrm{TWSA}_{\mathrm{ERA5L}} = \mathrm{SMS} + \mathrm{SWS} + \mathrm{CWS} - \overline{\mathrm{TWS}}_{0409}, \tag{2}$$

where SMS, SWS, and CWS are soil moisture, snow water, and canopy water storage, respectively, $\overline{\mathrm{TWS}}_{0409}$ denotes the 2004–2009 mean.

2.3. Time series data detrending

The GRACE TWSA time series may exhibit long-term declining/rising trends caused by the human interventions and/or changing climate. This presents challenges for TWSA prediction, as the hydroclimatic predictor data may not be able to fully capture these trends (Humphrey and Gudmundsson, 2019; Li et al., 2020; Sun et al., 2020). For the gap-filling task considered here, fortunately, the GRACE data before (April 2002–June 2017) and after (June 2018-) the gap are available. Therefore, we can obtain directly the long-term trends covering the gap period from existing data. Then we predict in the gap-filling task the detrended TWSAs instead, which is generally less challenging relative to predicting the original signals (Humphrey and Gudmundsson, 2019; Li et al., 2020). Mathematically, the GRACE TWSA time series are decomposed via linear detrending into two components:

$$\mathrm{TWSA}_{\mathrm{GRACE}} = \mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE}} + \mathrm{trend}_{\mathrm{GRACE}}, \tag{3}$$

where $\mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE}}$ is the detrended data, $\mathrm{trend}_{\mathrm{GRACE}}$ is the linear long-term trend obtained by linear fitting with the available GRACE (-FO) data between 2002 and 2020. Correspondingly, the driving data described in Section 2.2 are also detrended. In the prediction task, BCNN learns to predict the $\mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE}}$ signals and the predictions for the original TWSAs are then obtained by adding the GRACE trend:

$$\mathrm{TWSA}_{\mathrm{BCNN}} = \mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{BCNN}} + \mathrm{trend}_{\mathrm{GRACE}}. \tag{4}$$

3. Methods

3.1. BCNN deep learning model

The BCNN model is proposed to learn the underlying relationship between $\mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE}}$ and its four predictors (i.e., the detrended P, T, CWSC, and $\mathrm{TWSA}_{\mathrm{ERA5L}}$). Here we denote the network inputs (i.e., predictors) and outputs (i.e., $\mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE}}$) as $\mathbf{x}$ and $\mathbf{y}$, respectively. The global fields of GRACE TWSA and hydroclimatic predictors are arranged as images. The learning of high-dimensional and complex mapping between the outputs and inputs becomes an image regression problem and can leverage CNN's robust capability in image processing (Gu et al., 2018; Mo et al., 2019; Shen, 2018), as represented by

$$\boldsymbol{\eta} : \mathbf{x} \in \mathbb{R}^{n_x \times H \times W} \longrightarrow \mathbf{y} \in \mathbb{R}^{n_y \times H \times W}, \tag{5}$$

where $\boldsymbol{\eta} = \boldsymbol{\eta}(\mathbf{x}, \mathbf{w})$ is a BCNN model, with $\mathbf{w}$ denoting all trainable network parameters. The inputs $\mathbf{x}$ and outputs $\mathbf{y}$ become n_x and n_y images, respectively, all with $H \times W$ pixels (grids). It is worth mentioning that the spherical CNN (Su and Grauman, 2017) may be an alternative to vanilla CNN for the image regression task considered here, as the GRACE mascons are defined on a sphere. This, though out the scope of this work, deserves further investigation.

The network predictions are inevitably associated with epistemic uncertainties induced by a lack of training data. To quantify the predictive uncertainties, we treat the network parameters $\mathbf{w}$ as random variables. Given a set of training data $\mathscr{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N_{\mathrm{train}}}$, the network training is to infer the posterior distribution of $\mathbf{w}$, $p(\mathbf{w}|\mathscr{D})$. Consequently, one can obtain the predictive distribution of the target $\mathbf{y}$: $p(\mathbf{y}|\mathbf{w})$, $\mathbf{w} \sim p(\mathbf{w}|\mathscr{D})$, and in particular the mean $\mathbb{E}(\mathbf{y}|\mathbf{w})$ and standard deviation $\mathtt{Std}(\mathbf{y}|\mathbf{w})$.

In BCNN, a Bayesian training strategy called stein variational gradient descent (SVGD) (Liu and Wang, 2016; Zhu and Zabaras, 2018) is employed to estimate the posterior distribution $p(\mathbf{w}|\mathscr{D})$. Mathematically, the BCNN model is expressed as follows:

$$\hat{\mathbf{y}} = \boldsymbol{\eta}(\mathbf{x}, \mathbf{w}) + \mathbf{n}(\mathbf{x}, \mathbf{w}), \tag{6}$$

where $\hat{\mathbf{y}}$ denotes BCNN's prediction and $\mathbf{n}(\cdot)$ is an additive Gaussian noise term modeling the aleatoric uncertainty. The SVGD algorithm is similar to standard gradient descent while maintaining the particle methods' high efficiency (Liu and Wang, 2016). In implementation, we use N_S particles of $\mathbf{w}$ to approximate the posterior distribution. The N_S samples $\{\mathbf{w}_i\}_{i=1}^{N_S}$ are respectively optimized using the Adam optimizer (Kingma and Ba, 2015), whose gradient derives from SVGD. The predictive mean and standard deviation (i.e., uncertainty) of BCNN for an arbitrary input $\mathbf{x}$ can be then computed using the N_S predictions ($\hat{\mathbf{y}}^{(i)} = \boldsymbol{\eta}(\mathbf{x}, \mathbf{w}_i) + \mathbf{n}(\mathbf{x}, \mathbf{w}_i)$, $i = 1, \ldots, N_S$). For more details regarding the SVGD Bayesian training strategy, one can refer to Liu and Wang (2016) and Zhu and Zabaras (2018).

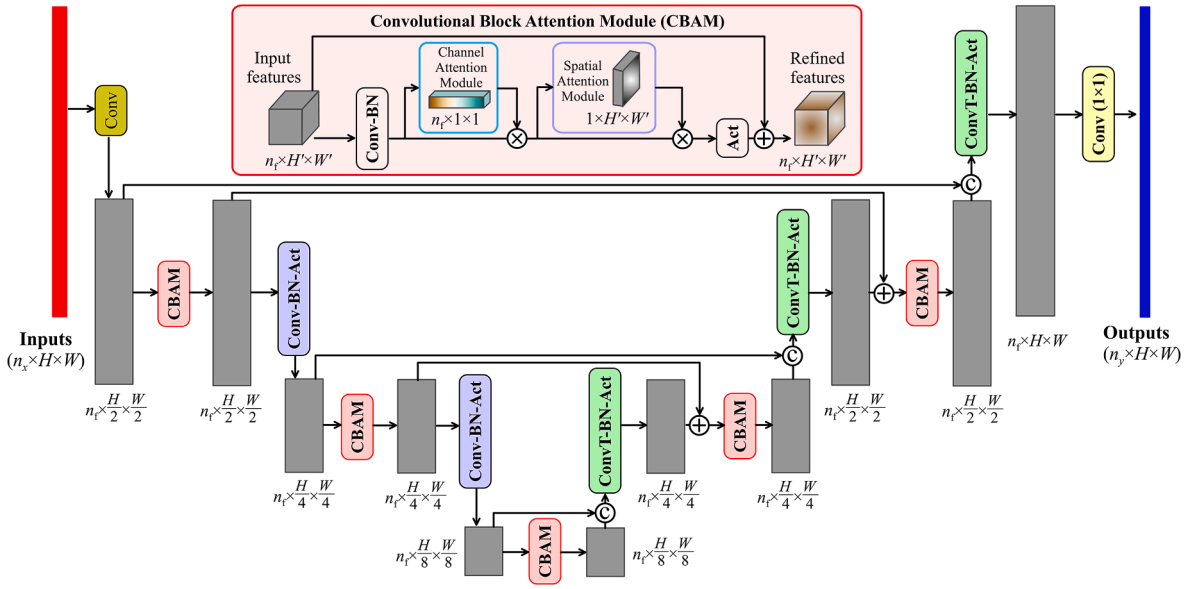

Fig. 2. Illustration of the Bayesian convolutional neural network (BCNN) architecture. It takes n_x images with a size of $H \times W$ as inputs and generates n_y images with the same size. It is an alternating cascade of convolutional (Conv)/transposed convolutional (ConvT) layers and convolutional block attention modules (CBAM), each of which outputs $n_f = 48$ feature maps. The size of feature maps is sequentially halved in each Conv layer from $H \times W$ to $\frac{H}{8} \times \frac{W}{8}$ to extract multi-scale features, and then sequentially recovered to $H \times W$ with ConvT layers. The symbols $\oplus$, ©, and $\otimes$ denote the addition (i.e., residual connection), concatenation (i.e., skip connection), and multiplication (i.e., attention connection) operations, respectively. Act and BN denote activation and batch normalization, respectively.

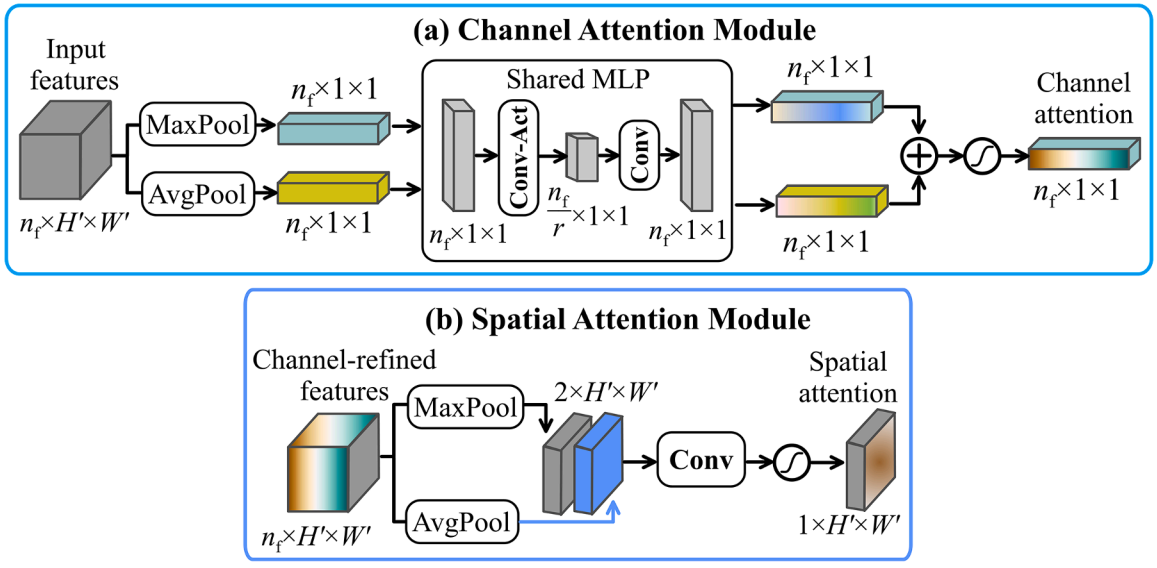

Fig. 3. Diagrams of the channel and spatial attention modules. The inputs to each module are n_f feature maps with a size of $H' \times W'$. The channel module utilizes both max-pooling (MaxPool) and average-pooling (AvgPool) outputs with a shared multi-layer perceptron (MLP) to produce a channel attention map ($n_f \times 1 \times 1$). The spatial module utilizes similar two outputs to produce a spatial attention map ($1 \times H' \times W'$). The sigmoid activation is used to guarantee the output values are between 0 and 1. Conv and Act denote the Convolution and Activation operations, respectively.

3.2. BCNN architecture design and training

The BCNN network architecture is depicted in Fig. 2. The convolutional block attention module (CBAM) (Woo et al., 2018) is used as the basic block. Given n_x images with a resolution of $H \times W$ as inputs to the network, they are passed through an alternating cascade of convolutional/transposed convolutional layers and CBAMs, each of which produces n_f feature maps with a resolution of $H' \times W'$, to extract multi-scale and hierarchical features to finally predict n_y images for the targets. The CBAM block contains two attention modules, namely the channel and spatial attentions as depicted in Fig. 2 and detailed in Fig. 3. More specifically, the channel module outputs n_f weights between 0 and 1 assigning to the n_f feature maps to tell the network 'what' (i.e., which maps) to attend; the spatial module outputs a $H' \times W'$ weight matrix assigning to the $(H' \times W')$-pixel feature maps to tell the network 'where' (i.e., which regions) to emphasize or suppress. As such, the network is able to automatically focus on important features and suppress unnecessary ones (Woo et al., 2018).

The residual (He et al., 2016) and skip (Ronneberger et al., 2015) connections are also adopted in our BCNN model. It has been extensively shown that they can effectively resolve the vanishing gradient problem and enhance information flow through the deep networks, substantially

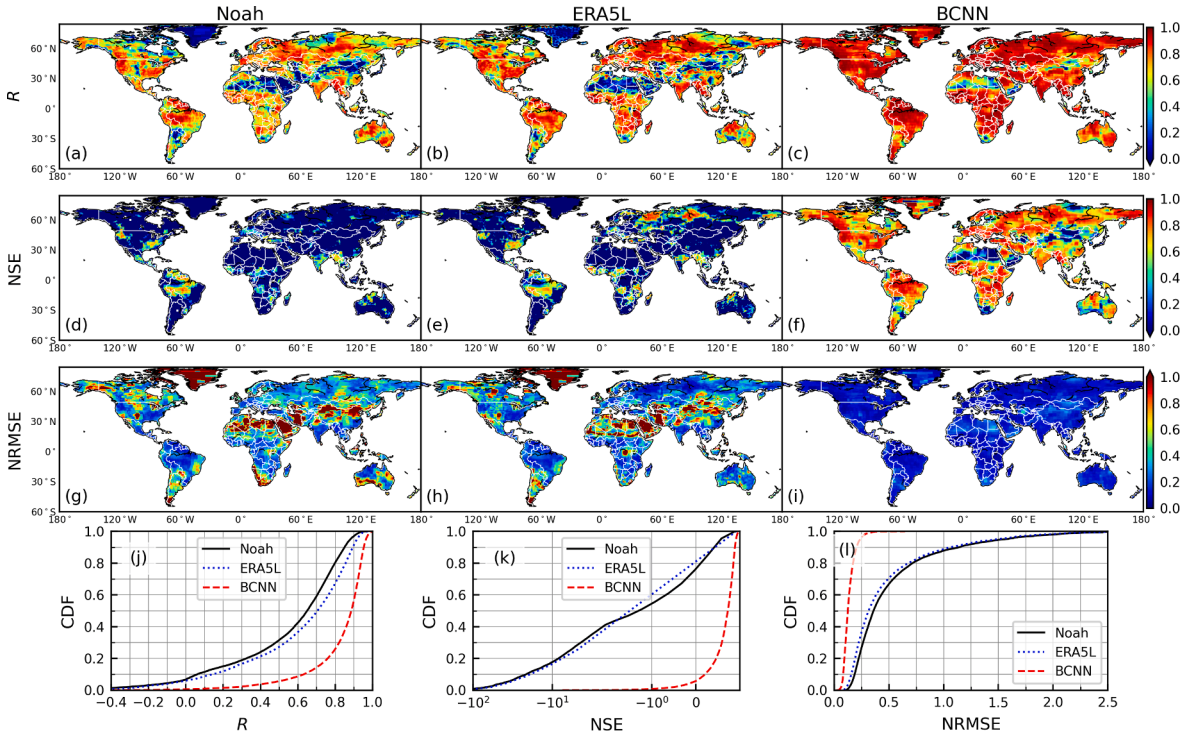

Fig. 4. Spatial maps of R (row 1), NSE (row 2), and NRMSE (row 3) values between the GRACE TWSAs and the Noah- (column 1), ERA5L- (column 2), and BCNN-derived (column 3) TWSAs during the testing periods (April 2014–June 2017, June 2018–August 2020). (j-l) Cumulative distribution functions (CDFs) of the R, NSE, and NRMSE values in (a-i).

improving the network performance. In the residual connection, the feature maps with the same shape ($n_f \times H' \times W'$) but at different layers are connected by applying element-wise addition (He et al., 2016). In the skip connection, the feature maps with the same size ($H' \times W'$) but at different layers are cascaded together and subsequently fed as inputs into the next layer (Ronneberger et al., 2015) (Fig. 2). The Mish function (Misra, 2019) is employed in BCNN as the activation function unless otherwise stated.

We use twelve years of monthly GRACE TWSA data from April 2002 to March 2014 (i.e., 144 months, ~69%) to train the BCNN network, and those from April 2014 to June 2017 and June 2018 to August 2020 (i.e., 66 months, ~31%) to test the performance. We set the number of lags for predictors to 2 after preliminary test experiments, as the increased lag did not clearly improve the performance (not shown). That is, for month t, the inputs to BCNN are the four predictors in months $t-2$ to t. Thus, each sample contains $n_x = 12$ input images and $n_y = 1$ output image (i.e., $\mathrm{TWSA}^{\mathrm{detrend}}_{\mathrm{GRACE},t}$). The region spanning from 60°S to 84°N and 180°W to 180°E (i.e., $H \times W = 144 \times 360$) is considered. During network training, we use $N_S = 20$ particles of $\mathbf{w}$ in the SVGD algorithm to approximate the posterior distribution, as suggested in Zhu and Zabaras (2018). The network is trained for 200 epochs, with a mean squared error loss function quantifying the predictive accuracy, an initial learning rate of 0.0025, and a batch size of 12. The training performed on a single GPU (NVIDIA Tesla V100) takes ~80 min. The network performance is evaluated with the testing data using three commonly used metrics, namely the correlation coefficient (R), Nash–Sutcliffe efficiency coefficient (NSE), and normalized root mean squared error (NRMSE):

$$R = \frac{\sum_{i=1}^{N_{\mathrm{test}}} (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N_{\mathrm{test}}} (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^{N_{\mathrm{test}}} (\hat{y}_i - \bar{\hat{y}})^2}}, \tag{7}$$

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N_{\mathrm{test}}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N_{\mathrm{test}}} (y_i - \bar{y})^2}, \tag{8}$$

$$\mathrm{NRMSE} = \frac{\sqrt{\frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} (y_i - \hat{y}_i)^2}}{y_{\max} - y_{\min}}, \tag{9}$$

where y denotes the observations and $\hat{y}$ the predictions, with $\bar{y}$ and $\bar{\hat{y}}$ respectively denoting their means, N_{test} is the number of testing samples, $y_{\max}$ and $y_{\min}$ represent the maximum and minimum values of y, respectively. A R or NSE value closer to 1.0 and a NRMSE value closer to 0 indicate better performances.

4. Results and discussion

4.1. Accuracy assessment with testing GRACE data

The prediction accuracy is assessed using the GRACE data in 66 testing months. To illustrate the performance of BCNN, the R, NSE, and NRMSE metrics are also computed for the ERA5L-reanalyzed TWSAs (Muñoz Sabater, 2019) and Noah-simulated TWSAs (Rodell et al., 2004).

Fig. 4 shows the spatial maps of the accuracy metrics obtained by Noah, ERA5L, and BCNN. While Noah and ERA5L both show relatively good correlations with GRACE in most regions except Greenland and the hyper-arid areas like Sahara, Gobi, and Arabian (Figs. 4(a,b)), BCNN's R values are clearly higher than those of Noah and ERA5L in almost all regions (Fig. 4c). For the NSE metric, which measures directly the matching quality between the predicted and observed values, both Noah and ERA5L obtain unsatisfactorily low values (<0) in most regions except in some humid regions like Amazon and Southeastern United States (Fig. 4(d, e)). In contrast, BCNN provides relatively high values (>0.5) in most regions (Fig. 4f). Note that although BCNN achieves

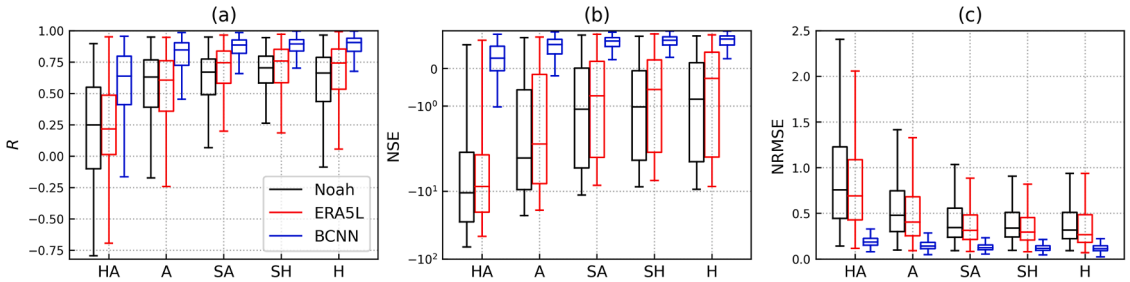

Fig. 5. Boxplots of the (a) *R*, (b) NSE, and (c) NRMSE values at grids in the hyper-arid (HA), arid (A), semi-arid (SA), semi-humid (SH), and humid (H) regions obtained by Noah, ERA5L, and our BCNN. The outliers are not shown.

Table 1
Medians of the *R*, NSE, and NRMSE values at grids in the hyper-arid (HA), arid (A), semi-arid (SA), semi-humid (SH), and humid (H) regions obtained by Noah, ERA5L, and our BCNN. Bold value indicates the best performance.

	R					NSE					NRMSE				
	HA	A	SA	SH	H	HA	A	SA	SH	H	HA	A	SA	SH	H
Noah	0.25	0.63	0.67	0.71	0.66	−10.50	−3.23	−1.08	−1.02	−0.82	0.76	0.48	0.35	0.34	0.32
ERA5L	0.22	0.61	0.75	0.76	0.74	−8.51	−2.01	−0.73	−0.56	−0.26	0.69	0.41	0.31	0.29	0.27
BCNN	**0.64**	**0.85**	**0.89**	**0.90**	**0.91**	**0.27**	**0.63**	**0.73**	**0.75**	**0.78**	**0.18**	**0.14**	**0.12**	**0.12**	**0.11**

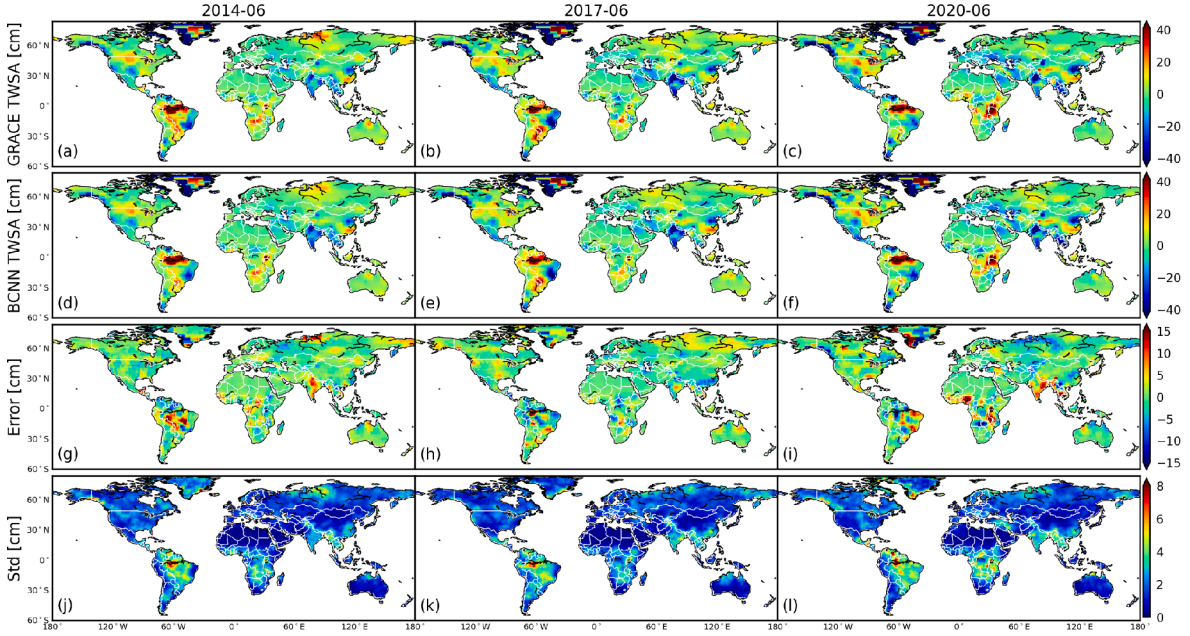

Fig. 6. The GRACE (a–c) and BCNN (d–f) TWSAs in three testing months (June 2014, June 2017, and June 2020). (g–i) BCNN's predicted error (i.e., $TWSA_{GRACE}$-$TWSA_{BCNN}$) and (j-l) standard deviation (Std). The Std is calculated using an ensemble of $N_S = 20$ BCNN predictions. BCNN's training period is April 2002-March 2014.

higher accuracy than Noah and ERA5L in the hyper-arid regions, the NSE values are still low relative to other regions. This is due to the fact that the TWSA signals in these regions is dominated by noise (Humphrey et al., 2016). The improved performance of BCNN over Noah and ERA5L can be also illustrated by the NRMSE maps depicted in Figs. 4(g-i) and the cumulative distribution functions of the three metrics depicted in Fig. 4(j–l), which indicate that BCNN provides significantly better accuracy (i.e., much higher *R* and NSE values and much lower NRMSE values). Note that the outperformance of BCNN benefits not only from its own robust capability in learning complex mappings, but also from the use of GRACE data for training. The inability of Noah and ERA5L to consider the groundwater and surface water components in their water storage estimates may be another cause for the inconsistency.

It can be seen from Fig. 4 that the performance is highly dependent on the regional climate conditions. We further compare the three models' *R*, NSE, and NRMSE values at grids in the hyper-arid, arid, semi-arid, semi-humid, and humid regions (see Fig. 1 for the climate regions). The results are summarized in the boxplots depicted in Fig. 5, with the metric medians being listed in Table 1. In general, higher performances are achieved as expected in regions with more humid climate. This is probably because the arid regions usually have relatively low signal-to-noise ratios and are often associated with heavy human interventions (e.g., groundwater extractions and reservoir operations) (Humphrey et al., 2016; Sun et al., 2020). Likewise, BCNN clearly outperforms Noah and ERA5L in all climate regions. For example, the median NSE values of Noah, ERA5L, and BCNN in the hyper-arid region are −10.50, −8.51,

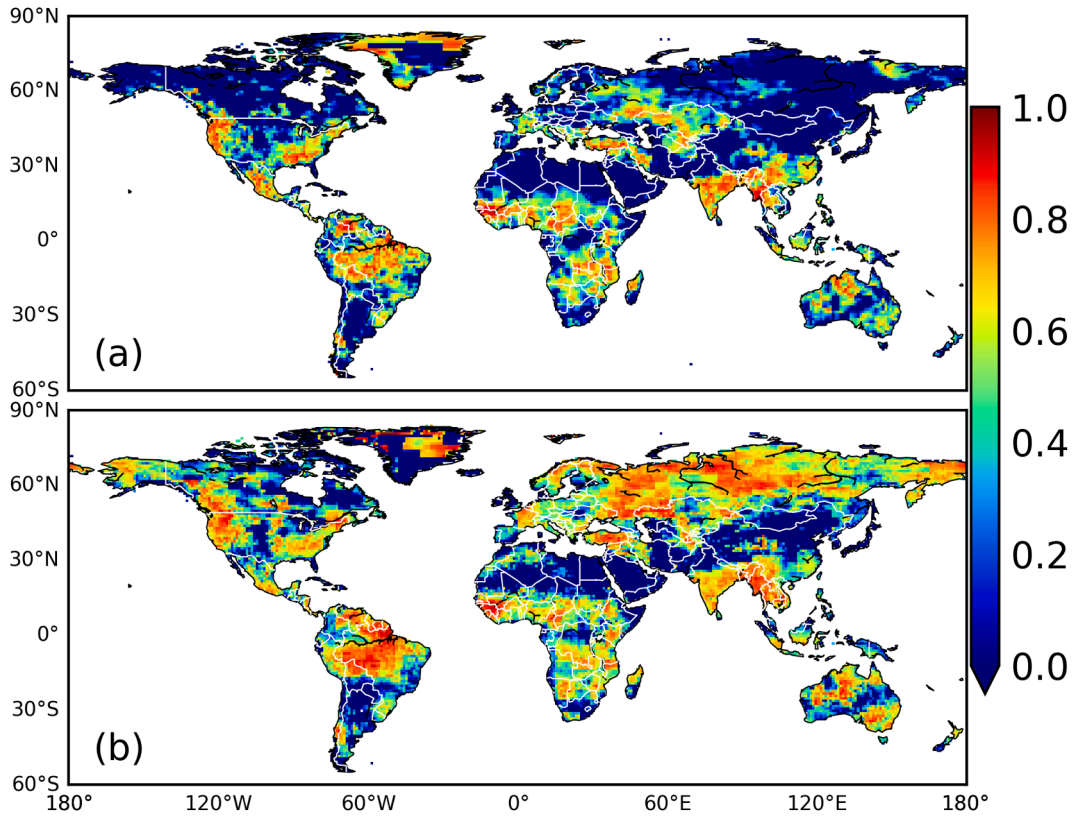

Fig. 7. The maximum NSE values (among three lagged months, $t-2$, $t-1$, t) (a) between the standardized GRACE TWSA and CPC soil moisture and (b) between the standardized GRACE and Noah TWSAs during the testing periods (April 2014–June 2017 and June 2018–August 2020). The data are standardized to exclude the influence of amplitude and magnitude differences.

and 0.27, respectively (Table 1).

Fig. 6 depicts BCNN's TWSA predictions for three testing months in June 2014, June 2017, and June 2020. Note that the GRACE data of the three months were not seen by BCNN during model training. For comparison, the reference GRACE TWSA fields are also shown. Due to the Bayesian nature of BCNN, the predictive uncertainties can be quantified and are depicted as standard deviation in the plot. It can be seen that BCNN successfully captures the spatial patterns of GRACE TWSA and provides close predictions in the three months (Figs. 6(a-f)). The predictive errors and uncertainties in humid regions (e.g., Amazon, Central Africa, South Asia, and Greenland) are generally larger compared to other regions (Figs. 6(g-l)), which are mainly because of the relatively high signal variability in the humid regions. The BCNN's TWSA predictions for all of the 66 testing months are shown in the GIF animation attached as supporting materials.

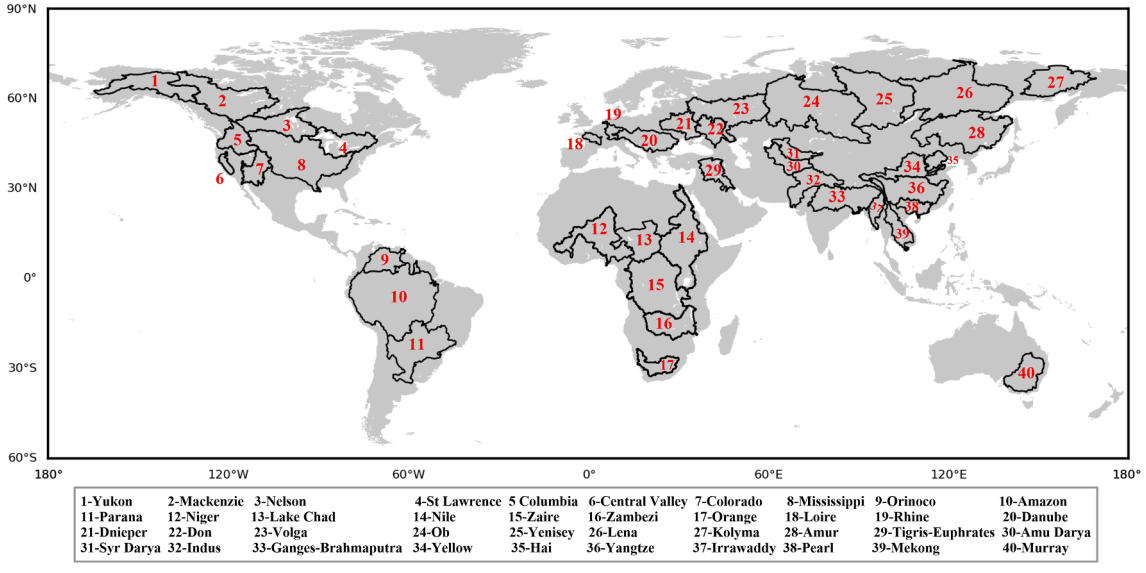

Fig. 8. The 40 selected major river basins. The basin boundaries are available at https://datacatalog.worldbank.org/dataset/major-river-basins-world.

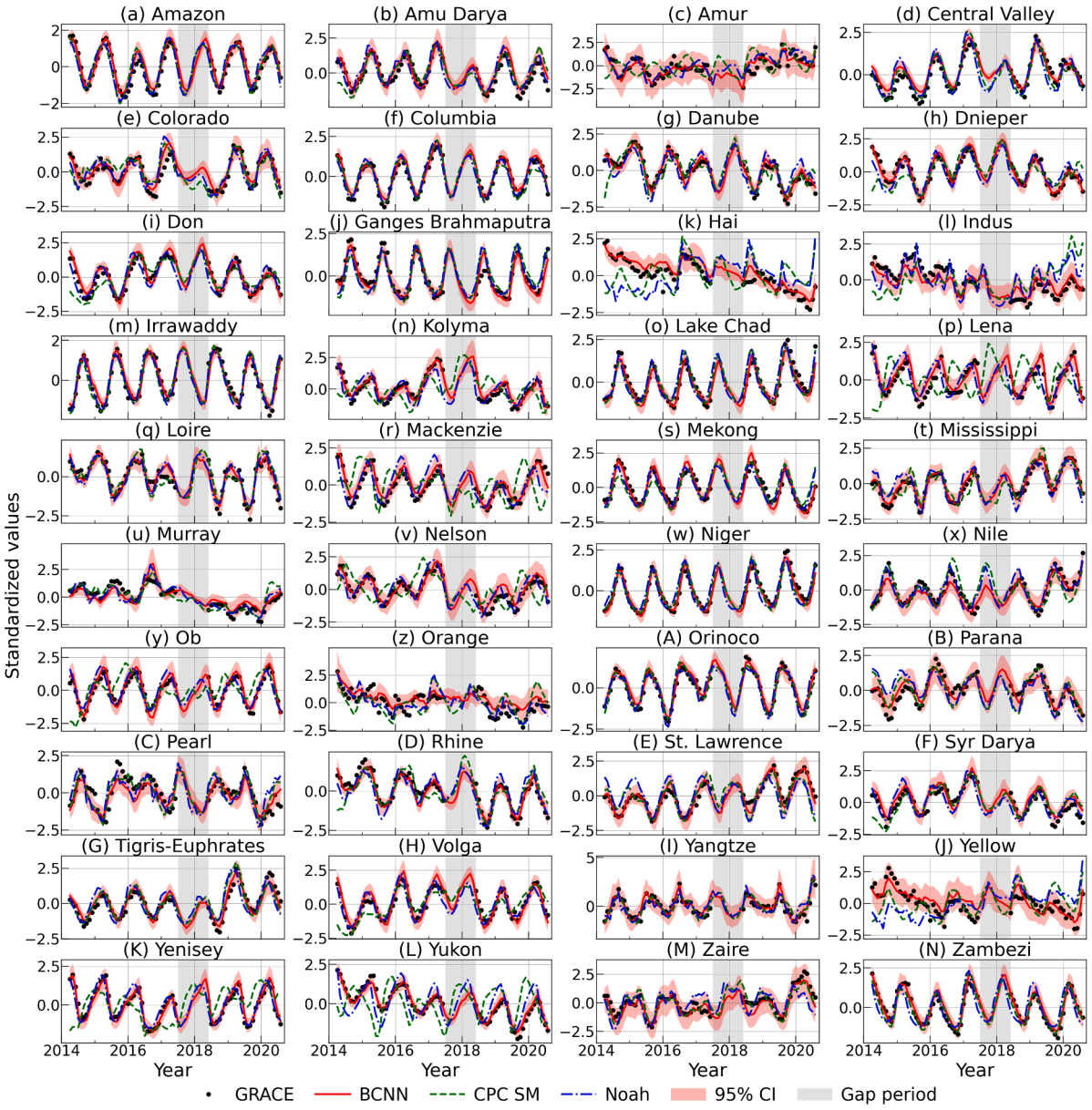

Fig. 9. The standardized time series of basin-averaged GRACE TWSA, BCNN TWSA, CPC soil moisture (SM), and Noah TWSA during the testing and gap periods. All time series are standardized to exclude the influence of amplitude and magnitude differences (GRACE and BCNN TWSAs are both standardized using the mean and standard deviation of GRACE for consistency). The red shaded area denote the 95% confidence interval (CI) of BCNN's predictions. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Evaluation of gap-filling quality at the basin scale

The results presented in Section 4.1 indicate that BCNN is able to provide close predictions to the testing GRACE data in the pre- and post-gap periods. To further validate that BCNN's infilling data are GRACE-like and reliable, two independent datasets, namely the CPC (Climate Prediction Center) soil moisture (van den Dool et al., 2003) and Noah-simulated TWSAs (Rodell et al., 2004), are used for verification at the basin scale. More specifically, we compare the basin-averaged time series of these datasets. As shown in Section 4.1 and by Scanlon et al. (2019), the amplitude and magnitude differences of two time series may lead to poor consistency. Therefore, these basin-averaged time series are respectively standardized to make the data comparable:

$$y = \frac{y - \mu_y}{\sigma_y}, \tag{10}$$

where μ_y and σ_y are the mean and standard deviation, respectively, of the time series. For consistency, the BCNN and GRACE TWSA time series are both standardized using GRACE's μ_y and σ_y. The grid-scale NSE values between the standardized GRACE TWSAs and soil moisture/Noah TWSAs during the testing periods indicate that they agree relatively well in most regions (Fig. 7). Particularly, after excluding the influence of amplitude and magnitude differences by standardization, the consistency between Noah and GRACE is significantly improved (see Figs. 4d and 7b). The idea behind the validation is, suppose that the standardized time series of soil moisture/Noah TWSA agree well with that of GRACE TWSA in the pre- and post-gap testing periods, the gap-filling results can be thought to be reliable if the BCNN TWSA time series fit well with those of soil moisture/Noah TWSA during the testing and gap periods.

The basin-averaged standardized time series of the GRACE TWSA, BCNN TWSA, CPC soil moisture, Noah TWSA for 40 major river basins (see Fig. 8 for the basin locations) are compared in Fig. 9, with the NSE

Table 2
The NSE values between basin-averaged time series of GRACE (G)/BCNN (B) TWSA and CPC soil (S) moisture/Noah (N) TWSA in 40 river basins with humid (H), semi-humid (SH), semi-arid (SA), or arid (A) climates. All time series are standardized to exclude the influence of amplitude and magnitude differences (GRACE and BCNN TWSAs are both standardized using the mean and standard deviation of GRACE for consistency). $trend_G$ and $trend_P$ denote the trends [cm/year] of GRACE TWSA in April 2014-August 2020 and of annual precipitation in 2014–2020, respectively. Basins with NSE_{G-S}, NSE_{B-S}, NSE_{G-N}, and NSE_{B-N} all less than 0.4 are bolded. NSE_{G-B}, NSE_{G-S}, and NSE_{G-N} are calculated for the testing periods (April 2014-June 2017 and June 2018-August 2020). NSE_{G-S} and NSE_{B-N} are calculated for the testing and gap periods (April 2014-August 2020).

Basin	ID	Climate	$trend_G$ ($trend_P$)	NSE_{G-B}	NSE_{G-S} (NSE_{B-S})	NSE_{G-N} (NSE_{B-N})
Amazon	10	H	0.04 (−1.33)	0.93	0.80 (0.85)	0.80 (0.83)
Amu-Darya	30	SA	−0.61 (-0.55)	0.85	0.61 (0.71)	0.71 (0.79)
Amur	28	SH	0.20 (1.42)	0.71	−0.78 (−0.47)	0.39 (0.63)
Central Valley	6	SA	2.19 (0.28)	0.85	0.85 (0.81)	0.71 (0.67)
Colorado	7	A	0.04 (−0.80)	0.85	0.53 (0.51)	0.53 (0.72)
Columbia	5	H	0.47 (−1.01)	0.94	0.88 (0.88)	0.89 (0.89)
Danube	20	H	−1.99 (−1.40)	0.93	0.61 (0.57)	0.60 (0.77)
Dnieper	21	H	−0.85 (0.50)	0.93	0.23 (0.50)	0.62 (0.77)
Don	22	SH	−0.16 (−0.98)	0.91	0.52 (0.48)	0.73 (0.74)
Ganges–Brahmaputra	33	H	−1.09 (3.02)	0.88	0.69 (0.81)	0.74 (0.83)
Hai	35	SA	−1.59 (−0.17)	0.73	**−0.85 (−1.03)**	**−0.99 (−1.33)**
Indus	32	SA	−1.62 (0.65)	0.67	**−1.03 (−1.67)**	**−0.07 (−0.39)**
Irrawaddy	37	H	0.44 (−0.11)	0.94	0.92 (0.91)	0.93 (0.91)
Kolyma	27	SH	−0.73 (−0.27)	0.91	−0.24 (−0.01)	0.61 (0.63)
Lake Chad	13	SA	0.64 (0.50)	0.91	0.84 (0.86)	0.75 (0.74)
Lena	26	SH	−0.03 (−0.15)	0.87	−1.50 (−1.98)	0.46 (0.45)
Loire	18	H	−0.93 (0.07)	0.87	0.78 (0.77)	0.77 (0.95)
Mackenzie	2	SH	−0.12 (1.57)	0.87	−0.82 (-0.79)	0.47 (0.56)
Mekong	39	H	−1.00 (0.14)	0.89	0.63 (0.77)	0.74 (0.84)
Mississippi	8	SH	1.39 (1.38)	0.95	0.71 (0.75)	0.69 (0.79)
Murray	40	SA	−1.24 (−1.06)	0.73	0.69 (0.46)	0.75 (0.81)
Nelson	3	SH	−0.64 (−0.87)	0.85	−0.36 (−0.81)	0.74 (0.63)
Niger	12	SA	1.04 (0.45)	0.95	0.87 (0.93)	0.76 (0.77)
Nile	14	SA	0.48 (1.65)	0.84	0.69 (0.11)	0.77 (0.61)
Ob	24	SH	0.10 (−0.53)	0.90	−0.23 (−0.06)	0.71 (0.80)
Orange	17	A	−0.36 (0.12)	0.44	**−0.06 (−1.25)**	**0.38 (−0.20)**
Orinoco	9	H	0.41 (−1.46)	0.94	0.92 (0.91)	0.88 (0.84)
Parana	11	H	0.10 (−6.29)	0.82	0.54 (0.58)	0.16 (0.31)
Pearl	38	H	−0.67 (−1.72)	0.67	0.60 (0.69)	0.27 (0.54)
Rhine	19	H	−1.98 (−0.48)	0.90	0.57 (0.52)	0.67 (0.71)
St. Lawrence	4	H	1.81 (1.32)	0.92	0.57 (0.72)	0.26 (0.60)
Syr-Darya	31	SA	−0.50 (−0.89)	0.77	0.57 (0.51)	0.64 (0.69)
Tigris-Euphrates	29	SA	1.34 (1.10)	0.90	0.74 (0.88)	0.53 (0.69)
Volga	23	H	0.44 (0.61)	0.86	0.38 (0.32)	0.78 (0.82)
Yangtze	36	H	0.05 (0.41)	0.78	0.73 (0.81)	0.59 (0.82)
Yellow	34	SA	−0.53 (0.87)	0.63	**−0.52 (−0.97)**	**−1.16 (−1.89)**
Yenisey	25	H	−0.43 (0.65)	0.94	−0.42 (−0.25)	0.81 (0.81)
Yukon	1	SH	−3.13 (0.90)	0.96	**−0.89 (−1.16)**	**0.35 (0.34)**
Zaire	15	H	0.95 (1.44)	0.81	0.70 (0.66)	0.17 (0.28)
Zambezi	16	SA	−1.06 (1.46)	0.92	0.86 (0.92)	0.58 (0.64)

values between them being summarized in Table 2. These plots manifest that BCNN TWSAs show favorable consistency with GRACE TWSAs during the testing periods in the 40 basins. The NSE values between them are generally larger than 0.7, with the exception of Indus (NSE = 0.67), Orange (NSE = 0.44), Pearl (NSE = 0.67), and Yellow (NSE = 0.63) River Basins. The mismatch in these basins is probably because of the insufficient quality and/or ability of driving data to capture the impacts of anthropogenic activities on water cycle. It is worthy noting that although BCNN may slightly underestimate/overestimate GRACE TWSAs, the GRACE curves are almost completely enveloped within BCNN's 95% prediction interval (calculated using an ensemble of N_S = 20 BCNN predictions; Section 3.1). Additionally, the standardized time series of soil moisture and Noah TWSA agree relatively well with those of GRACE TWSA in most basins during the GRACE-covered periods, as also indicated by the NSE values listed in Table 2. Therefore, the good consistency between BCNN TWSAs and soil moisture/Noah TWSAs in these basins suggests the reliability of BCNN's gap-filling results. For the remaining few basins with the NSE values between the GRACE/BCNN TWSA and soil moisture/Noah TWSA time series all less than 0.4 (marked in bold in Table 2), it is found that all of these basins have declining GRACE TWSA trends while the precipitation exhibits oppositely rising trends (an exception is the Hai River Basin, but the TWSA time series are more significantly declining). The decreased water storage in these basins are mainly resulted from glacier retreating (Yukon River Basin) or groundwater depletion (Hai, Indus, and Yellow River Basins) (Rodell et al., 2018). The inability of soil moisture and Noah to reflect the TWSA declines induced by glacier retreating and groundwater depletion leads to the poor consistency. Despite this, BCNN's close predictions to GRACE TWSAs in the pre- and post-gap periods still suggest the reliability of gap-filling results (Figs. 9(k, l, z, I, L)).

For the gap-filling purpose, one can also simply bridge the gap with the long-term trend and seasonal signals derived from the available GRACE observations (i.e., $TWSA^{T+S}_{GRACE}$=$trend_{GRACE}$+$season_{GRACE}$). Fig. 10 depicts the basin-averaged time series of the original (without standardizing) GRACE TWSA, BCNN TWSA, and $TWSA^{T+S}_{GRACE}$ for the 40 river basins. It is observed that BCNN achieves a better consistency (higher NSE values) with GRACE than $TWSA^{T+S}_{GRACE}$ in all basins. Fig. 11 displays the NSE field between GRACE TWSA and $TWSA^{T+S}_{GRACE}$. The regions with high NSE values are as expected generally the seasonal component-dominant regions (see Fig. 9 in Humphrey et al. (2016) for the spatial distribution of seasonal component-dominant regions).

One popular use of the TWSA data is to estimate the long-term trend of water storage (Chen et al., 2014; Feng et al., 2013; Feng et al., 2018; Scanlon et al., 2018; Tapley et al., 2019). Here we investigate the impact

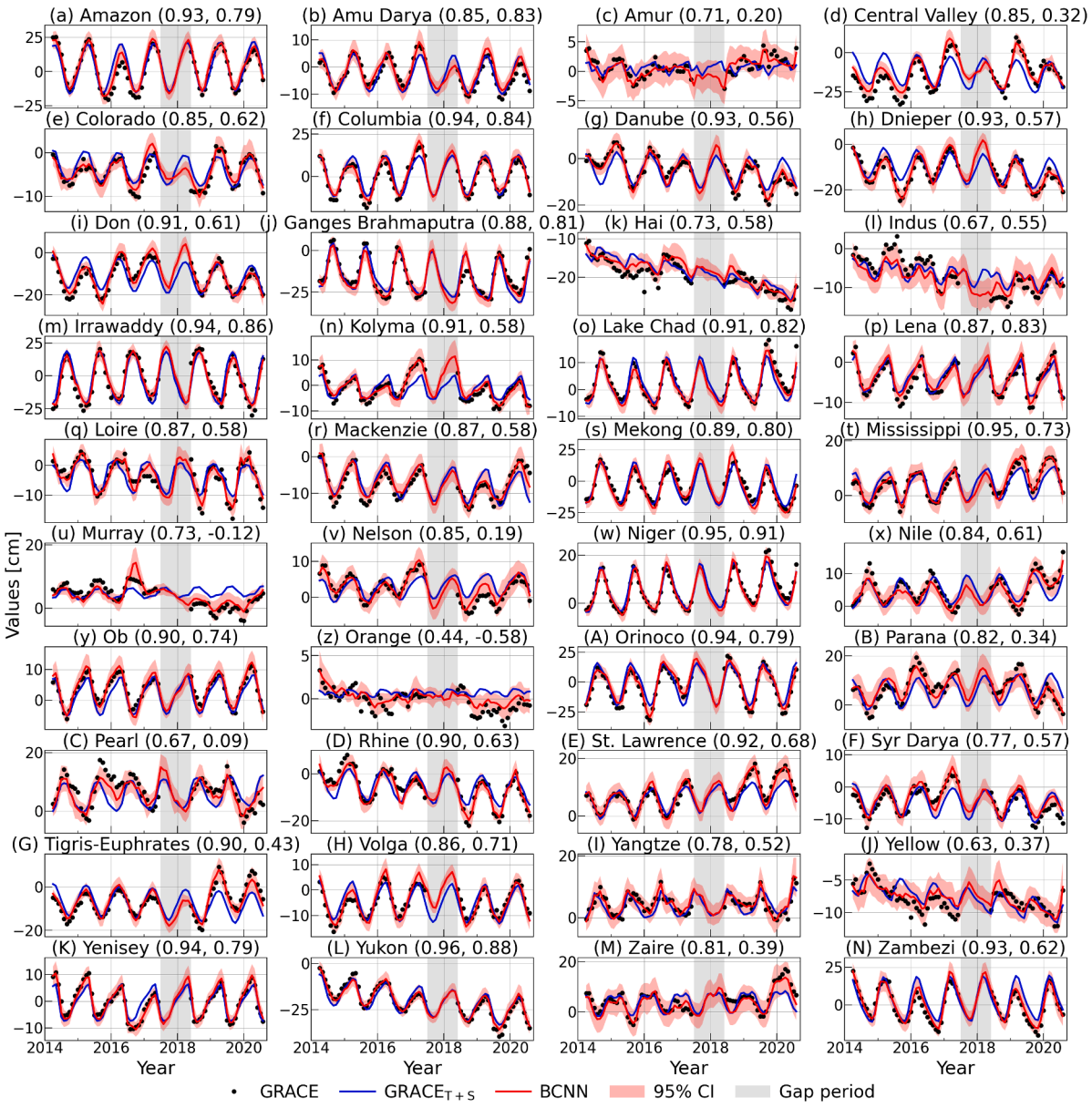

Fig. 10. Comparison of the basin-averaged time series of GRACE, BCNN, and GRACE$_{T+S}$ TWSAs (sum of the trend and seasonal signals derived from GRACE TWSA) during the testing and gap periods. The first and second numbers in the bracket denote the NSE values of BCNN and GRACE$_{T+S}$, respectively, with GRACE.

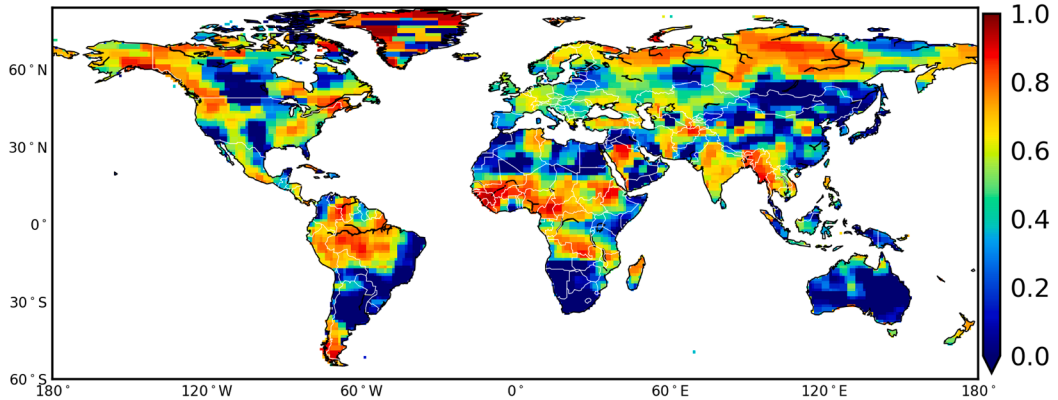

Fig. 11. NSE values between GRACE and GRACE$_{T+S}$ TWSAs (sum of the trend and seasonal signals derived from GRACE TWSA) during the testing periods (April 2014–June 2017 and June 2018–August 2020).

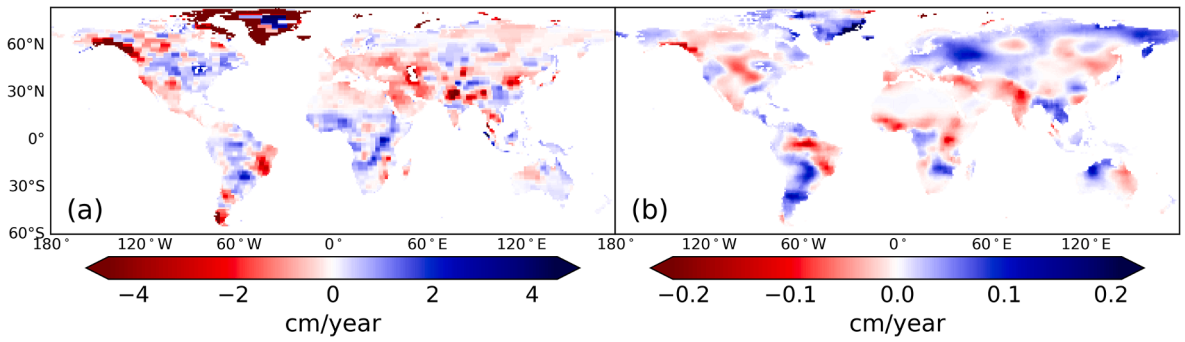

Fig. 12. (a) The TWSA trend field after the gap between GRACE and GRACE-FO is filled with BCNN predictions. (b) The difference (i.e., $trend_1$-$trend_0$) between the trends before ($trend_0$) and after ($trend_1$) gap filling.

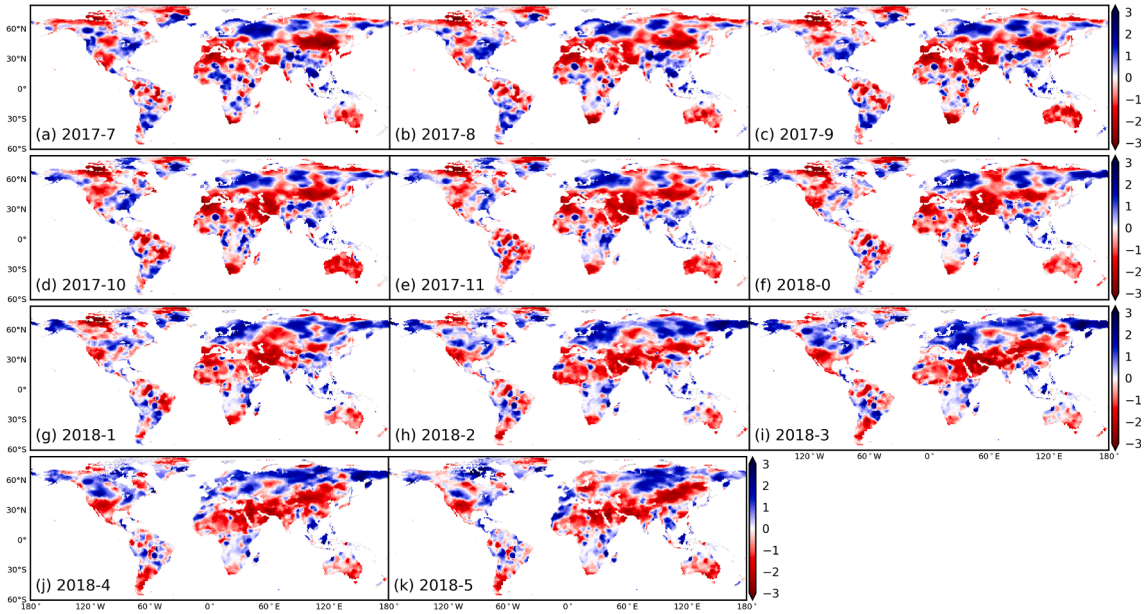

Fig. 13. The 6-month standardized precipitation-evapotranspiration index (SPEI-6) fields during the gap period (July 2017–May 2018).

of the 11-month gap on the long-term (April 2002-August 2020) trend estimation. The trend field after the gap is filled with BCNN-predicted TWSAs is illustrated in Fig. 12. The difference between the trends before and after gap filling is also shown, which can be as large as ±0.2 cm/year in many regions. Taking the Amazon River Basin (the area is over 6×10^6 km^2) as an example, a bias of 0.2 cm/year in the trend estimate would lead to a deviation of over 12 Gt/year in the water storage loss/gain estimate. The differences in trend estimates may be partially caused by dry or wet climate conditions during the gap period, as they usually lead to decreased or increased TWSA signals, respectively. To examine this, the 6-month standardized precipitation-evapotranspiration index (SPEI-6; available at https://spei.csic.es/spei_database) fields during the gap period are plotted in Fig. 13. It is observed that the trend difference patterns (Fig. 12b) are spatially similar to the wet/dry patterns in the SPEI-6 fields, indicating that the overestimation/underestimation of the original long-term trends is mainly related to the the dry/wet conditions during the gap period. BCNN reproduces the dry/wet condition-induced abnormal TWSA signals from hydroclimatic inputs and thus contributes to improving the trend estimation. It should be mentioned that for the extremely dry/wet conditions caused by climate extremes, BCNN's generalization on such extreme-induced abnormal signals may be limited due to the well-known long-tail distribution issue in deep learning (the extremes are associated with only a few samples because of their rare occurrence) (Menon et al., 2021). Integrating the long-tailed learning (Menon et al., 2021) with BCNN may be a better solution.

4.3. Comparison with swarm solution

In this section, we conduct a comparison between BCNN- and Swarm-derived TWSAs to illustrate the merits of BCNN in providing more reliable gap-filling products. The Swarm satellite provides the observations of TWSAs since December 2013 but at much lower resolution relative to GRACE (Friis-Christensen et al., 2008). The data quality is not stable in early years but expected to increase as the mission progresses (da Encarnação et al., 2016). The monthly Swarm level-2 gravity field model, provided by the Astronomical Institute at the Czech Academy of Sciences (ASU) (Bezděk et al., 2016), in terms of spherical harmonic coefficients up to degree and order 40, is used to estimate global TWSAs. To reduce the high magnitude noise in the Swarm solution, we apply a 1000 km Gaussian filter to smooth the data fields. Considering the low resolution of Swarm data, we make a basin-scale comparison for only 15 world's major river basins. The time series of TWSAs from GRACE, Swarm, and BCNN during April 2014 and December 2019 are compared in Fig. 14. The NSE values of Swarm and BCNN TWSAs with the reference GRACE TWSAs are also attached in each subplot. The figure manifests that although the Swarm TWSAs can generally capture the variability patterns of GRACE TWSAs, the

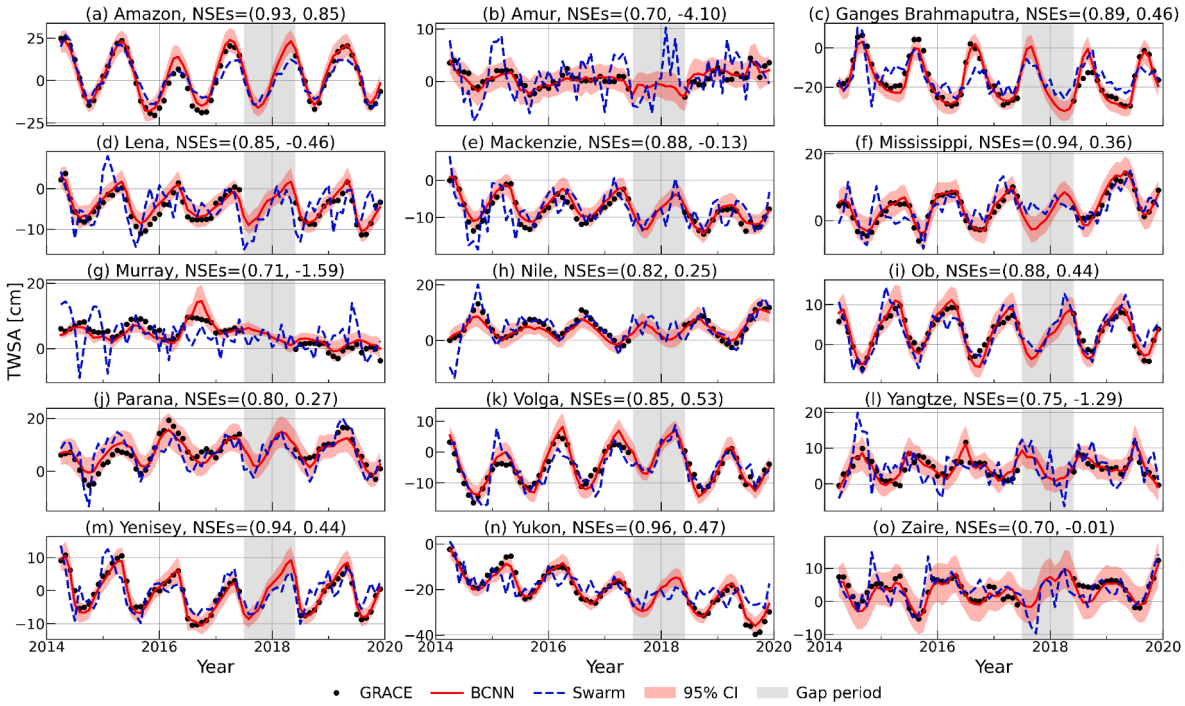

Fig. 14. Basin-averaged TWSA time series derived from GRACE, BCNN, and Swarm during April 2014 and December 2019. The first and second numbers in the bracket are the NSE values of BCNN and Swarm, respectively, with GRACE.

Table 3
Summary of the experimental settings in previous studies.

Reference	GRACE product	Spatial resolution	Training period	Testing period
Humphrey and Gudmundsson (2019)[a]	JPL mascon RL06	0.5° × 0.5°	–	Apr 2014–Jun 2017; Jun 2018–Jul 2019
Sun et al. (2020)	CSR mascon RL06 v01	1° × 1°	Apr 2002–Jan 2014	Feb 2014–Jun 2017
Li et al. (2021)	CSR mascon RL06 v02	0.5° × 0.5°	Apr 2002–Jun 2017	Jun 2018–Jun 2020

[a] The product driven by the ERA5 hydroclimatic data is used for comparison.

differences in their amplitudes lead to relatively large deviations in many basins (e.g., Amur, Murray, Yangtze). On the contrary, BCNN TWSAs fit appreciably better with those of GRACE with much higher NSEs, suggesting BCNN's higher reliability in bridging the GRACE and GRACE-FO gap.

4.4. Comparison with previous studies

Here we restrict the comparison with Humphrey and Gudmundsson (2019), Sun et al. (2020), and Li et al. (2021), who provided publicly accessible global-scale TWSA prediction products. The predicted TWSA product by Humphrey and Gudmundsson (2019) is known as GRACE-REC. The original GRACE-REC dataset provides the detrended and deseasonalized TWSAs. We add the trend and seasonal signals obtained from the GRACE TWSAs and Humphrey et al. (2017), respectively, to the original GRACE-REC TWSAs for consistency. The TWSA product generated in Sun et al. (2020) with a deep fully-connected neural network is used here for comparison. The spatial resolution of predicted TWSAs in Humphrey and Gudmundsson (2019) and Li et al. (2021) is 0.5° × 0.5°. For consistency, we predict the TWSAs at the same resolution and thus the input/output image size of BCNN is $H \times W = 288 \times 720$. For a fair comparison, the GRACE TWSA data and the training periods used for BCNN network training are respectively the same as those employed in the three studies. The detailed descriptions of the three TWSA products are summarized in Table 3.

The comparison results are shown in Fig. 15. For simplicity, we show in the plot only the NSE metric as it measures directly the matching quality in terms of both magnitude and phase between the predicted and target time series. In addition, we also compare separately the performances in the hyper-arid, arid, semi-arid, semi-humid, and humid regions in Fig. 16, which summarizes the boxplots of the gridded NSE values in the five climate regions. The medians of the boxplots are listed in Tables 4. Figs. 15 and 16 clearly suggest BCNN's better performance relative to the three previous methods (Humphrey and Gudmundsson, 2019; Li et al., 2021; Sun et al., 2020), which obtain relatively high accuracy in the humid/semi-humid regions but their performances decrease in the hyper-arid/arid/semi-arid regions. Our BCNN method successfully improves the prediction accuracy in these hyper-arid/arid/semi-arid regions to a relatively high level. For instance, compared to Humphrey and Gudmundsson (2019), our BCNN improves the median NSE values in the hyper-arid, arid, and semi-arid regions from −0.41 −0.01, and 0.39, respectively, to 0.17, 0.57, and 0.69 (Table 4).

The results suggest BCNN's superior performance in providing improved TWSA predictions to bridge the GRACE and GRACE-FO gap, which is attributed jointly to the use of GRACE trend (Section 2.3) and BCNN's outstanding capability in learning the high-dimensional and highly-complex mappings between the TWSA and hydroclimatic inputs. Two additional noteworthy merits of BCNN compared to prior methods are the few assumptions/preprocessing involved and its ability to handle directly the global scale. Note that we set in BCNN for simplicity all inputs and outputs to the same spatial resolution and consider only the land components. It is flexible and straightforward for BCNN to handle inputs/outputs with different sizes and ocean components (e.g., sea surface temperature). This property benefits from the flexibility of CNN in architecture design and performing downsampling/upsampling, concatenation, and many other operations.

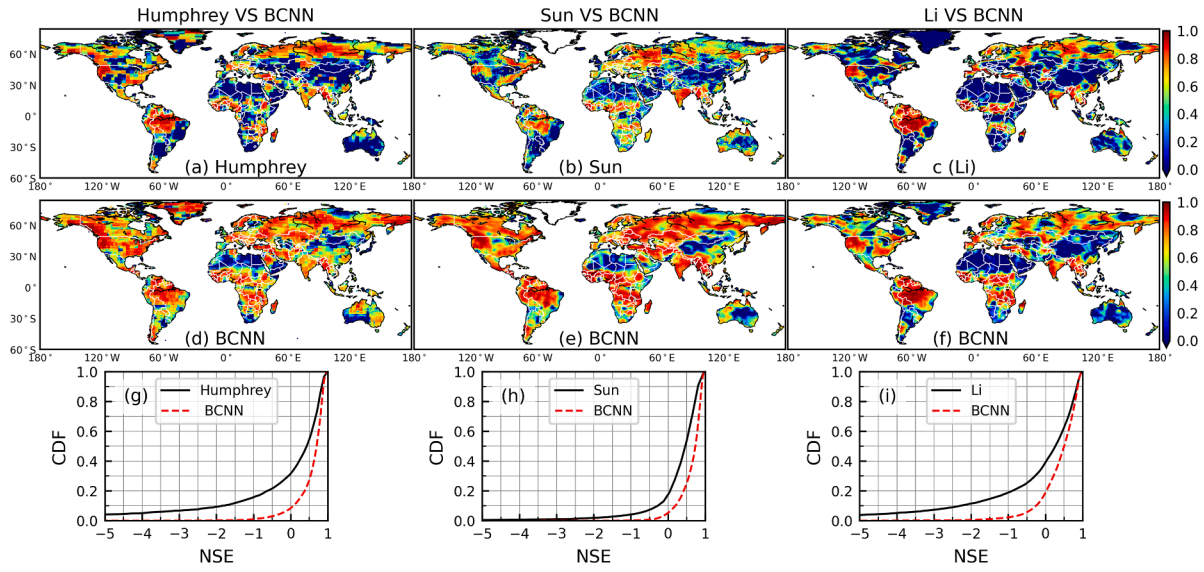

Fig. 15. (a–f) Comparison of BCNN's NSE values with those obtained by Humphrey and Gudmundsson (2019) (left), Sun et al. (2020) (middle), and Li et al. (2021) (right). The cumulative distribution functions (CDFs) of NSE values are compared in (g-i).

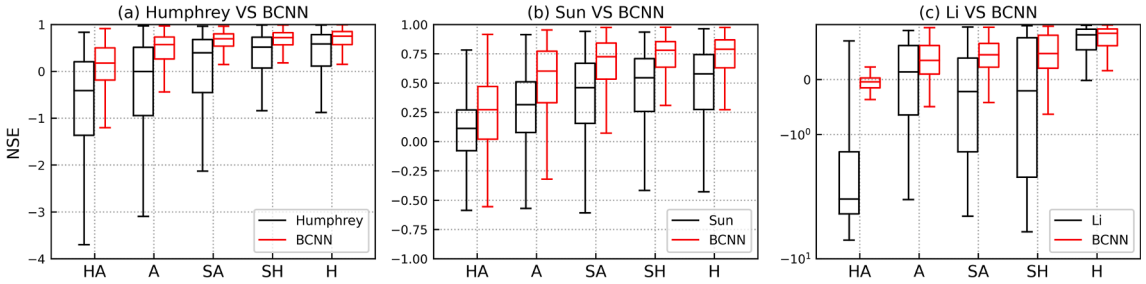

Fig. 16. Comparison of BCNN's boxplots of the NSE values at grids in the hyper-arid (HA), arid (A), semi-arid (SA), semi-humid (SH), and humid (H) regions with those obtained by (a) Humphrey and Gudmundsson (2019),Sun et al. (2020), and (c) Li et al. (2021).

Table 4
Comparison of BCNN's medians of the NSE values at grids in the hyper-arid (HA), arid (A), semi-arid (SA), semi-humid (SH), and humid (H) regions with those obtained by Humphrey and Gudmundsson (2019), Sun et al. (2020), and Li et al. (2021). BCNN's results are shown in the brackets and bold value indicates the better performance.

Comparison	HA	A	SA	SH	H
Humphrey VS BCNN	−0.41 (**0.17**)	−0.01 (**0.57**)	0.39 (**0.69**)	0.51 (**0.72**)	0.58 (**0.75**)
Sun VS BCNN	0.11 (**0.27**)	0.31 (**0.60**)	0.46 (**0.72**)	0.54 (**0.78**)	0.58 (**0.79**)
Li VS BCNN	−2.49 (**−0.04**)	0.14 (**0.35**)	−0.22 (**0.45**)	−0.21 (**0.47**)	0.81 (**0.84**)

5. Conclusions

In this study, we propose a deep learning-based BCNN method, driven by ERA5L hydroclimatic data, to fill the one-year TWSA observation gap between the GRACE and GRACE-FO satellites at the global scale. The integration of residual-skip connections, spatial-channel attentions, and Bayesian training strategy in BCNN enables it to effectively extract informative features for TWSA predictions from multiple predictor data and quantify the predictive uncertainties. Results show that BCNN successfully captures TWSA's complex spatiotemporal patterns. The comparisons with reanalyzed/simulated TWSA products, Swarm solution, and three previous studies further suggest BCNN's clearly higher gap-filling performance, particularly in the relatively arid regions. The gap-filling quality in maintaining the data continuity is further validated and confirmed through comparison with the standardized CPC soil moisture and Noah-simulated TWSA at the basin scale. The improvements in restoring the missing TWSA signals can be of great significance for applications desiring continuous data records in the time series analysis, correcting the bias in long-term trend estimates due to missing data, and enhancing the reliability of hydrological model predictions.

The outperformance of BCNN is mainly attributed to the use of TWSA trends, which are derived from the available GRACE (-FO) data in the pre- and post-gap periods, and its outstanding performance in feature extraction. The long-term TWSA trends induced by anthropogenic and/or natural factors are usually challenging-to-learn (Humphrey and Gudmundsson, 2019; Li et al., 2020; Sun et al., 2020). The utilization of this trend information makes full use of the existing data and essentially eases the learning task for BCNN. The BCNN's capability for informative feature extraction inherits the outstanding performance of CNN in image processing, which is further enhanced by integrating recent advances in deep learning. Note that we are concerned with bridging the gap between GRACE and GRACE-FO in the current work. For the task reconstructing TWSAs in the pre-GRACE period, which is beyond the scope of this study, the trend information is unavailable. The performance of BCNN for such a task remains to be explored.

CRediT authorship contribution statement

Shaoxing Mo: Conceptualization, Methodology, Investigation,

Formal-analysis, Validation, Writing-original-draft. **Ehsan Forootan:** Data-curation, Validation, Writing-review-editing. **Nooshin Mehrnegar:** Data-curation, Writing-review-editing. **Xin Yin:** Data-curation, Writing-review-editing. **Jichun Wu:** Supervision, Funding-acquisition, Writing-review-editing. **Wei Feng:** Supervision, Data-curation, Validation, Writing-review-editing. **Xiaoqing Shi:** Supervision, Funding-acquisition, Writing-review-editing.

Declaration of Competing Interest

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

Acknowledgments

The JPL and CSR GRACE Mascon data used in this study are available at https://podaac.jpl.nasa.gov/dataset/TELLUS_GRACGRFO_MASCON_CRI_GRID_RL06_V2 and http://www2.csr.utexas.edu/grace/RL06_mascons.html, respectively; ERA5-Land data are available at https://doi.org/10.24381/cds.68d2bb30; Noah TWSA dataset is downloaded from https://disc.gsfc.nasa.gov/datasets/GLDAS_NOAH10_M_2.1/summary. We thank Dr. Vincent Humphrey, Dr. Zhangli Sun, and Dr. Fupeng Li for sharing their TWSA prediction products. This work was funded by the National Natural Science Foundation of China (41730856, 41874095, 41977157, 42002248, 42004073), China Postdoctoral Science Foundation (2020M681550), Jiangsu Planned Projects for Postdoctoral Research Funds (2020Z133), and Fundamental Research Funds for the Central Universities (020614380106). S. Mo thanks Dr. Yinhao Zhu from Qualcomm AI Research for his valuable suggestions on BCNN. We are also grateful to the Editor, Associate Editor, and three anonymous reviewers for their constructive comments. The predicted TWSA dataset generated in this work is available at https://doi.org/10.5281/zenodo.4589755.

References

AghaKouchak, A., Farahmand, A., Melton, F.S., Teixeira, J., Anderson, M.C., Wardlow, B. D., Hain, C.R., 2015. Remote sensing of drought: progress, challenges and opportunities. Rev. Geophys. 53 (2), 452–480.

Ahmed, M., Sultan, M., Elbayoumi, T., Tissot, P., 2019. Forecasting GRACE data over the African watersheds using artificial neural networks. Remote Sens. 11 (15), 1769.

Bezděk, A., Sebera, J., Teixeira da Encarnação, J., Klokočník, J., 2016. Time-variable gravity fields derived from GPS tracking of Swarm. Geophys. J. Int. 205 (3), 1665–1669.

Chen, J., Li, J., Zhang, Z., Ni, S., 2014. Long-term groundwater variations in Northwest India from satellite gravity measurements. Global Planet. Change 116, 130–138.

da Encarnação, J.T., Arnold, D., Bezděk, A., Dahle, C., Doornbos, E., van den IJssel, J., Jäggi, A., Mayer-Gürr, T., Sebera, J., Visser, P., Zehentner, N., 2016. Gravity field models derived from Swarm GPS data. Earth, Planets Space 68 (1), 127.

Famiglietti, J.S., Lo, M., Ho, S.L., Bethune, J., Anderson, K.J., Syed, T.H., Swenson, S.C., de Linage, C.R., Rodell, M., 2011. Satellites measure recent rates of groundwater depletion in California's Central Valley. Geophys. Res. Lett. 38 (3), L03403.

Feng, W., Shum, C.K., Zhong, M., Pan, Y., 2018. Groundwater storage changes in China from satellite gravity: An overview. Remote Sens. 10 (5), 674.

Feng, W., Zhong, M., Lemoine, J.-M., Biancale, R., Hsu, H.-T., Xia, J., 2013. Evaluation of groundwater depletion in North China using the Gravity Recovery and Climate Experiment (GRACE) data and ground-based measurements. Water Resour. Res. 49 (4), 2110–2118.

Forootan, E., Kusche, J., Loth, I., Schuh, W.-D., Eicker, A., Awange, J., Longuevergne, L., Diekkrüger, B., Schmidt, M., Shum, C.K., 2014. Multivariate prediction of total water storage changes over West Africa from multi-satellite data. Surveys Geophys. 35 (4), 913–940.

Forootan, E., Schumacher, M., Mehrnegar, N., Bezděk, A., Talpe, M.J., Farzaneh, S., Zhang, C., Zhang, Y., Shum, C.K., 2020. An iterative ICA-based reconstruction method to produce consistent time-variable total water storage fields using GRACE and Swarm satellite data. Remote Sens. 12 (10), 1639.

Friis-Christensen, E., Lühr, H., Knudsen, D., Haagmans, R., 2008. Swarm – an Earth observation mission investigating geospace. Adv. Space Res. 41 (1), 210–216.

Gentine, P., Green, J.K., Guérin, M., Humphrey, V., Seneviratne, S.I., Zhang, Y., Zhou, S., 2019. Coupling between the terrestrial carbon and water cycles—a review. Environ. Res. Lett. 14 (8), 083003.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. Pattern Recogn. 77, 354–377.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Humphrey, V., Gudmundsson, L., 2019. GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. Earth Syst. Sci. Data 11 (3), 1153–1170.

Humphrey, V., Gudmundsson, L., Seneviratne, S.I., 2016. Assessing global water storage variability from GRACE: trends, seasonal cycle, subseasonal anomalies and extremes. Surveys Geophys. 37 (2), 357–395.

Humphrey, V., Gudmundsson, L., Seneviratne, S.I., 2017. A global reconstruction of climate-driven subdecadal water storage variability. Geophys. Res. Lett. 44 (5), 2300–2309.

Jing, W., Zhao, X., Yao, L., Di, L., Yang, J., Li, Y., Guo, L., Zhou, C., 2020. Can terrestrial water storage dynamics be estimated from climate anomalies? Earth and space. Science 7 (3) e2019EA000959.

Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (Eds.), 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, B., Rodell, M., Kumar, S., Beaudoing, H.K., Getirana, A., Zaitchik, B.F., de Goncalves, L.G., Cossetin, C., Bhanja, S., Mukherjee, A., Tian, S., Tangdamrongsub, N., Long, D., Nanteza, J., Lee, J., Policelli, F., Goni, I.B., Daira, D., Bila, M., de Lannoy, G., Mocko, D., Steele-Dunne, S.C., Save, H., Bettadpur, S., 2019. Global GRACE data assimilation for groundwater and drought monitoring: advances and challenges. Water Resour. Res. 55 (9), 7564–7586.

Li, F., Kusche, J., Chao, N., Wang, Z., Löcher, A., 2021. Long-term (1979-present) total water storage anomalies over the global land derived by reconstructing GRACE data. Geophys. Res. Lett. 48 (8), e2021GL093492.

Li, F., Kusche, J., Rietbroek, R., Wang, Z., Forootan, E., Schulze, K., Lück, C., 2020. 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 (5), e2019WR026551.

Liu, Q., Wang, D., 2016. Stein variational gradient descent: a general purpose Bayesian inference algorithm. In: Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R. (Eds.), Advances in Neural Information Processing Systems (NeurIPS), vol. 29. Curran Associates Inc., pp. 2378–2386

Long, D., Scanlon, B.R., Longuevergne, L., Sun, A.Y., Fernando, D.N., Save, H., 2013. GRACE satellite monitoring of large depletion in water storage in response to the 2011 drought in Texas. Geophys. Res. Lett. 40 (13), 3395–3401.

Long, D., Shen, Y., Sun, A., Hong, Y., Longuevergne, L., Yang, Y., Li, B., Chen, L., 2014. Drought and flood monitoring for a large karst plateau in Southwest China using extended GRACE data. Remote Sens. Environ. 155, 145–160.

Mehrnegar, N., Jones, O., Singer, M.B., Schumacher, M., Bates, P., Forootan, E., 2020. Comparing global hydrological models and combining them with GRACE by dynamic model data averaging (DMDA). Adv. Water Resour. 138, 103528.

Mehrnegar, N., Jones, O., Singer, M.B., Schumacher, M., Jagdhuber, T., Scanlon, B.R., Rateb, A., Forootan, E., 2021. Exploring groundwater and soil water storage changes across the CONUS at 12.5 km resolution by a Bayesian integration of GRACE data into W3RA. Sci. Total Environ. 758, 143579.

Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S., 2021. Long-tail learning via logit adjustment.

Meyer, U., Sosnica, K., Arnold, D., Dahle, C., Thaller, D., Dach, R., Jäggi, A., 2019. SLR, GRACE and Swarm gravity field determination and combination. Remote Sens. 11 (8).

Misra, D., 2019. Mish: a self regularized non-monotonic activation function. arXiv e-prints, arXiv:1908.08681.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2019. Deep autoregressive neural networks for high-dimensional inverse problems in groundwater contaminant source identification. Water Resour. Res. 55 (5), 3856–3881.

Mo, S., Zabaras, N., Shi, X., Wu, J., 2020. Integration of adversarial autoencoders with residual dense convolutional networks for estimation of non-Gaussian hydraulic conductivities. Water Resour. Res. 56 (2) e2019WR026082.

Mo, S., Zhu, Y., Zabaras, N., Shi, X., Wu, J., 2019. Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. Water Resour. Res. 55 (1), 703–728.

Muöoz Sabater, J., 2019. ERA5-Land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Nie, W., Zaitchik, B.F., Rodell, M., Kumar, S.V., Arsenault, K.R., Li, B., Getirana, A., 2019. Assimilating GRACE into a land surface model in the presence of an irrigation-induced groundwater trend. Water Resour. Res. 55 (12), 11274–11294.

Rateb, A., Scanlon, B.R., Pool, D.R., Sun, A., Zhang, Z., Chen, J., Clark, B., Faunt, C.C., Haugh, C.J., Hill, M., Hobza, C., McGuire, V.L., Reitz, M., Müller Schmied, H., Sutanudjaja, E.H., Swenson, S., Wiese, D., Xia, Y., Zell, W., 2020. Comparison of groundwater storage changes from GRACE satellites with monitoring and modeling of major U.S. aquifers. Water Resour. Res. 56 (12) e2020WR027556.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204.

Richey, A.S., Thomas, B.F., Lo, M.-H., Reager, J.T., Famiglietti, J.S., Voss, K., Swenson, S., Rodell, M., 2015. Quantifying renewable groundwater stress with GRACE. Water Resour. Res. 51 (7), 5217–5238.

Richter, H.M.P., Lück, C., Klos, A., Sideris, M.G., Rangelova, E., Kusche, J., 2021. Reconstructing GRACE-type time-variable gravity from the Swarm satellites. Sci. Rep. 11, 1117.

Rodell, M., Famiglietti, J.S., Wiese, D.N., Reager, J.T., Beaudoing, H.K., Landerer, F.W., Lo, M.H., 2018. Emerging trends in global freshwater availability. Nature 557 (7707), 651–659.
Rodell, M., Houser, P.R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J.K., Walker, J.P., Lohmann, D., Toll, D., 2004. The global land data assimilation system. Bull. Am. Meteorol. Soc. 85 (3), 381–394.
Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241.
Scanlon, B.R., Zhang, Z., Rateb, A., Sun, A., Wiese, D., Save, H., Beaudoing, H., Lo, M.H., Müller-Schmied, H., Döll, P., van Beek, R., Swenson, S., Lawrence, D., Croteau, M., Reedy, R.C., 2019. Tracking seasonal fluctuations in land water storage using global models and GRACE satellites. Geophys. Res. Lett. 46 (10), 5254–5264.
Scanlon, B.R., Zhang, Z., Save, H., Sun, A.Y., Müller Schmied, H., van Beek, L.P.H., Wiese, D.N., Wada, Y., Long, D., Reedy, R.C., Longuevergne, L., Döll, P., Bierkens, M. F.P., 2018. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. Proc. Nat. Acad. Sci. 115 (6), E1080–E1089.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resour. Res. 54 (11), 8558–8593.
Soltani, S.S., Ataie-Ashtiani, B., Simmons, C.T., 2021. Review of assimilating GRACE terrestrial water storage data into hydrological models: advances, challenges and opportunities. Earth Sci. Rev. 213, 103487.
Su, Y.-C., Grauman, K., 2017. Learning spherical convolution for fast features from 360 imagery. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. vol. 30. Curran Associates Inc., pp. 529–539.
Sun, A.Y., Scanlon, B.R., 2019. How can big data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environ. Res. Lett. 14 (7), 073001.
Sun, A.Y., Scanlon, B.R., Save, H., Rateb, A., 2020. Reconstruction of GRACE total water storage through automated machine learning. Water Resour. Res. 57 e2020WR028666.
Sun, A.Y., Scanlon, B.R., Zhang, Z., Walling, D., Bhanja, S.N., Mukherjee, A., Zhong, Z., 2019. Combining physically based modeling and deep learning for fusing GRACE satellite data: can we learn from mismatch? Water Resour. Res. 55 (2), 1179–1195.
Sun, Z., Long, D., Yang, W., Li, X., Pan, Y., 2020. Reconstruction of GRACE data on changes in total water storage over the global land surface and 60 basins. Water Resour. Res. 56 (4) e2019WR026250.
Tapley, B.D., Watkins, M.M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., Sasgen, I., Famiglietti, J.S., Landerer, F.W., Chambers, D.P., Reager, J.T., Gardner, A. S., Save, H., Ivins, E.R., Swenson, S.C., Boening, C., Dahle, C., Wiese, D.N., Dobslaw, H., Tamisiea, M.E., Velicogna, I., 2019. Contributions of GRACE to understanding climate change. Nat. Clim. Change 9, 358–369.
van den Dool, H., Huang, J., Fan, Y., 2003. Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001. J. Geophys. Res.: Atmosp. 108 (D16).
Wang, F., Shen, Y., Chen, Q., Wang, W., 2021. Bridging the gap between GRACE and GRACE follow-on monthly gravity field solutions using improved multichannel singular spectrum analysis. J. Hydrol. 125972.
Watkins, M.M., Wiese, D.N., Yuan, D.-N., Boening, C., Landerer, F.W., 2015. Improved methods for observing Earth's time variable mass distribution with GRACE using spherical cap mascons. J. Geophys. Res.: Solid Earth 120 (4), 2648–2671.
Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV).
Yan, X., Zhang, B., Yao, Y., Yang, Y., Li, J., Ran, Q., 2021. Grace and land surface models reveal severe drought in eastern china in 2019. J. Hydrol. 601, 126640.
Yi, S., Sneeuw, N., 2021. Filling the data gaps within GRACE missions using singular spectrum analysis. J. Geophys. Res. Solid Earth 126 (5) e2020JB021227.
Yin, W., Han, S.-C., Zheng, W., Yeo, I.-Y., Hu, L., Tangdamrongsub, N., Ghobadi-Far, K., 2020. Improved water storage estimates within the North China Plain by assimilating GRACE data into the CABLE model. J. Hydrol. 590, 125348.
Yin, Z., Xu, Y., Zhu, X., Zhao, J., Yang, Y., Li, J., 2021. Variations of groundwater storage in different basins of China over recent decades. J. Hydrol. 598, 126282.
Zaitchik, B.F., Rodell, M., Reichle, R.H., 2008. Assimilation of GRACE terrestrial water storage data into a land surface model: results for the Mississippi River Basin. J. Hydrometeorol. 9 (3), 535–548.
Zhong, Y., Zhong, M., Feng, W., Zhang, Z., Shen, Y., Wu, D., 2018. Groundwater depletion in the West Liaohe River Basin, China and its implications revealed by GRACE and in situ measurements. Remote Sens. 10 (4), 493.
Zhu, Y., Zabaras, N., 2018. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. J. Comput. Phys. 366, 415–447.