Science of the Total Environment 830 (2022) 154701

Contents lists available at ScienceDirect

Science of the Total Environment

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

Bridging the gap between GRACE and GRACE-FO missions with deep learning aided water storage simulations

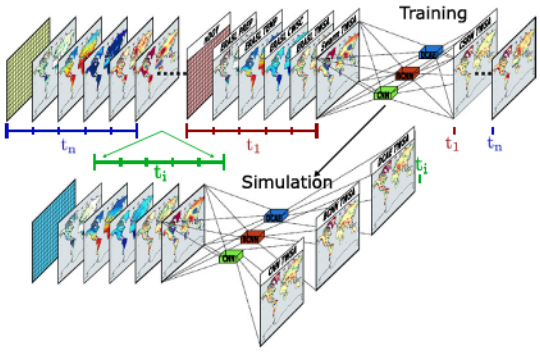
Metehan Uz [a], Kazım Gökhan Atman [b], Orhan Akyilmaz [a,*], C.K. Shum [c,d], Merve Keleş [a], Tuğçe Ay [a], Bihter Tandoğdu [a], Yu Zhang [c], Hüseyin Mercan [e]

[a] Dept. of Geomatics Eng., Istanbul Technical University, Istanbul, Turkey
[b] School of Mathematical Sciences, Queen Mary University of London, London, UK
[c] Division of Geodetic Science, School of Earth Sciences, Ohio State University, Columbus, OH, USA
[d] Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan, China
[e] Dept. of Geomatics Eng., Çanakkale Onsekiz Mart University, Çanakkale, Turkey

HIGHLIGHTS

- Data gaps within and between GRACE/-FO are effectively filled at high-resolution
- Using Swarm coarse resolution gravity solutions improved the TWSA simulations
- The importance end necessity of external validation with independent data are shown
- Climate-induced extreme hydrologic signals within the gap are successfully detected
- The simulated infilling product can reliably maintain the data continuity

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 13 December 2021
Received in revised form 13 March 2022
Accepted 16 March 2022
Available online 23 March 2022

Editor: Christian Herrera

Keywords:
GRACE
GRACE-FO
Swarm
Deep learning neural networks
Terrestrial water storage anomaly
Groundwater storage

ABSTRACT

The monthly high-resolution terrestrial water storage anomalies (TWSA) during the 11-months of gap between GRACE (Gravity Recovery And Climate Experiment) and its successor GRACE-FO (-Follow On) missions are missing. The continuity of the GRACE-like TWSA series with commensurate accuracy is of great importance for the improvement of hydrologic models both at global and regional scales. While previous efforts to bridge this gap, though without achieving GRACE-like spatial resolutions and/or accuracy have been performed, high-quality TWSA simulations at global scale are still lacking. Here, we use a suite of deep learning (DL) architectures, convolutional neural networks (CNN), deep convolutional autoencoders (DCAE), and Bayesian convolutional neural networks (BCNN), with training datasets including GRACE/-FO mascon and Swarm gravimetry, ECMWF Reanalysis-5 data, normalized time tag information to reconstruct global land TWSA maps, at a much higher resolution (100 km full wavelength) than that of GRACE/-FO, and effectively bridge the 11-month data gap globally. Contrary to previous studies, we applied no prior detrending or de-seasoning to avoid biasing/aliasing the simulations induced by interannual or longer climate signals and extreme weather episodes. We show the contribution of Swarm and time inputs which significantly improved the TWSA simulations in particular for correct prediction of the trend component. Our results also show that external validation with independent data when filling large data gaps within spatio-temporal time series of geophysical signals is mandatory to maintain the robustness of the simulation results. The results and comparisons with previous studies and the adopted DL methods demonstrate the superior performance of DCAE. Validations of our DCAE-based TWSA simulations with independent datasets, including in situ groundwater level, Interferometric Synthetic Aperture

* Corresponding author.
E-mail address: akyilma2@itu.edu.tr (O. Akyilmaz).

http://dx.doi.org/10.1016/j.scitotenv.2022.154701

Radar measured land subsidence rate (e.g. Central Valley), occurrence/timing of severe flash flood (e.g. South Asian Floods) and drought (e.g. Northern Great Plain, North America) events occurred within the gap, reveal excellent agreements.

1. Introduction

GRACE (Gravity Recovery And Climate Experiment) and its successor, GRACE-FO (-Followon) gravimetry satellite missions are designed to detect changes in Earth's global gravity field at monthly sampling and a spatial scale longer than 333 km (half-wavelength) (Tapley et al., 2004, 2019). The GRACE gravity-inferred mass change data were collected for a much longer than expected mission time span (expected lifetime was 5 years) at 15 years, from 17 March 2002 through June 2017. GRACE-FO twin-satellite gravity mission was launched in May 2018 as a successor mission to ensure the continuity of GRACE mission (Flechtner et al., 2016; Kornfeld et al., 2019). However, there is a significant data gap between the GRACE and GRACE-FO (GRACE/-FO) missions, which covers a period of successive 11 months. In addition, there exist other multiple data gaps within GRACE/-FO solution time series for individual months either due to satellite maneuvers performed to extend the mission's lifetime, or lack of healthy data/insufficient ground track coverage such as short repeat-cycling orbits (Klokočník et al., 2015). One possible solution addressed by the geoscience community is to fill these gaps (at least beyond 2014) using data from the European Space Agency's Earth Explorer Mission Swarm three-satellite constellation. Swarm satellites were launched in November 2013 and currently in operation, with its primary scientific objective to map Earth's magnetic field and its temporal variations (Olsen et al., 2013). Additional objective is the mapping of Earth's temporal gravity field via high (GPS)-low (Swarm spacecrafts) satellite-to-satellite tracking (hl-SST) at low spatial resolution (Encarnação et al., 2016, 2020), at spherical harmonics (SH) complete to degree and order (d/o) 12, longer than 3000 km (full wavelength).

Several institutes have produced monthly gravity field solutions from Swarm hl-SST data using different approaches and orbit solutions (Guo et al., 2015; Bezděk et al., 2016; Jäggi et al., 2016; Lück et al., 2018; Encarnação et al., 2019). Five official centers including Astronomical Institute of the Czech Academy of Science (ASU), Astronomical Institute of the University of Bern (AIUB), Institute of Geodesy (IfG) of the Graz University of Technology, Institute of Geodesy and Geoinformation (IGG) of the University of Bonn and Division of Geodetic Science, School of Earth Sciences, Ohio State University (OSU) have been routinely providing monthly SH Swarm solutions. These Swarm solutions effectively are capturing long wavelength temporal gravity field only at resolutions commensurate to SH d/o 12 (Lück et al., 2018; Encarnação et al., 2020) although they are typically expanded up to d/o 40 terms.

Contemporary studies so far used Swarm gravimetry data aided by statistical approaches, such as the Independent Component Analysis (ICA, Forootan et al., 2020), and Multichannel Singular Spectrum Analysis (MSSA, Wang et al., 2021). More recently, Swarm gravimetry was used in a purely data-driven approach to reveal the return of the rapid mass loss state in West Antarctica during the absence of GRACE in 2017–2018 (Zhang et al., 2021). These studies thus succeeded in bridging satellite gravimetry data gaps at long spatial scales, limited by Swarm data resolution (and accuracy). The lack of GRACE-like resolution (666 km full-wavelength) terrestrial water storage anomaly (TWSA) data during data gaps would likely degrade assimilative hydrological studies resulting in large uncertainties including biases. Thus, it is crucial to fill the TWSA data gap with realistic simulations and at least at GRACE-type resolutions. Other studies used different methods to directly or indirectly fill the gap between GRACE/-FO, and focused either only on temporal (mostly regional or basin-wide average) or on spatio-temporal (grid-wise and covering all or most part of the Earth) TWSA modeling. Among these studies, Humprey and Gudmundsson (2019) first computed monthly TWSA using land surface temperature (TEMP) and precipitation (PPT) from three different meteorological data sources adopting basic principles of hydrologic modeling at each grid cell of corresponding GRACE mascon (mass concentration) solutions. Then for each grid cell, a scale factor but between the de-trended and de-seasoned time series of simulated TWSA and mascon TWSA is empirically estimated.

Sun et al. (2019) used convolutional neural networks (CNN) to model the discrepancy between the TWSA from GLDAS NOAH (Rodell et al., 2004) land surface model (LSM) and the GRACE TWSA over India; however not focused on gap filling but mentioned its potential. Yu et al. (2021) performed a similar study over Canada, using three different deep learning (DL) architectures namely CNN, conditional generative adversarial networks (cGAN) and deep convolutional auto encoders (DCAE); they modelled the relationship between (input) regional LSM-derived TWSA and output GRACE TWSA and used their final models to reconstruct GRACE-like TWSA prior to the GRACE era, (from 1979 to 2002). They also showed that reconstruction using LSM-based deep learning modeling produces significantly better results than modeling using only GRACE data. Li et al. (2020), at the first step has applied signal decomposition using ICA and principal component analysis to 1° × 1° gridded time series of all input data variables (including PPT, TEMP, sea surface temperature-SST and 17 other climate indices) and output variable GRACE TWSA to separate the signals into spatial patterns and temporal modes; then in the next step these temporal modes are further separated into trend, annual, inter-annual and residual components. In the third step, the temporal modes (except the trend signal) of the input variables are used to predict those of GRACE TWSA through three methods; artificial neural networks (ANN), autoregressive exogenous (ARX) and multiple linear regression (MLR). Finally, the GRACE-like TWSA were reconstructed by adding spatial patterns and trend from GRACE TWSA back to the three temporal modes predicted in the previous step.

Sun et al. (2020a) used three methods, namely deep neural networks (DNN), MLR and seasonal autoregressive integrated moving average with exogenous variables (SARIMAX) to model the nonlinear relationship between the input hydrometeorological variables plus the GLDAS-NOAH derived TWSA and the output GRACE-derived TWSA (both from mascon solutions and official L2 data products), and demonstrated that DNN and SARIMAX achieve (both globally and regionally) comparable results and outperform MLR method in 60 basins Worldwide. Besides, they used all the GRACE data (covering 2002–2018) and did not attempt to fill the gap but noted that the studied methods can be used for such a task. Forootan et al. (2020) first applied ICA to entire GRACE (2003–2018) 1° × 1° gridded TWSA time series to retrieve the dominant modes of the TWSA signal, and then these modes are used for reconstruction process in combination with Swarm data to obtain refined Swarm temporal gravity fields complete to d/o 40 with considerably reduced noise. Sun et al. (2020b) has used several machine learning (ML) techniques to reconstruct GRACE-like TWSA over the U.S. by using several input data variables including GLDAS TWSA, TEMP and PPT from ERA5-Land (ERA5-L), sea surface temperature, and two climate indices [North Atlantic Oscillation (NAO), and Multivariate ENSO Index (MEI)] with multiple monthly time lags of various lengths, and arguably concluded that no single method is superior to another throughout the U.S. and suggested the combination of results from multiple ML methods.

Richter et al. (2021) combined the principal components of global (including oceans) gridded GRACE/-FO mass change time series (in terms of EWH, or equivalent water height) at low resolution (d/o 12) monthly Swarm solutions to reconstruct de-noised temporal gravity field models up to d/o 40 between December 2013 and December 2018. Wang et al. (2021) applied MSSA using all the monthly GRACE/-FO RL06 spherical harmonics solutions (April 2002–March 2020) provided by University of

Texas Center for Space Research (UTCSR) to fill the data gaps; monthly Swarm solutions were only used for comparisons, which confirm that the TWSA computed from reconstructed fields are more consistent with hydrologic models than those computed from Swarm-only solutions. Li et al. (2021) basically extended the work in Li et al. (2020) beyond 26 basins to global (excluding polar regions) land surface and additionally estimated/determined scaling factors (which they call 'application scale') for each grid cell defined on the land to refine the monthly TWSA reconstruction for the period 1979–2020. Yi and Sneeuw (2021) used singular spectrum analysis (SSA) to fill the data gap within and between GRACE/-FO, but in spherical harmonics domain and showed that their approach is able to retain GRACE-like resolution temporal gravity fields up to d/o 30 for the missing months. They used Swarm solutions, again for comparison only. Löcher and Kusche (2021) combined the coarse resolution (d/o 10) monthly Satellite Laser Ranging (SLR) gravity solutions with the spatial patterns derived from the available GRACE mission by decomposing the series of monthly gravity field solutions into empirical orthogonal functions and produced monthly SH gravity field solutions (including the data gap period) which have same spatial resolution with GRACE.

Mo et al. (2021) used a Bayesian CNN (BCNN) method to identify the relationship between the input (hydroclimatic/meteorological observation data as well as TWSA computed from ERA5-L) and the output (1° × 1° gridded) GRACE/-FO mascons both de-trended between April 2002 and August 2020. Their model established the relationship between de-trended (input and output) signals. Moreover, they interpolated the intermittent monthly gaps which exist within individual GRACE and GRACE-FO data spans using mascon solutions of neighboring months and used all the GRACE/-FO mascon time series, including interpolated epochs, except the long inter-mission data gap of 11 months, to estimate the linear trend. The estimated mascon trend was then added backed to the de-trended TWSA simulations of BCNN model to reconstruct the original TWSA signal. We hypothesize that such a de-trending or de-seasoning processes may induce large errors prohibiting the exact capturing of the interannual or longer climate events, and the extreme regional climate signals, occurred within the 11-month long GRACE/-FO data gap.

The all the data driven methods (including ML/DL type methods) dedicated to fill the gap between GRACE/-FO listed above restore either the trend signal or even also the spatial patterns retrieved from existing GRACE/-FO data and assess the performance of their results based on residual signal which is the difference between their simulated TWSA and the observed GRACE/-FO TWSA. In this study, three DL architectures namely CNN, DCAE, and BCNN through TensorFlow (Abadi et al., 2016), Keras (Chollet et al., 2015) and PyTorch (Paszke et al., 2017) implementations, respectively, have been developed to reconstruct or simulate all the missing monthly TWSA maps within and between GRACE/-FO beyond January 2014 and until December 2020. Our study differs from the aforementioned studies in several aspects which also draw the novelty of the study: (i) No prior de-trending, de-seasoning, signal decomposition or interpolation of intermittent gaps is applied neither to the input nor to the output data set; all the data are used as they are, in order to avoid biasing or aliasing the simulations induced by interannual or longer term climate signals as well as extreme weather episodes. Instead, we introduced the normalized time epochs of the associated input data as additional input variables. The rationale for this idea is that almost all geophysical signals are functions of time, and it would be logical to allow the deep learning process to retrieve such trend and seasonal signals (which may show different spatio-temporal patterns) from the data itself. Another advantage of this approach is the ability of direct implementation of data in DL models. (ii) We use monthly coarse resolution (complete to d/o 12) Swarm TWSA solutions directly as another input to retrieve the long-wavelength component of the TWSA signal through deep learning. (iii) We used shorter time series of data for our modeling work; all the input and output data are those in the period December 2013–December 2020 since the Swarm data is available beyond December 2013. (iv) Finally, we compared results from all 3 DL methods to ensure robustness in our approach, and to identify the best DL approach for effective GRACE/-FO data gap bridging which eventually shows the necessity and importance of validation with independent (non-GRACE) data.

The rest of the paper is organized as follows. The data, methods and the performance metrics used in the study are briefly described in Section 2. The numerical results and discussions are given along with internal and external (with comparisons to independent data, hydrologic models and the results from previous studies) quality assessment of the simulated TWSA are given in Section 3. Finally, conclusions are drawn and some practical hints to be considered when using ML/DL methods in modeling geophysical processes are summarized in Section 4.

2. Material and methods

2.1. Data

Our DL models basically include six input and one output data variables. The single output variable is the monthly GRACE TWSA, while the input variables are monthly coarse resolution Swarm-derived TWSA, four hydroclimatic/meteorological parameters (PPT, TEMP, cumulative water storage change and model-derived TWSA) from ERA5-L hydrologic model and the time (in terms of normalized Day of Year – nDOY) epochs of the corresponding input data, respectively. The nDOY of a month is simply computed by dividing the DOY of the mid-day of that month by 365 (or 366), i.e. the number of days in a year. The data and the pre-processing steps applied before adjusting the parameters (also known as training or learning process) of the DL models are briefly described in the following. All three DL methods use the same input-output architecture; that is the input data at two successive monthly time epochs *t-1*, and *t* are used to approximate the corresponding output data, i.e. GRACE TWSA at time epoch *t*. The flowchart of the overall methodology and complete data preparation scheme and adopted input-output data patterns are presented in Figs. A1 and A2 (see Appendix), respectively.

2.1.1. GRACE TWSA data

The monthly mascon TWSA solutions of GRACE/-FO, from December 2013 till December 2020, released by UTCSR are used in this study (Save et al., 2016; Save, 2020). CSR RL06 (Release 06) mascons (CSRM) are the global 0.25° × 0.25° (noting that this is not the true spatial resolution, which is at 666 km, but the resampling size) gridded monthly TWSA computed as differences with respect to a mean-field in the period 2004.0–2009.999 and can be fragmented into two main parts regarding the coverage period of the two missions; GRACE (from April 2002 to July 2017) and GRACE-FO (from May 2018 to present). The gap between these missions includes a total of 11 months, however, there are also some intermittent missing months within each mission's individual data span. The standard corrections including degree-1 (Swenson et al., 2008), replacement of C_{20} and C_{30} coefficients using satellite laser ranging (SLR) solution (Cheng et al., 2013), ICE-6GD (VM5a) glacial isostatic adjustment forward model (Peltier et al., 2018) and ellipsoidal corrections (Ditmar, 2018) are all applied before the CSRM TWSA data are published (Save et al., 2016; Save, 2020). We resampled the original mascons to 1° × 1° grids to be consistent with current native resolution of CSR RL06 mascon solutions. We emphasize again that the CSR RL06 mascon solution at 25 km, is a re-sampled or interpolated gridded data, and not the real GRACE/-FO resolution which is at 666 km (full wavelength). Here our targeted 100-km resolution for our TWSA simulation is argued to be realistic, with the additional and independent datasets aided by DL approach. Further, the resampled monthly CSRM TWSA data constitutes the single output variable of the DL models in this study.

2.1.2. Swarm TWSA data

The monthly SH gravity field models (Level 2 - L2 data products) recovered from Swarm orbit tracking data up to d/o 40 published by International Center for Global Earth Models (ICGEM - http://icgem.gfz-potsdam.de/series/02_cost-g/swarm) are used (Encarnação et al., 2019, 2020). These models are made available in quarterly basis starting from

December 2013. Swarm L2 data is not directly used but converted to TWSA (Swarm TWSA) applying the same corrections as those in the CSRM processing chain using the updated version of GRACE MATLAB Toolbox (GRAMAT) Software (Feng, 2019). In order to suppress the noise in higher degree SH coefficients, we truncated the monthly SH models at d/o 12 and thus the long wavelength components of the gravitational signal could be obtained from Swarm solutions. Encarnação et al. (2020) suggested that a 750 km smoothing radius can be applied to retrieve time-variable gravity signals on land from Swarm models. Therefore, we applied Gaussian smoothing with a radius of 1000 km to the truncated Swarm L2 data to obtain the monthly TWS change. To make the Swarm-derived mass changes consistent with mascon TWSA, the mean-field of CSR RL06 L2 models between 2004.0 and 2009.999 is calculated and removed from monthly Swarm models. Finally, Swarm TWSA between December 2013 and December 2020 are resampled to the 1° × 1° grids. This data serves as one of the input variables of our DL architectures.

2.1.3. ERA5-L data

ERA5-L data set is released by European Centre for Medium-Range Weather Forecast (ECMWF - https://cds.climate.copernicus.eu) and has been produced by replaying the land components from the ECMWF ERA5 climate reanalysis dataset. ERA5-L data are publicly available from 1950 to present both as hourly products and monthly aggregates of hourly products and has a global coverage with 0.1° × 0.1° spatial resolution (Muñoz Sabater, 2019). The ERA5-L data set include various surface variables at such high resolution which are retrieved from vast amount and types of historic observations including those from satellites and in situ data sources using advanced modeling and data assimilation systems. The data from ERA5-L used in our DL modeling, in analogy to Mo et al. (2021), are monthly PPT, TEMP, cumulative water storage changes (CWSCs), and TWSA computed solely from ERA5-L. We applied exactly the same approach of Mo et al. (2021) for the computation of CWSC and ERA5-L TWSA: the TWSA is calculated from the water storage variables of soil moisture (in four layers from surface down to 289 cm), snow and canopy by summing up these variables and then removing their long term mean in the period 2004.0–2009.999 to be consistent with CSRM TWSA as described in Section 2.1.1. The CWSC is calculated by aggregating the cumulative differences between inflow (i.e. PPT) and outflow (i.e. evapotranspiration (ET), and runoff (RO)) at each grid cell through the water balance equation, i.e., Water Storage Change (WSC = PPT - ET - RO). Note that irrigation is not explicitly included in the water balance equation as ERA5-L ET data products include the irrigation effect on soil water storage to some extent. For details of the overall data preparation steps, the reader is referred to Eqs. (1–2) of Mo et al. (2021). Finally, all four ERA-derived hydro-/climatic input data variables (PPT, TEMP, CWSC and ERA5-L TWSA) are resampled to 1° × 1° grids to be consistent with other (input and output) variables of our DL architectures before running the learning process. The whole ERA5-L data suit provides more than seventy years of various meteorological and climate-related data which can be useful to study the impact of climate change, e.g. on watershed hydrology (Ekwueme and Agunwamba, 2020), streamflow dynamics (Oo et al., 2020), etc.

2.2. Deep learning models

2.2.1. CNN

CNN (Cun et al., 1989), is a class of neural networks that is advantageous in processing data which have a grid-like topology such as images or in the form of time-series. The main reason for this advantage, compared to conventional (feed-forward) ANN is due to the fact that the network utilizes mathematical linear operations which are called convolution. Additionally, CNNs are widely used in problems such as image change detection (Li et al., 2022), machine health diagnosis (Mukherjee and Tallur, 2021), land surface temperature reconstruction (Wu et al., 2019) and filling the remote sensing data gaps (Zhang et al., 2018). Particularly, in CNNs input is convoluted by the set of filters through the convolution layers. In this context, the relation between successive layers in the CNNs is represented as follows

$$a^{(l+1)} = \sigma\left(a^{(l)} * W^{(l)} + b^{(l)}\right) \tag{1}$$

where * is the convolution operator and σ denotes the activation function. In this way, it may be said that the use of CNN leads to effective way of feature extraction. Therefore, taking the so-called advantages into account, the CNN model has been developed to extract spatial features of input data. In this study, the network architecture of the CNN model consists of six convolutional layers with gradually decreasing filters size from 128 to 8 neurons which run through the activation function, namely exponential linear unit. Thus, the influence of nonlinearity is taken into account at this stage of the layers which is also called as detector stage. Additionally, since the use of pooling helps to ensure invariance of the CNN to the small translations of the input data, convolutional layers are followed by a max-pooling layer with 2 × 2 pool size (Zhou and Chellappa, 1988). Moreover, each layer is followed by dense layer which has 64, 32, 32, 32, 16, 1 neurons, respectively, and the first convolutional layer has also a regularizer which applies penalties on layer parameters, to prevent overfitting. In this study, all three DL models including the CNN model were trained for 250 epochs and the mean square error (MSE) was selected as the loss function. For the calculation of error by the loss function, Adamax optimizer, which is a variant of Adam algorithm (Kingma and Ba, 2015) was utilized.

2.2.2. DCAE

In addition to the CNN model, we also developed a DCAE network (Hinton and Zemel, 1994; Alain and Bengio, 2014; Kamyshanska and Memisevic, 2014) for reconstruction/simulation of GRACE-like TWSA maps. In general, DCAEs are the type of CNN which are used for representation learning by dimensionality reduction. DCAE mainly consists of two parts: encoder part for feature representation which represents by h = f (x) function and decoder part for reconstruction of input from representation by r = g(h). According to this, DCAE model takes input and maps to h

$$\boldsymbol{h} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}) \tag{2}$$

where σ represents the activation function, $\boldsymbol{W}$ and $\boldsymbol{b}$ are weight matrix and bias vector of encoder, respectively. Further, output of decoder is given as follows:

$$\boldsymbol{r} = \sigma\left(\hat{\boldsymbol{W}}\boldsymbol{h} + \hat{\boldsymbol{b}}\right)\Big) \tag{3}$$

where σ is activation function, $\widehat{W}$ corresponds to the weight matrix and $\widehat{b}$ is the bias vector of decoder. In this study, the network architecture of our model consists of five convolutional layers with gradually increasing filter size from 32 to 400. These are followed by flatten layer and a fully connected layer which has 256 neurons. In this manner, in the decoding part, the fully connected layer of 400 neurons is followed by the four transposed convolutional layers. Similar to the CNN model, the MSE metric is chosen as a loss function and the Adamax optimizer is employed for the training process.

2.2.3. BCNN

Besides the standard neural networks, surrogate models and Bayesian approaches to CNN can be seen as major developments for treating problems which have limited data for training. The underlying idea of the Bayesian approaches is the fact that the posterior probability is proportional to its prior probability and likelihood as $P(\theta|X) \propto P(X|\theta)P(\theta)$ (Goodfellow et al., 2016). Therefore, it may be said that the main difference between non-Bayesian and Bayesian approaches is that Bayesian neural networks (BNN) take epistemic uncertainty of model parameters into accounts by making use of the probability distributions. In the Bayesian treatment of

the learning, model weights are considered as unknown parameters with uncertainties and might be represented by probability distribution. Therefore, adding adaptive noise into weights may be considered as an effective way of reflecting this uncertainty. For this reason, we employ the model architecture which is proposed by Zhu and Zabaras (2018). In this approach, the process is described by probabilistic mapping as follows:

$$y = h(x, w) + \eta \tag{4}$$

Here, h(x,w) corresponds to an output of the neural network and η to additive noise. According to the proposed approach, the architecture of the baseline network is developed by making use of a dense convolutional encoder-decoder network (DenseNet) (Huang et al., 2017) which broadens the idea of ResNet (He et al., 2016). In this study, for simulation of GRACE-like TWSA, we adopt the proposed baseline network to our problem in a similar manner. Therefore, model architecture consists of the encoder part with convolution layer through the dense block and decoding part with consecutions of the encoding part, as shown in Fig. A3 (see Appendix). Additionally, Stein variational gradient descent (SVGD) is utilized to compute posterior distribution by minimizing the *Kullback–Leibler* (*KL*) divergence for the target distribution $p(\theta,x)$ and simpler distribution $Q(\theta;\lambda)$ such as $KL[Q(\theta;\lambda\|p(\theta,x)]$, where θ are stochastic parameters which are thus represented as random variables and λ represents observations such as $\lambda = [x^i, y^i]_{i=1}^{N_{mini}}$.Hence, we use the samples $\{\theta_i\}_{i=1}^{N_{mini}}$ to approximate the posterior distribution $p(\theta,x)$ which are optimized by making use of Adam algorithm. Details of the network architectures for all three DL models are given in Fig. A3.

2.3. Performance metrics

The performance of the simulations has been evaluated by two metrics: Nash - Sutcliffe efficiency (NSE, Nash and Sutcliffe, 1970) and root mean square error (RMSE), commonly used measures in hydrological studies and computed as follows

$$NSE = 1 - \frac{\sum_{i=0}^{N}(Y_i - O_i)^2}{\sum_{i=0}^{N}(O_i - \overline{O})^2}, \mathrm{NSE}\in(-\infty, 1] \tag{5}$$

where Y_i represents the values simulated by the model, O_i and $\overline{O}$ denote the observed and its mean values, respectively. Hence, NSE closer to 1 indicates a model with accurate simulative skill.

RMSE is a metric which is frequently used to measure regression performance of the models. In contrast to NSE, RMSE closer to 0 suggests a model with better performance.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - O_i)^2}, \mathrm{RMSE}\in[0, +\infty) \tag{6}$$

3. Results and discussions

Adopting the same input-output data patterns, the training of the three DL models is performed on a single NVIDIA GeForce GTX 2060 GPU. It is worth noting that the computer runtimes are approximately 4 h, 5 min and 75 min for CNN, DCAE and BCNN models, respectively.

Here, we organized the validation and evaluation of the simulation (reconstruction) results into three parts: The internal validation, comparison with previous studies and the external validation. The internal validation includes the global and regional (basin-wise) performance assessment of the simulated TWSA versus CSRM TWSA through the metrics given in Section 2.3. It basically shows the goodness of fit of the model simulations to the observed CSRM TWSA for chosen testing dataset, which is not used for training the DL models. It also includes comparisons with TWSA derived from independent hydrological models. The reconstructed TWSA from two previous studies (Li et al., 2021; Mo et al., 2021) are compared to our simulated TWSA using the test data set common for all three studies. The external validation, on the other hand, includes the comparison of the simulated TWSA with other non-gravimetry data set such as in situ groundwater level (GWL) measurements, the occurrence of extreme weather events such as flash floods and droughts, and annual land subsidence rates from InSAR (Interferometric Synthetic Aperture Radar), in particular during the GRACE/-FO data gap.

3.1. Internal validation

During the study period (December 2013–December 2020), though covering 84 months, only 61 months of mascon TWSA solutions are available. Considering the time lags for input data, i.e. *t-1* and *t* where *t* denotes the month, the input-output patterns corresponding to the so-called 61 months of CSRM TWSA have been generated. The 48 out of 61 months of the input-output data have been used to train the proposed DL models, while the remaining 13 months which were randomly chosen, have been used for testing/validating the model parameters. Therefore, roughly >20% of the entire data set have been separated for validation, which is necessary to avoid overfitting phenomena. The entire data gap between December 2013 and December 2020, including the 11 months of intermission period, consists of 23 months which are to be simulated using proposed DL methods.

The first part of internal validation is carried out considering the chosen 13 test months. Monthly mean global NSE and RMSE scores and their overall averages (dashed lines) for the entire test months are presented in Fig. 1. The overall average values from each DL models are 0.98 (DCAE), 0.98 (BCNN), 0.97 (CNN) for NSE and 2.7, 2.8, 3.5 cm for RMSE, respectively. While RMSE and NSE metrics for BCNN and DCAE simulations are consistent with each other, CNN shows slightly worse performance, especially in terms of RMSE. It is interesting to see the months with the highest RMSE and the lowest NSE values common for all three DL models in Fig. 1 are November 2016 and May 2017 (see green shaded area). While the gray shaded area shows the data gap between GRACE/-FO, the green shaded area is the battery turn-off period of the GRACE-B satellite from November 2016 to June 2017 (Save et al., 2018; Bandikova et al., 2019).

Therefore, the slightly worse performance metrics at these months can plausibly be explained by the battery issue of GRACE-B. According to Save et al. (2018), CSRM products between November 2016 and June 2017 are calculated using the operation with only one single working accelerometer (ACC), which is based on transplanting ACC data considering the attitude and time correction, to mitigate the battery issue of GRACE-B. In addition, CSRM solutions may not exactly be derived from the GRACE observations within a particular month; some solutions are computed using the integrated GRACE data partially observed in successive months depending on the observation quality and error sources of GRACE satellites, especially towards the end of GRACE mission lifetime. According to GRACE Science Data System Report (SDS) of November 2016 (https://isdc.gfz-potsdam.de/grace-isdc/grace-gravity-data-and-documentation/), the transplanted ACC for GRACE-B were used in the solution of CSR RL05 L2 models.

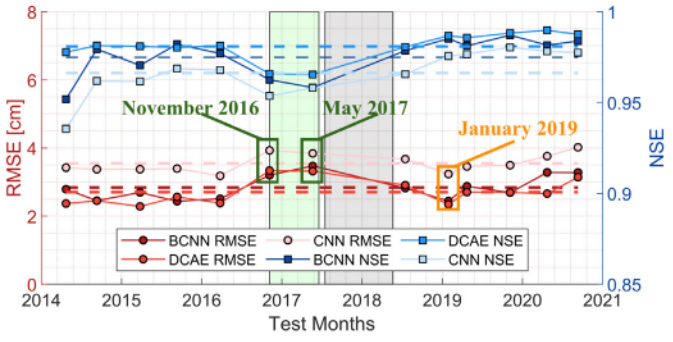


Fig. 1. Monthly mean and overall average global RMSE and NSE scores for 13 test months.

Moreover, the time span of the CSR products for this particular month is actually based on 28 days of data acquired between November 13 and December 10. The same data is used in the production of CSR RL06 L2 and CSRM TWSA. Similarly, for May 2017, only 20 days of GRACE data (between 3 and 23 May) have been used to compute CSRM solutions. Therefore, this may also explain the relatively low agreement of the TWSA simulations with CSRM at these test months as the corresponding CSRM solutions do not actually represent the mean TWSA within the pronounced months while the input data, on the other hand, are exactly the monthly mean values of daily observations within corresponding months.

The maximum (3.5 cm) and the minimum (2.6 cm) monthly mean global RMSE values (both from DCAE and BCNN) are obtained for May 2017 and January 2019, respectively (see Fig. 1). Simulated TWSA along with the corresponding CSRM TWSA for these two months are shown in Fig. 2a and b, respectively. Furthermore, differences between the simulated and the CSRM TWSA are also presented. In Fig. 2, the performance of the three DL models can be compared spatially; they all seem to be capable of capturing the spatio-temporal patterns of the TWSA. However, it appears that the differences (DIFF TWSA) of BCNN (Fig. 2a, b-v) from CSRM TWSA has a similar pattern with its simulated TWSA signal, which reveals that there is still significant mass anomaly signal in the residual (DIFF) TWSA; the amplitudes are underestimated especially at basins which show large seasonal variability such as Amazon, Zambezi and Greenland. This result also explains why Mo et al. (2021) applied BCNN to the de-trended data set and focused only on predicting seasonal component of TWSA signal within the gap period. In other words, BCNN is not effective to retrieve both trend and amplitude signal from the original GRACE TWSA when the normalized time and long wavelength gravity (such as TWSA from Swarm) data have jointly given as additional inputs. In order to achieve good simulations with BCNN without de-trending or de-seasoning the input-output data, the long-wavelength gravity information should be included as an additional input to the other four input variables,

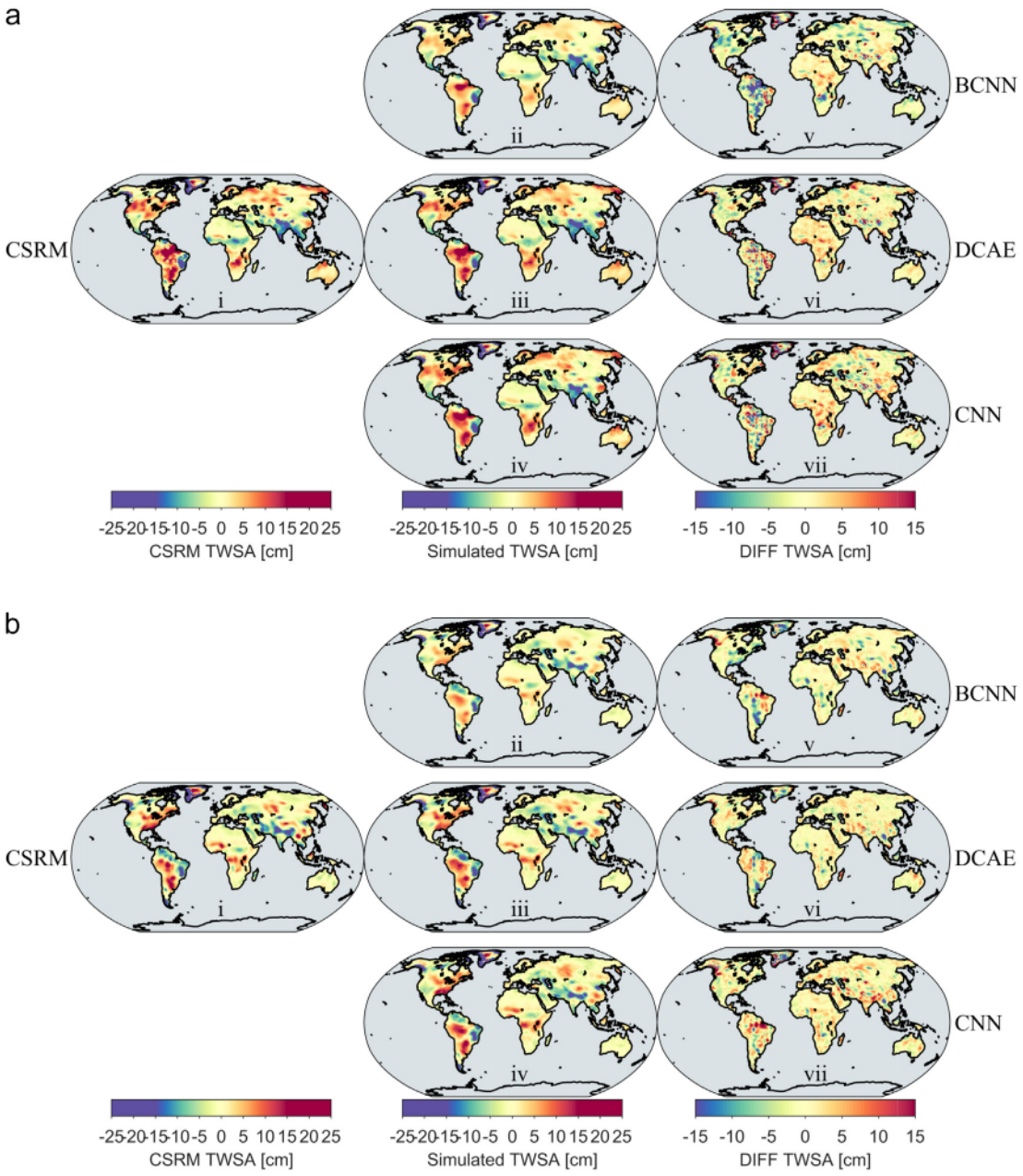


Fig. 2. Simulation results (a) for May 2017 and (b) for January 2019. The first columns include CSRM TWSA for these two test months and the second columns represent (ii) BCNN, (iii) DCAE and (iv) CNN simulations of TWSA for the corresponding months, respectively. The third columns show the differences (v–vi–vii) between CSRM and the corresponding simulated TWSA on the left (CSRM: CSR RL06 Mascon Solution, DIFF TWSA: differences between CSRM and simulated TWSA).

but not the normalized time since the time information is not a stochastic variable and does not conform to the probability distributions which violates the basic principles of BCNN (Keleş et al., 2021).

Relatively larger differences between the all simulated and CSRM TWSA are observed at hydrologically active basins such as Amazon and Greenland where the high temporal mass variations exist. We believe that the high differences at these basins are mainly due to the higher uncertainty of the TWSAs computed from hydrologic models (Scanlon et al., 2018) which are used as one of the major input variables in our DL-based simulative models. However, in general, high correlation between the simulated and CSRM TWSA can be clearly seen for both test months.

The DL-based TWSA simulations as well as ERA5-L and NOAH derived TWSA for entire test months are compared to corresponding CSRM TWSA and the spatial distributions of their RMSE and NSE are presented in Fig. 3. From top to bottom, the left column shows the RMSE for BCNN, DCAE, CNN, ERA5-L and NOAH, respectively, while the right column illustrates the NSE for the corresponding method or hydrologic model. Though ERA5-L and NOAH TWSA show reasonable correlations with CSRM at the regions which exhibit dominant hydrological signal (e.g. Amazon basin, Brahmaputra-Ganges basin) except for Greenland, they almost have no correlation at hyper-arid regions. On the other hand, the DL simulations of TWSA are clearly more consistent and have higher correlations with the CSRM almost at all regions.

An interesting observation from the simulated TWSA is that, regardless of the method used, not only in this study, but also in previous studies (e.g., see Fig. 2a of Humprey and Gudmundsson, 2019; Fig. 3b of Li et al., 2020; Fig. 2d of Li et al., 2021 and Fig. 4j-l of Mo et al., 2021), the highest RMSE values always appear in the same regions, namely, Amazon, Brahmaputra, Ganges and Zambezi basins, respectively. This result is consistent with the findings of Scanlon et al. (2018), which states that the hydrologic models generally underestimate the GRACE-TWSA trends; and the largest discrepancy between GRACE-TWSA and TWSA predicted by hydrologic models exists at these four large hydrologic basins. This is also evident from Fig. 4, where the (2014–2020) time series of CSRM, simulated TWSA from DL models and TWSAs computed from hydrologic models (ERA5-L and NOAH) at 10 selected basins [the boundaries of the river basins are derived from Total Runoff Integrating Pathway (TRIP) database (Oki and Sud, 1998)] with different hydrological characteristics (humid, semi-humid, arid, semi-arid) are shown. It can be seen that the hydrologic models underestimate the seasonal amplitudes at Amazon and partially at

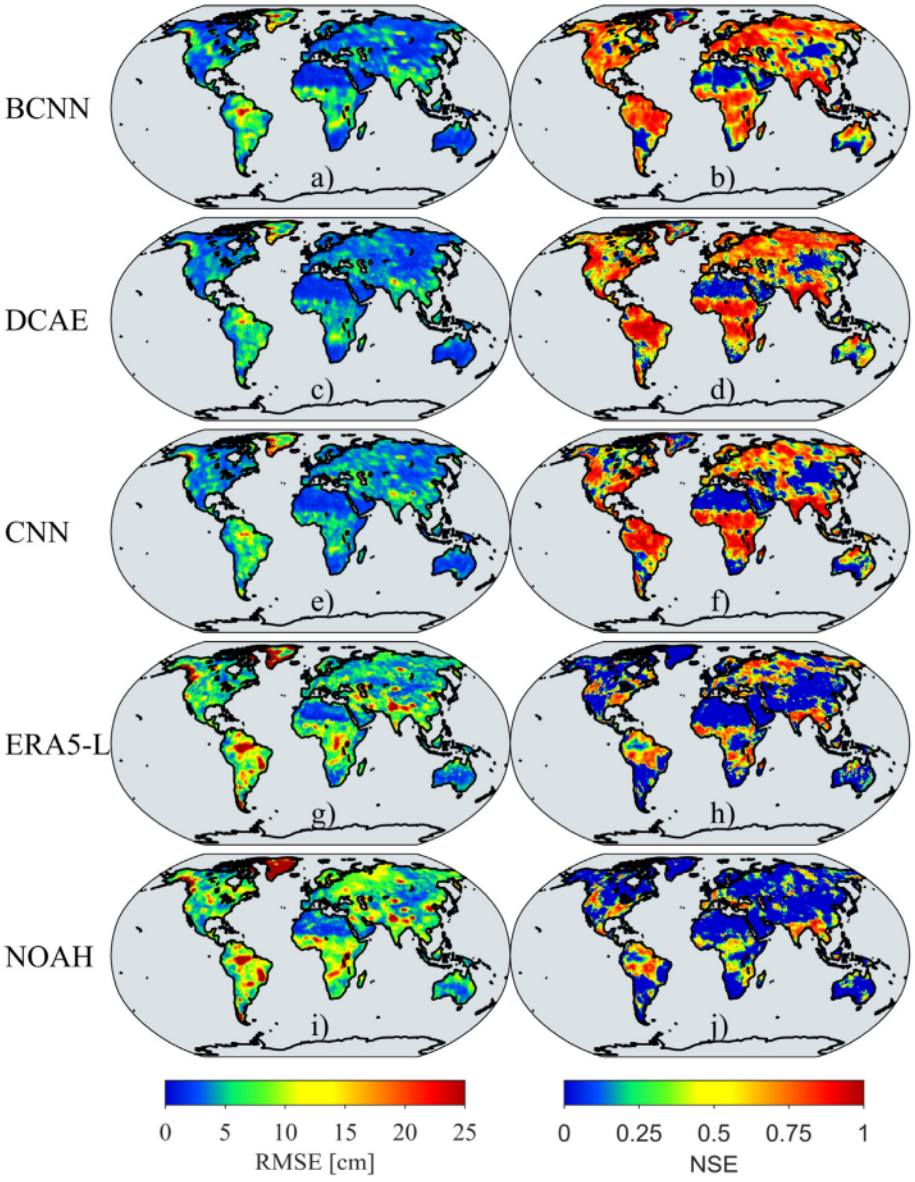


Fig. 3. Spatial distribution of RMSE (left) and NSE (right) scores for TWSA simulations of: (a, b) BCNN, (c, d) DCAE, (e, f) CNN, (g, h) ERA5-L and (i, j) NOAH, respectively. The performance metrics are computed using differences to the CSRM TWSA for 13 test months.

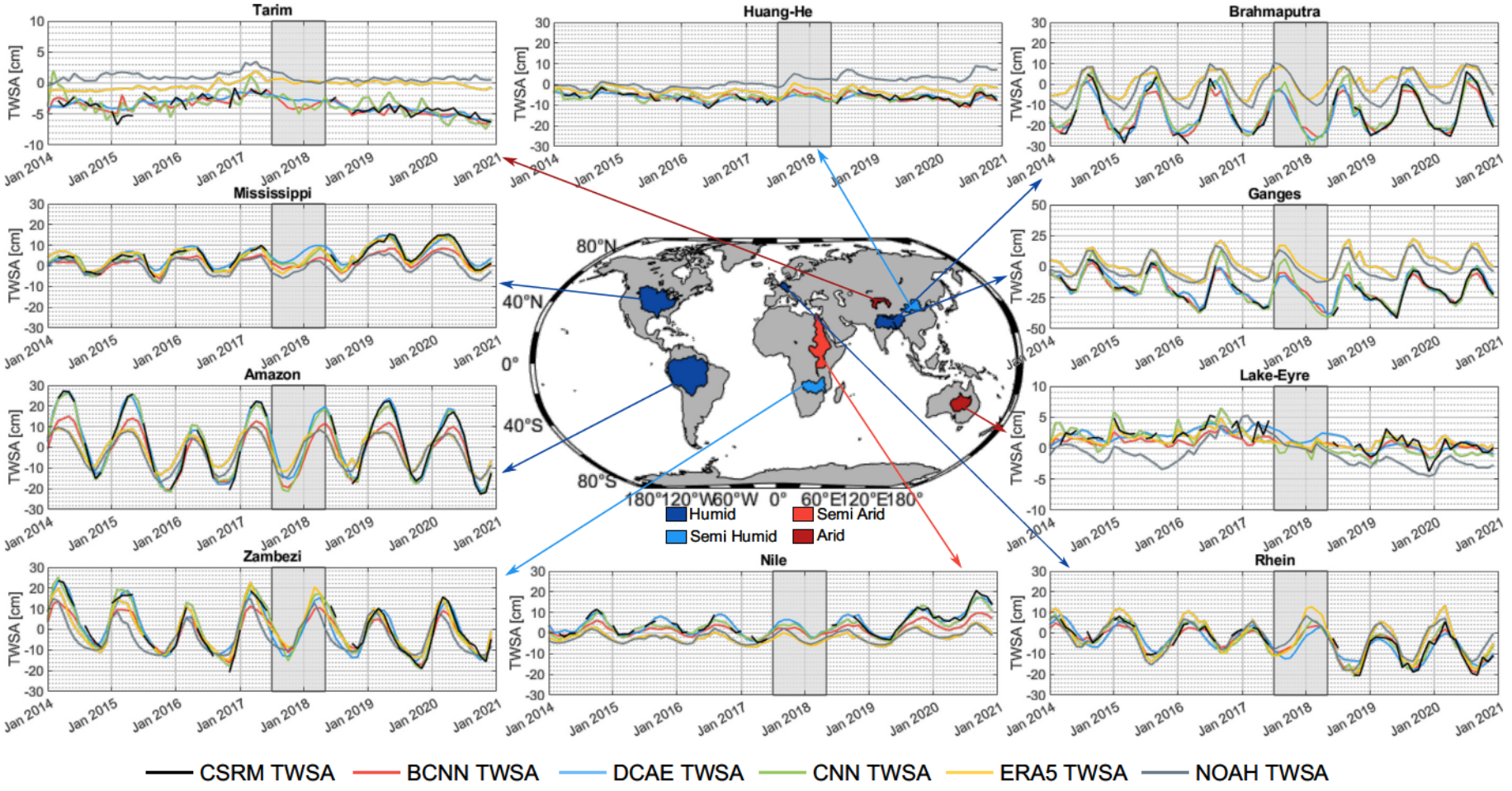


Fig. 4. Time series of TWSA signals from January 2014 to December 2020 for CSRM, BCNN, DCAE, CNN, ERA5-L and NOAH models. The basin borders are from Total Runoff Integrating Pathway (TRIP) database.

Zambezi basins compared to GRACE (both CSRM and simulated) TWSA, while in addition, a significant bias (~25 cm) exist at Ganges and Brahmaputra basins. The bias between the GRACE TWSA and TWSA from hydrologic models at Ganges basin is due to the human intervention attributed to significant ground water (GW) abstraction as this basin includes heavily irrigated land for agricultural purposes. Moreover, the signal leakage from the High Mountain Asia glaciers introduces additional systematic errors in the GRACE TWSA which also propagates into the error budget of the DL simulations. Similarly, Brahmaputra is also affected by human interventions through irrigation (Long et al., 2016), which is likely to be not adequately accounted for in the assimilative hydrologic models but captured by GRACE/-FO. It is also worth noting that all four basins with large errors have strong TWSA signals compared to other regions which cover the majority of Earth's terrestrial hydrologic regions. The higher uncertainty of the input hydrologic models at these regions results in higher simulation errors, because all the (DL or other) algorithms aim to achieve mathematically best fit to the CSRM TWSA globally. In order to achieve better TWSA simulations over these basins, either region-specific (i.e., only using the input-output data for that region) models should be studied or regionally improved hydrologic models should be used as inputs. For example, a DL model constructed to simulate the GRACE-like TWSA in Amazon basin alone, may improve the simulation performance. An alternative approach for improved simulations could be making use of location dependent prior uncertainty information of the input hydrologic models, however a thorough analysis of the models may be needed in order to estimate realistic uncertainty of the input data throughout the Earth. Nevertheless, the DL simulations match quite well with the observed CSRM, although the BCNN performed slightly worse to capture the signal amplitude, particularly at basins with strong annual TWS change signal such as Amazon, Zambezi, and partially at Mississippi. Among the three DL models, DCAE seems to provide the best simulations at almost all ten selected basins. Finally, in terms of temporal patterns, the simulated TWSAs are consistent with both hydrologic models, ERA5-L and NOAH.

3.2. Comparison with previous studies

In order to compare our results, we chose two recently published similar studies, namely Li et al. (2021) and Mo et al. (2021), who provided publicly available monthly gridded TWSA data products (also excluding Antarctica). These two studies are known as the best available reconstruction models so far, dedicated to fill the gap between mascon type GRACE/-FO TWSA time series at grid cell scale. Similar to our study, Li et al. (2021) used CSRM, but performed the TWSA reconstruction at a spatial resolution of 0.5° × 0.5°. Therefore, before the comparison, their TWSA products are resampled to 1° × 1° grids to be consistent with our resolution. It is worth noting that Li et al. (2021) used the entire GRACE CSRM data (April 2002 to June 2017) for training while they used the GRACE-FO CSRM data (from June 2018 to June 2020) for testing their model. Furthermore, they computed the TWSA trend from the whole available GRACE CSRM data, applied their method using de-trended series and restored the CSRM TWSA trend back to their original reconstructions of the seasonal TWSA signals.

Comparing the entire (training and testing) data span of Li et al. (2021) and our 13 test months shown in Fig. 1, there are 12 overlapping months. Despite the fact that the first seven of these 12 months are included within the training data set of Li et al. (2021), we calculated both RMSE and NSE values using these 12 months of CSRM TWSA grids and the corresponding simulated TWSAs from this study and from Li et al. (2021). Here, we restricted comparison with our DCAE model simulations as it shows the best performance (see Section 3.1) among the three DL models we experimented. The resulting RMSE and NSE are shown in Fig. 5 (a–d). While both solutions show comparable results in general, the significant improvements (i.e. lower RMSE and higher NSE) by DCAE are obvious at glaciers (e.g. eastern Greenland, and western Alaska), Zambezi basin in Africa, northern Asia, and eastern Australia. On the other hand, Li et al. (2021) seems to be slightly better in north Africa covered by the hyper-arid Sahara Desert if evaluated with NSE values. This can be explained with the weak TWSA signal and most likely the Li et al.'s (2021) assumption of the constant long-term trend holds in this region, which means that the trend of TWSA does not change significantly in time. In hyper-arid regions such as north Africa, the seasonal TWSA signal is very weak, yielding a very low signal-to-noise ratio in GRACE TWSA products in such regions. However, the borders of the area with low NSE pattern (dark blue area in Fig. 5c) in north Africa from DCAE coincides perfectly with the borders of Sahara Desert and this might actually imply that our DCAE model adequately filtered out the noise in the CSRM TWSA solutions in the region. This is an interesting result and deserves to be investigated in a future study.

The lower NSE of Li et al. (2021) in eastern Australia is due to the widespread flooding in the region occurred in February 2020. The extreme rainfall caused an exceptional increase in the TWS in February and lasted for the following few months (https://earthobservatory.nasa.gov/images/146284/extreme-rain-douses-fires-causes-floods-in-australia) which violates the assumption of constant TWSA trend by Li et al. (2021), thus resulting in high reconstruction errors e.g. in April 2020 which is one of the 12 test months used to calculate the RMSE and NSE values in Fig. 5. We also show the trend maps calculated from 55 (43 training and 12 test) months

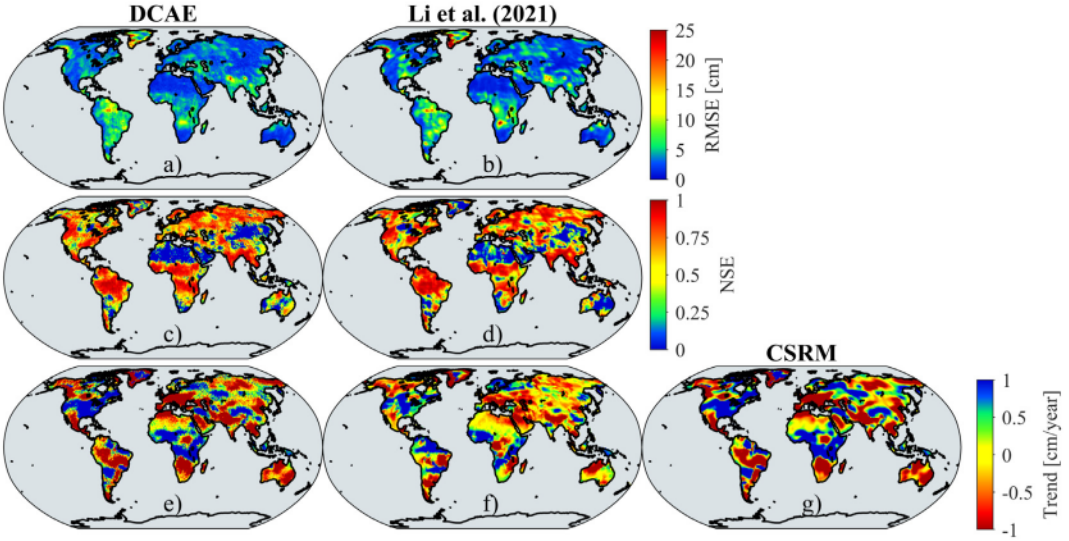


Fig. 5. Spatial distribution of RMSE and NSE values calculated from 12 test months: (a, c) for DCAE and (b, d) for Li et al. (2021), respectively. The long-term linear trend maps estimated from 55 (including 43 training and 12 test) months of (e) TWSA simulations by DCAE, (f) by Li et al. (2021) and (g) by the corresponding CSRM TWSA solutions.

of data between January 2014 and June 2020 (i.e. the last month in Li et al., 2021) using the reconstructed monthly TWSA from DCAE and Li et al. (2021) in Fig. 5e and f, respectively. The *true* trend map calculated from the corresponding CSRM TWSA is also presented in Fig. 5g for reference. Although we did not apply any de-trending to either inputs or the output CSRM and retrieved the TWSA trend through the training process, DCAE seems to have successfully captured the TWSA trend; a perfect agreement is clearly seen with that of CSRM (see Fig. 5e and g). This success in the trend retrieval is not only due to the DCAE method but also is the joint contribution of coarse resolution Swarm data and in particular the normalized time which are included as input data variables in our DL architectures.

Our second comparison is made against the reconstruction products of Mo et al. (2021). Contrary to Li et al. (2021), Mo et al. (2021) used JPL (Jet Propulsion Laboratory) mascon (JPLM) TWSA solutions averaged to 1° × 1° grids as the target TWSA data. Since the notable differences between CSRM and JPLM solutions exist (Chen et al., 2019; Sun et al., 2020a), in order to make a fair comparison, we retrained our DCAE model using the JPLM TWSA as the output data instead of CSRM TWSA without changing the input data or the model architecture.

The RMSE and the NSE metrics for DCAE simulations and for those of Mo et al. (2021) were calculated using the same 12 test months that we used for comparison with Li et al. (2021) and shown in Fig. 6 (a–d), respectively. It is clearly seen that DCAE shows significant improvement particularly in regions which exhibit strong water mass variations such as west and southern Greenland, western Alaska and Amazon basin with lower RMSE and higher NSE values in these regions. However, Mo et al. (2021) seems to be performing arguably better in hyper-arid regions, especially in terms of NSE in northwest Africa and Arabian Peninsula. This is again because the assumption of the constant long-term TWSA trend is valid at these regions; Mo et al. (2021) estimated and removed the TWSA trend computed from the entire GRACE/-FO TWSA time series of JPLM (between April 2002 and August 2020) and applied a BCNN approach using the input hydroclimatic and output JPLM data after removing the corresponding trend signals. This means that Mo et al. (2021) actually reconstructed the seasonal component of the TWSA, and the original GRACE-like TWSA products were then obtained by adding the JPLM-estimated trend component back to their output from BCNN model. Therefore, it is reasonable to assume that Mo et al.'s (2021) reconstructed TWSA data products should be free of trend error. However, the similarity of the NSE patterns shown in Figs. 5c and 6c in north Africa and Arabian Peninsula is noteworthy. Although different mascon data were used in modeling, both DCAE simulations resulted in almost the same NSE values in these regions. This similarity cannot be attributed to DCAE model; the same is observed (see Fig. 3, right column) not only from the TWSA simulations of other two (BCNN and CNN) DL models but also those derived from the hydrologic models (ERA5-L and NOAH). Apparently, in hyper-arid regions, our DCAE model results are more consistent with hydrologic models in terms of TWSA signal amplitude while retaining the long-term trend accurately from GRACE TWSA. This is a reasonable result; as the weak seasonal TWSA signal in hyper-arid regions causes low signal-to-noise ratio in the recovered TWSA from GRACE; that is GRACE can capture adequately the long-term signal, but the seasonal component of the signal is largely or almost completely dominated by noise. This result in Sahara Desert is also supported by, e.g. Klees et al. (2008) and Eicker et al. (2020); both studies reveal that the root mean square of the monthly mean water mass change signal over Sahara Desert is below the accuracy expected from GRACE/-FO missions. Therefore, low (almost zero or negative) NSE of TWSA simulations over hyper-arid regions should not be interpreted as the failure of the methodology; indeed, it may reflect the efficiency of the DCAE model in filtering the inherent noise of JPLM TWSA products. We note that deep autoencoders have already been confirmed with their success in denoising image data (Vincent et al., 2008; Zhang et al., 2017).

Similar to the Fig. 5 (e–g), we calculated and plot the linear trend from the corresponding monthly TWSA time series of 57 (45 training and 12 test) months between January 2014 and August 2020 (i.e. the last month in Mo et al., 2021). The trend map from DCAE simulations and that from Mo et al. (2021) are shown in Fig. 6e and f, respectively. The *true* trend map computed from JPLM is also presented in Fig. 6g for reference. Although the (*true*) trend signal computed from JPLM was directly added to the de-trended TWSA products of Mo et al. (2021), the spatial pattern and the magnitude of the trend from DCAE simulations seem to be more coherent with the corresponding JPLM trend. Note that, unlike Mo et al. (2021), we did not de-trend either the input data variables (PPT, TEMP, CWSC, ERA5-L TWSA, Swarm TWSA and normalized time) or the output (mascon TWSA) data in our DL model, however the trend signal is successfully retrieved from the data through the training process. We believe that, including the non-stochastic normalized time as one of the input data has led the model resolve the complex temporal correlations between input and output data better and yielded an excellent trend retrieval. Since the computed linear trends depend on the length of the time series used, in particular under the impacts of global climate change which is frequently pronounced nowadays, we suggest not to de-trend the mascon or other types of GRACE TWSA when performing reconstruction studies.

Comparisons with both studies above show that our DCAE model produces comparable or even better GRACE-like TWSA simulations although

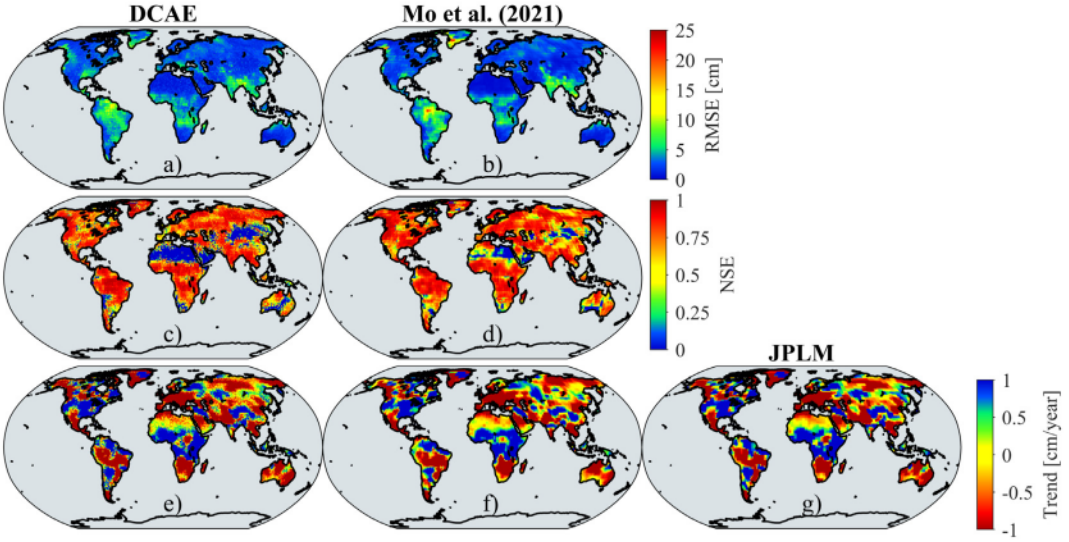


Fig. 6. Spatial distribution of RMSE and NSE values calculated from 12 test months: (a, c) for DCAE and (b, d) for Li et al. (2021), respectively. The long-term linear trend maps estimated from 57 (including 45 training and 12 test) months of (e) TWSA simulations by DCAE, (f) by Li et al. (2021) and (g) by the corresponding JPLM TWSA solutions.

we used less number of data within a shorter time span and without the need of preprocessing such as prior de-trending and assumptions such as constant linear trend during the entire time series. Nevertheless, it would give more sound opinion if we compare the simulations/reconstructions within the gap period where independent (non-GRACE) geophysical observations are available to evaluate the performances of the studies. Such a comparison is carried out and presented in Section 3.3.3.

3.3. External validation

The internal validation is commonly performed in almost all data-driven modeling studies (including DL methods) to ensure the generalization capability of the developed models. However, validations with independent observation data is of crucial importance, in particular for geophysical studies including simulations/reconstructions of geophysical parameters such as GRACE-like TWSA which aim to fill the data gap when no observation of the same type (e.g. CSRM TWSA) is available. In this section, the simulative performance of BCNN, DCAE and CNN algorithms are validated versus external measurements and/or models that are independent from GRACE observations/models. Thus, the simulated TWSA could be evaluated to see whether it could capture the spatio-temporal patterns of geophysical signals or not. For this purpose, three different types of independent data time series which are (i) in situ GWL measurements, (ii) the occurrences of extreme weather events, i.e., the floods and droughts attributed to abrupt changes of PPT and TEMP, and (iii) InSAR land subsidence rate are used.

3.3.1. Validation with in situ ground well observations

The in situ GWL measurements are downloaded from United States Geological Survey (USGS - https://groundwaterwatch.usgs.gov/) for chosen wells. Two different wells are selected considering that they include continuous daily GWL measurements with no gap within the study period and they are located at regions with different hydrologic characters. The daily GWL measurements at each well are first averaged to the monthly values, and the mean of the monthly GWL between January 2004 and December 2009 are removed from monthly averaged GWL in order to be consistent with simulated TWSA. The time series of simulated TWSA and TWSA from hydrologic models are also calculated by averaging the simulated TWSA at the nine neighboring ($1^\circ \times 1^\circ$) grid cells around the well locations, considering the native spatial resolution of GRACE/-FO. Since the TWSA and GWL are not exactly the observations of the same signal (Note that TWSA includes GW and the various surface water components which also have different temporal dynamics), the monthly time series between 2014 and 2020 of both GWL and region-averaged TWSAs are further detrended before the comparison. In order to allow better visual comparison, each time series is normalized and transformed to corresponding Z-scores using standard deviation and mean of each model/data.

The normalized time series of GWL and simulated TWSAs as well as those of available CSRM TWSA as described above are given for two chosen wells along with their location information in Fig. 7. However, direct visual comparison of simulated TWSAs with CSRM should be avoided since each time series are de-trended using their own estimated trend components and therefore the errors of the methods' capturing the trend signal in TWSA do not appear in Fig. 7. With a simple visual inspection, we can see that the temporal patterns of the simulated TWSA signals have good agreement with in situ GWL measurements, in particular at the well in Michigan while a significant difference between the trends of TWSA and GWL in Sacramento is still observed which is most likely due to the large year-to-year changes in the surface and soil water storage within the study period. This is also valid for the TWSA computed from both hydrologic models. In order to quantify this agreement, the cross correlations between TWSA and GWL signals are calculated separately both for the whole time span and for only 11-months data gap between GRACE/-FO. The computed correlations are listed in Table 1. The performance of DCAE seems to be the best among the three DL models in particular during the data gap, as the DCAE

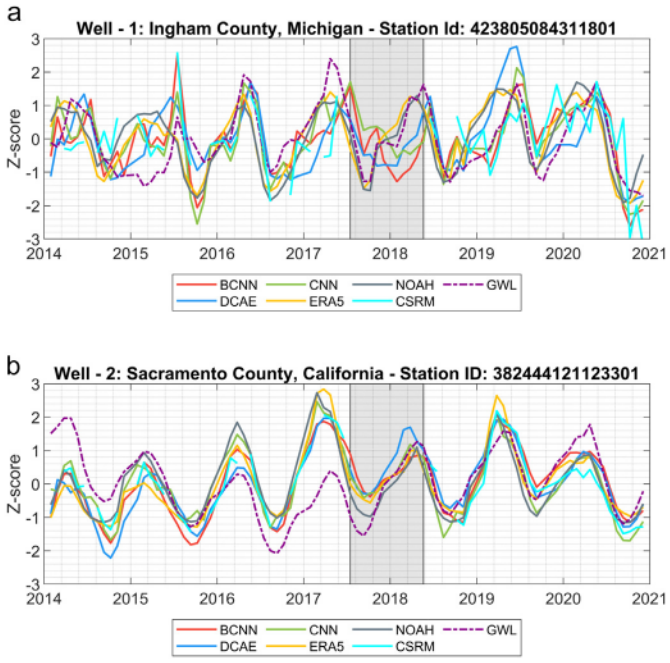


Fig. 7. Comparison of GWL measurements with the de-trended and normalized TWSA simulations, TWSAs from hydrologic models and from original CSRM solutions at in situ well stations (a) in Michigan and (b) in California. Gray shaded area represents the data gap between the GRACE/-FO.

simulated TWSA has the highest correlations. In addition, both ERA5-L and NOAH also have higher correlations with the GWL measurements as well as with the simulated TWSAs, in particular, with the DCAE derived TWSA, within the long GRACE/-FO data gap. The higher correlation of hydrologic model simulations with GWL is because they better spatially localize the surface water storage changes, such as those occurring in lakes and reservoirs than the GRACE/-FO estimates which suffer from leakage effects due to low spatial resolution. Surface storage changes can represent a large portion of the total water storage, such as those in Laurentian Great Lakes leading to leakage effects, and therefore resulting in surface water signals becoming erroneously assimilated into other water storage compartments of neighboring grids (Deggim et al., 2021). This also explains the short-term fluctuations of the CSRM TWSA time series in Michigan (see Fig. 7a), the surface water (Great Lakes reservoir) and the TWS signals are mixed up while they indeed have significant phase-lag in-between as well as different amplitudes. In order to separate and localize these two compartment signal in this region, Deggim et al. (2021) provided a correction data product for GRACE TWSA, generated from forward modeling of surface water volume estimates based on satellite altimetry and optical remote sensing, which can reach up to 30 cm (EWH) for individual months in the vicinity of Great Lakes. Nevertheless, DCAE seems to mitigate the effect of the Lake water on TWS to some extent; the short-term fluctuations are almost completely disappeared and the seasonal pattern can clearly be distinguished. Note that such a high-frequency fluctuation is not observed around the well in California (Fig. 7b); the CSRM time series clearly follow annual pattern as there is no large surface water bodies at the region.

Table 1
The cross correlations between TWSA signals of DL simulations and of hydrologic models and GWL measurements.

		BCNN	DCAE	CNN	ERA5-L	NOAH
2014–2020	Well - 1	0.57	0.65	0.65	0.65	0.62
	Well - 2	0.58	0.56	0.58	0.58	0.54
Gaps	Well - 1	0.02	0.70	−0.20	0.88	0.93
	Well - 2	0.69	0.91	0.91	0.92	0.95

3.3.2. Validation with extreme weather events: flood and drought examples

The extreme weather events can also be used to assess the simulative success of DL models, especially in 11-months gap. For this purpose, two different flash weather events were selected. With the expression *flash weather events*, we mean the events which differ in terms of occurrence time from the similar events regularly occurred at certain times with certain periods at the region in the past. Since almost all the data-driven methods learn the patterns from the available data, it is a challenge to capture such an in-phase (i.e. delayed) signal within the gap period.

South Asian Floods is a good example for such an extreme weather event. The severe South Asian Floods were the extreme floods at India, Nepal and Bangladesh in 2017, since floods hit the Ganges, Brahmaputra and Meghna river basins unusually in terms of both the precipitation levels and the time-span of monsoon seasons, i.e. while the normal monsoon season is around June–September, it occurred between April and October in 2017, contrary to the former events (Akanda et al., 2017; Palash et al., 2020). Thus, the South Asian Floods consist of different devastating flood events throughout monsoon period in 2017, but the severe flood occurred in early August in the Northern Bangladesh, Northern India and Eastern Nepal (Akanda et al., 2017; Philip et al., 2019; Palash et al., 2020).

The simulative performance of DL algorithms is compared to precipitation measurements, which is the total precipitation dataset of ERA5-L and used as input in the training process, to see whether the TWSA signals include the similar spatio-temporal behavior with the heavy rainfall on chosen flood region. For this purpose, nine tile grids within the red squares in Fig. 8a are chosen considering severely affected regions and shown along with the month-to-month differences of TWSA, PPT and TEMP during the monsoon period in 2017 to see their monthly evolution. The devastating effect of the floods is clearly seen in the region, especially from June to July and from July to August, from all TWSA simulations as well as those from hydrologic models. Fig. 8b shows the monthly normalized time series of simulated TWSA, TWSA from hydrologic models and the monthly mean PPT, all de-trended, for the whole study period (January 2014 and December 2020), averaged over the region within the nine tile grids. It is worth noting that the trend component is estimated from the 61 months of relevant data in each time series which coincide with the epochs of available CSRM within the study period; that is the simulations/observations at neither the intermission nor intermittent gap epochs have been used for trend reduction. As seen from Fig. 8b, TWSAs from ERA5-L and NOAH hydrologic models as well as simulated TWSAs from DL models except for CNN reach their peaks exactly in August 2017, i.e. at the reported peak time of severe flood within the gap. In contrast, CNN shows a peak signal with a one month of delay in September 2017.

The increasing and decreasing patterns of all TWSA series show high correlations with each other as well as with the PPT, though with an expected phase lag of 1 month which is also observed between PPT and CSRM TWSA. While the relatively larger hatched gray shaded area covers the 11 months of gap between GRACE/-FO, the narrower blue shaded areas represent the regular monsoon periods and red shaded area (in 2017) shows the time span of actual monsoon event (in 2017) which is shifted and extended in time with respect to the usual (or expected) occurrence time and period. It is seen that PPT and simulated TWSA signals are in very good agreement and the monsoon periods including the abnormal period during the data gap are clearly captured by the DL simulations. This result is also confirmed by high correlation coefficients, despite the reasonable 1 months of phase lag, computed between TWSA and PPT signals which are listed in Table 2 both for whole time series and only for 11-month gap, respectively. The correlation values computed after shifting the PPT series as much as the phase lag (i.e. 1 month) ahead in time were also given. Note that the correlation values reach all above 0.80 for entire study period and above 0.95 within the data gap when the phase shifting (PS) is applied.

The performance of the simulations capturing the extreme drought events was also investigated. To this end, the flash drought in Northern Great Plain (NGP) during the gap period is selected as a case event. The flash drought in NGP started approximately from early June covering the most part of Northeastern Montana, and North and South Dakota regions in United States (Svoboda et al., 2002; Gerken et al., 2018). The drought reached to its peak around early September (Gerken et al., 2018) and the drought conditions of the most part of Montana was reported (Svoboda et al., 2002) as extreme (D3) and/or exceptional (D4). The details of month-to-month spatio-temporal evolutions of NGP drought can be found in the United States Drought Monitoring – USDM database (https://droughtmonitor.unl.edu/Maps/MapArchive.aspx, Svoboda et al., 2002).

The nine (1° × 1°) tile grids (between 47°–50° North latitudes and 108°–105° West longitudes) covering Glasgow, Montana which is one of the regions reported to be exceptionally affected by drought (Svoboda et al., 2002; Gerken et al., 2018) are chosen. Similar to the flood case, detrending and normalization process applied to the time series of simulated TWSA, TWSA from hydrologic models and TEMP from ERA5-L model data in this region for comparison. Resulting time series are shown in Fig. 9. Here, the yellow shaded area represents the extreme or exceptional drought time-span in the region, while the hatched gray shaded area shows again the gap period between GRACE/-FO missions. The all TWSA signals have negative tendency within the yellow shaded time span, which means that the simulated TWSAs include the drought signal. However, the minimum simulated TWSAs in the yellow shaded area only from DCAE DL model and ERA5-L are observed in September, i.e. at the reported peak time of the drought in the region. In contrast, NOAH TWSA reaches its negative peak in August, while both BCNN and CNN exhibits the minimum TWSA simulations later in October. In addition, CNN seems to fail in simulations within the rest of the gap period (beyond September 2017). Moreover, the simulated TWSA signals are generally in good agreement with inverted time series of TEMP. The cross correlations computed between the TWSA and TEMP time series are given in Table 2 both for the entire study period and for only the gap period, respectively. All TWSA simulations as well as those from hydrologic models have negative correlations with varying values as expected, except for the CNN which incorrectly resulted in positive correlation during the gap period. This suggests that CNN could not produce reliable simulations, i.e. generalization of CNN model was actually not achieved despite the internal validation results tell otherwise. Therefore, validations with only using available input-output data (i.e. internal validation) alone can be misleading and must be supported by external validations with independent observations/data. The higher correlations of TWSA from hydrologic models with respect to BCNN and DCAE simulations with TEMP are due to their shorter phase lags from TEMP than those of DL results. The shorter phase lag of the hydrologic models is understandable as the surface water mass change and meteorological information (including soil moisture, temperature, precipitation, lakes, river water level data etc.) are directly assimilated in the models probably without considering the latency of contributions of those components into the TWS budget. The correlation values were recalculated after shifting the TEMP series ahead in time with an amount of the phase lag (i.e. 2 months) and the results are listed in Table 2. Here, we can see that the DCAE outperforms the other two DL methods and the hydrologic models both for the entire study period as well as for the gap period. Note that the respective values are −0.83 and −0.98 which are the highest negative correlations with TEMP among others.

3.3.3. Validation with InSAR subsidence rates

Annual subsidence rates in raster map format in Central Valley, California, computed from Sentinel-1 InSAR data are available and downloaded from https://data.cnra.ca.gov/. Each raster map covers a moving one-year period, produced at monthly intervals starting from 01 Jan 2015 to

Fig. 8. (a) The successive month-to-month differences of TWSA for DCAE, BCNN, CNN simulations as well as for ERA5 and NOAH, and of PPT and TEMP between April–October 2017. The red squares represent the chosen sub-region to compute, (b) the de-trended and normalized time series of TWSA and PPT and (c) simulated/reconstructed TWSAs vs. GRACE CSRM.

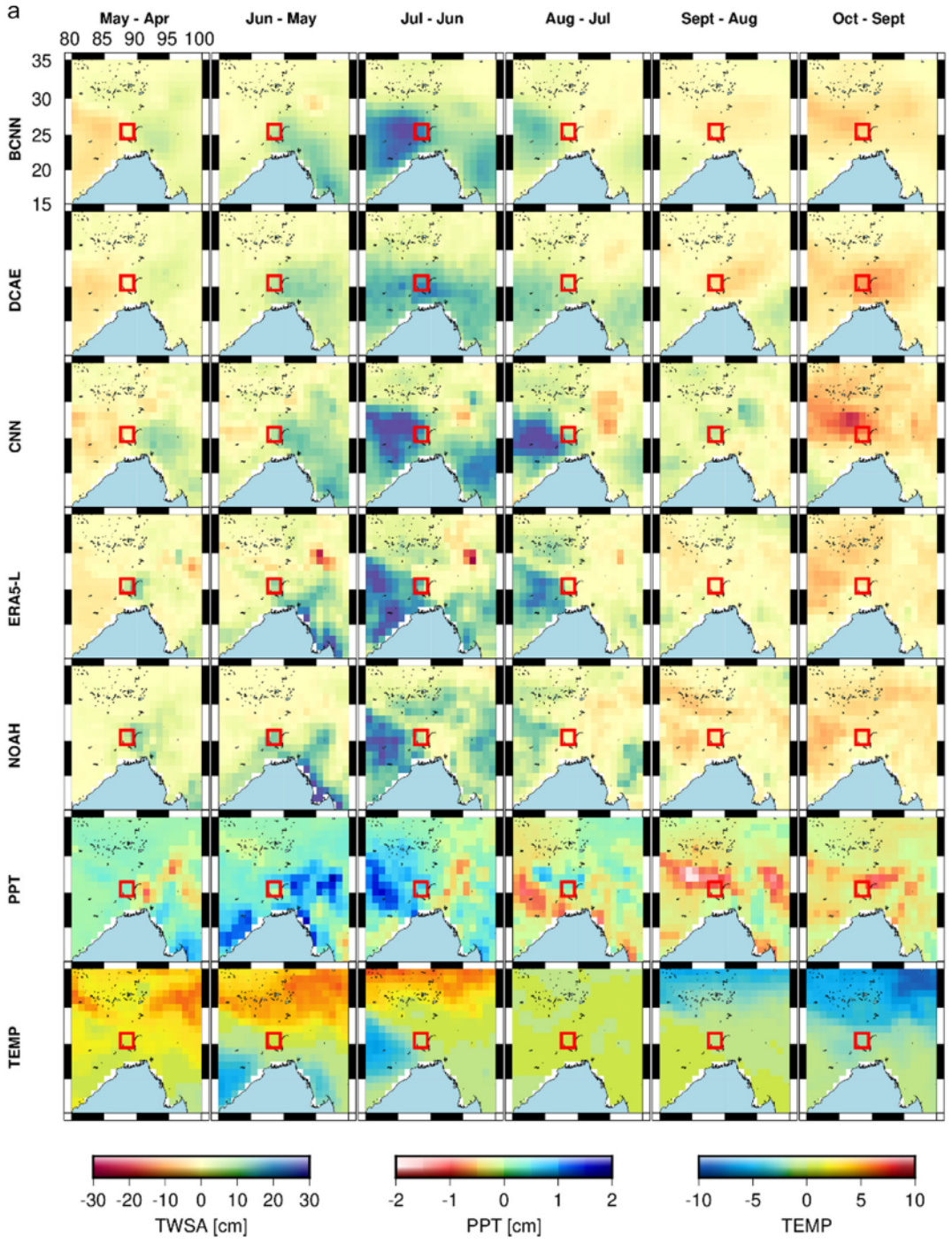

a
May - Apr
Jun - May
Jul - Jun
Aug - Jul
Sept - Aug
Oct - Sept
80 85 90 95 100
35 30 25 20 15
BCNN
DCAE
CNN
ERA5-L
NOAH
PPT
TEMP
-30 -20 -10 0 10 20 30
TWSA [cm]
-2 -1 0 1 2
PPT [cm]
-10 -5 0 5 10
TEMP
b
TWSA Z-score
PPT. Z-score
3 2 1 0 -1 -2 -3
2014 2015 2016 2017 2018 2019 2020 2021
BCNN
DCAE
CNN
ERA5
NOAH
CSRM
PPT

Table 2
The cross correlations both with (w/) and without (w/o) phase shifting (PS) between TWSA and PPT in South Asian Floods (top) and between TWSA and TEMP in Northern Great Plain Drought (bottom).

Event	Time	BCNN	DCAE	CNN	ERA5	NOAH
South Asian Floods	2014–2020 w/o PS	0.56	0.57	0.51	0.64	0.72
	Gap w/o PS	0.65	0.73	0.62	0.69	0.75
	2014–2020 w/ PS	0.87	0.84	0.88	0.85	0.81
	Gap w/ PS	0.95	0.96	0.97	0.97	0.97
Northern Great Plain Drought	2014–2020 w/o PS	−0.25	−0.39	−0.17	−0.47	−0.44
	Gap w/o PS	−0.22	−0.39	0.47	−0.45	−0.79
	2014–2020 w/ PS	−0.58	−0.83	−0.48	−0.65	−0.42
	Gap w/ PS	−0.97	−0.98	−0.50	−0.95	−0.91

01 October 2019. Thus, in each data file the subsidence rates computed using the successive 12 months of subsidence series from the SAR interferograms are given. As such, a time series of annual subsidence rates with monthly time steps at synthetic measurement points (SMP) with grid resolution of approximately 0.001° (~100 m) can be generated. The InSAR-observed subsidence was calibrated by vertical displacement from continuous GPS monitoring starting from mid-June 2015 (TRE Altamira Inc., 2021). Therefore, in our analysis, we used the InSAR-derived annual subsidence rate solutions starting from June 2015 till October 2019. The time series of these annual subsidence rates can be assumed as an approximation to the time derivative of the secular component of the subsidence time series observed at Central Valley between 2015 and 2020.

San Joaquin Valley and Tulare basins are highly productive agricultural regions in Central Valley, both receive minimal precipitation within entire Central Valley and account for at least 10% of the extracted GW in the US. GW depletion in these regions has resulted in the permanent loss of GW storage (GWS), manifested by declined water tables and land subsidence due to ongoing GW overdraft and aquitard compaction (Agarwal, 2020). The highest subsidence rates in Central Valley are observed in southern part of San Joaquin Valley and in particular at Tulare basin (covering the area between 35.5°–36.5° North latitudes and 119°–120° West longitudes), and thus chosen in our comparisons.

Liu et al. (2019) has shown that there is high correlation (as much as 0.72 in average), between the secular signals (but not between the seasonal components) of land subsidence and of GWS anomaly (GWSA) in San Joaquin Valley and Tulare basins, though there is not a strictly-linear relationship between the two, which suggests that a good portion of GRACE-observed GWSA change must reflect the GW loss due to inelastic compaction in the aquitard layers. Therefore, one should also expect a similarity between the patterns of the annual rates of subsidence and those of the GWSA to some extent.

In order to compare with the mean annual subsidence rates in San Joaquin Valley and Tulare basins, we estimated the mean annual rates (trends) with monthly time steps of GWSA of the two neighboring (1° × 1°) grid tiles aligned in the North-South direction covering the subsidence area. It is worth noting that the mean InSAR subsidence rates were calculated by averaging the original subsidence rates within these two grids. The GLDAS NOAH soil moisture anomaly (SMA), canopy water and snow water components (from ERA5-L) were first removed from the simulated TWSA series to obtain the corresponding GWSA series. Then annual GWSA rates were estimated. The reservoir water storage was not taken into account as it shows very little annual trends within the studied time span and thus ignored (Ojha et al., 2019). The GWSA rates computed as above were further filtered by 6-month moving average to smooth out the short-term fluctuations which are most likely due to seasonal variations of surface water.

Fig. 10 shows the time series of computed annual rates of both GWSA (from the three DL simulations) and the InSAR-estimated subsidence in the region. We also included the corresponding GWSA rates computed from reconstructed products of Li et al. (2021), Mo et al. (2021), Löcher and Kusche (2021) and those ($DCAE_{JPLM}$) estimated from our DCAE simulations using JPLM instead of CSRM (as described in Section 3.2) for

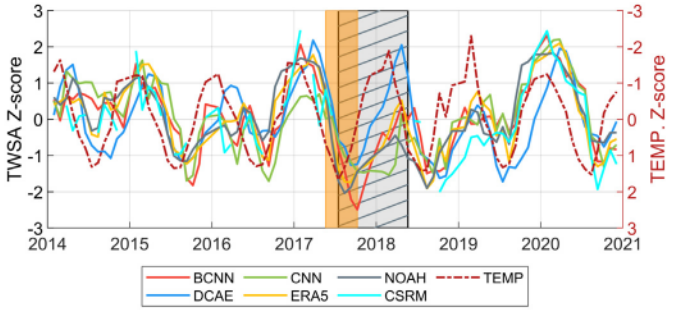


Fig. 9. Time series of TWSA and TEMP in North Great Plain, Montana, USA. Each series are normalized with its own mean and standard deviation. Note that the right axis for TEMP is inverted to improve the readability.

comparison after removing exactly the same SMA, canopy water and snow water components as applied to the CSRM-based DL simulations. The long-term TWSA trends computed from available CSRM, JPLM and postprocessed SH data products of Löcher and Kusche (2021) do not differ significantly from one another within the study area (i.e. Central Valley); the differences between the TWSA trends <0.4 cm/yr and thus are negligible for our comparison. Due to the 6-month moving average applied, we excluded the first and the last six months of the time series in Fig. 10, to avoid misinterpretation. Furthermore, note that each annual rate value is plotted versus the time epoch at the middle of the corresponding time span. For example, the rate value computed between January 2016 and December 2016 is plotted at the time epoch of June 2016.

Despite the scale differences, we can see a very good temporal correlation between the simulated GWSA rates and the subsidence rates. However, the GWSA rates computed from the BCNN and CNN simulations seem to be overestimated, which is due to their worse simulative performance at the region when compared to CSRM. BCNN additionally shows a phase lag of about 1 year between the GWSA and subsidence rates which is not realistic and most probably due to its shortcoming of learning trend information from the CSRM TWSA series, at least within this region. In contrast, DCAE produced more consistent TWSA simulations both with CSRM and JPLM (i.e. has the least RMSE and the highest NSE) at the region (not shown here) during the time span of the comparison. In general, the rise in the GWSA rate seems to be responded by a deceleration of the subsidence and the decrease of the GWSA rate yields an acceleration of the subsidence at the region, as expected. Regardless of the values of the GWSA rates, CNN also seems to have a good match, however it has negative correlation (~ −0.48) with subsidence rates within the gap period (gray shaded area) which suggests that CNN has rather focused on mathematically fitting to the available CSRM TWSA than retrieving the complex dynamics of the overall physical process. On the other hand, GWSA rates ($DCAE_{CSRM}$ and $DCAE_{JPLM}$) from DCAE have high positive correlations with InSAR subsidence rate both during the entire time span and during the gap period; correlation values reach >0.70 and >0.60, respectively in the gap period. Considering the 2 months delayed response of vertical displacements to changes in GWS during our study period in the region, as shown in Liu et al. (2019), we shifted the GWSA rate series 2 months ahead in time and recalculated the respective correlations with subsidence rate series and obtained the values > 0.80 for both $DCAE_{CSRM}$ and $DCAE_{JPLM}$.

The pattern of the annual GWSA rates from Mo et al. (2021) seems to be consistent with those of corresponding subsidence rates until the beginning of the gap period (mid-2017), however this consistency completely vanishes in the gap period; a negative correlation of −0.41 with subsidence rates is observed. This is due to the constant long-term trend assumption of Mo et al. (2021), which apparently does not hold during the entire gap period in this region; the subsidence has been accelerated during the gap period as a consequence of unusual GW decline. A similar observation can be made for the GWSA rates from Li et al. (2021), a negative correlation of −0.46 with subsidence rates is calculated. Note that both Li et al. (2021)

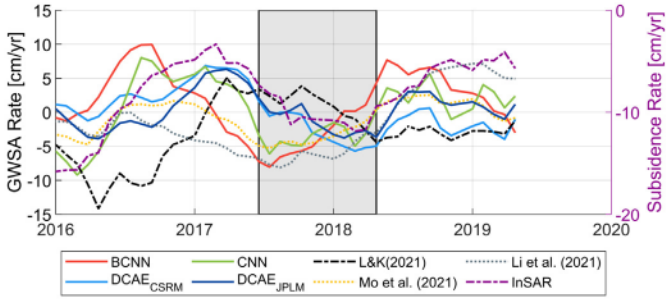


Fig. 10. The time series of annual GWSA rate vs Subsidence rate (from InSAR) for chosen region on Central Valley, California, USA. $DCAE_{CSRM}$: GWSA rate computed from CSRM based DCAE simulations; $DCAE_{JPLM}$: GWSA rate computed from JPLM based DCAE simulations; L&K (2021): Löcher and Kusche (2021).

and Mo et al. (2021) did not predict the trend signal; they rather computed the trend from available GRACE/-FO mascon TWSA data and assumed that this long-term linear trend would not change in time. When we apply 2 months shifting of the GWSA rates, these correlation values turn to be positive, but still as small as 0.31 and 0.19 for Mo et al. (2021) and Li et al. (2021), respectively. On the other hand, the correlation between the GWSA rates from SLR-GRACE combined solutions of Löcher and Kusche (2021) and the subsidence time series is slightly higher, at the level of 0.37, despite its lower spatial resolution (333 km half-wavelength) than mascons used in Mo et al. (2021) and Li et al. (2021), and its overall pattern is more consistent with those of our DCAE simulations. We believe that this is the contribution of the long-wavelength gravity signal from coarse resolution SLR solutions; SLR data adjusted the spatial patterns derived from available GRACE-only solutions to some extent through the combination process and the adjusted pattern is then projected to the monthly TWSA reconstructions in the gap period. Note that, although we used a different approach, we also utilized the long-wavelength gravity information as input, but from Swarm instead of SLR which eventually provided accurate simulations in the gap. Therefore, including coarse resolution gravity data which are available during the GRACE/-FO data gap such as Swarm (or SLR) yields physically more meaningful TWSA predictions. The overall comparison shows that DCAE provides more realistic simulations than the previous studies and has the highest generalization capability among the three DL methods investigated in this study.

4. Conclusions

In this study, we investigated the capability of three different DL neural network methods for filling the data gap between GRACE/-FO missions. We employed CNN, DCAE and BCNN DL models to simulate GRACE-like, but at a much higher resolution at 100 km than the natural GRACE/-FO spatial resolution (666 km), monthly gridded TWSA using four input hydro-/climatic model data sets (retrieved from ERA5-L), as well as long wavelength gravity data from the 3-satellite Swarm constellation solutions. In addition, we also included the normalized time information as the sixth input variable considering the fact that almost all geophysical processes/signals are functions of time.

Unlike previous studies, no prior de-trending or de-seasoning process to either input or output data was applied in order to avoid biasing/aliasing the simulations induced by interannual or longer term climate signals as well as extreme weather episodes. Therefore, we did not constrain our solutions with a pre-determined long-term trend or annual amplitude computed from the existing CSRM data as abrupt/unusual temporal changes within the data gap may deviate from long-term components of the signal. Instead, we allow DL models to learn the trend/amplitude and their temporal variation from the data during the training process.

We first tested the simulative performance of the proposed DL methods using global performance metrics (internal validation) such as RMSE and NSE which show the global goodness of fit of the simulations to the observed TWSA (i.e. CSRM which were not used in learning process) in least squares sense. Although the global metrics yield similar values with DCAE, BCNN seems to underestimate the trend and annual TWSA amplitudes, in particular in basins which have strong temporal TWS variations such as Amazon, Zambezi and partly in Mississippi, most likely due to using non-stochastic normalized time input which does not conform to the probability distributions and violates the basic principles of BCNN. We compared our simulation results to those from two hydrologic models, ERA5-L and GLDAS NOAH as well as from recently published similar studies. The global performances of the DL simulations are all found to be superior to the hydrologic models. Moreover, regardless of the method used, not only in this study, but also in previous studies, the highest RMSE values at grid cells appear at the same regions which are Amazon, Brahmaputra, Ganges and Zambezi basins, Western Alaska and coasts of Greenland, respectively. The same holds also for the hydrologic models which means that the uncertainty of the hydrologic models at these basins are larger than that at the other basins on the Earth. Therefore, since we use TWSA from a hydrologic model (ERA5-L) as an input variable, its relatively higher uncertainty also propagates into the simulated TWSA at the grid cell scale at these basins. However, the errors at grid cells dramatically reduce when the basin averages are calculated (see Fig. 4). In order to achieve better TWSA simulations at grid cell scale over these basins, either region-specific models should be studied or regionally improved hydrologic models should be used (as inputs). These are beyond the scope of this study and left for a future work. Furthermore, although we used less number of data within a shorter time span, our DCAE simulations outperform the CNN and BCNN and those of similar studies both in terms of internal (see Section 3.2) and external validations (see Section 3.3.3) which means that physically more meaningful TWSA predictions are obtained by DCAE model.

In addition to the global performance assessment, we also validated the simulation results with independent non-GRACE geophysical data sets (external validation) such as GWL observation records at in situ well stations, reported extreme weather events such as floods (South Asian Floods) and drought (Northern Great Plain Drought) attributed to heavy rainfall or extreme temperature, respectively and InSAR-observed land subsidence due to GW depletion in San Joaquin Valley and Tulare basins in Central Valley, occurred in particular within the gap period to check if the simulations can capture these signals. Among the three DL methods, DCAE is the only one which provided TWSA simulations consistent with all of the aforementioned independent data/observations. Therefore, using DL methods for filling large gaps in climate-related geophysical data such as TWSA is not trivial; the infilled data should be validated by independent observation data, as the internal validation based on global performance metrics alone may lead to the misjudgment of the models.

The overall validations and comparisons in this study show that: (i) DCAE is an effective DL approach for filling the gap between GRACE/-FO and (ii) one should avoid assumptions such as constant long-term linear trend computed from GRACE/-FO data in the pre-gap and post-gap periods, to obtain physically meaningful predictions of TWSA within the gap periods, (iii) the developed methodology and the model architecture has high potential to simulate/predict GRACE-like TWSAs in the pre-GRACE period, limited with the availability of long-wavelength gravity information; a possible candidate for such data could be monthly SLR gravity field solutions which are available since early 90s. Since our DCAE model is proven to efficiently simulate the trend information from the input data, the limitation due to unavailability of trend information for the pre-GRACE period as mentioned in Mo et al. (2021) can be successfully overcome. The performance of DCAE for such a task remains to be explored and left for a future study.

CRediT authorship contribution statement

Metehan Uz: Methodology, Software, Investigation, Data curation, Validation, Formal analysis, Visualization, Writing – original draft. **Kazım G. Atman**: Methodology, Software, Writing – review & editing. **Orhan**

Akyilmaz: Conceptualization, Supervision, Methodology, Writing – review & editing, Project administration. **C K Shum**: Conceptualization, Supervision, Methodology, Writing – review & editing. **Merve Keleş**: Methodology, Data curation, Investigation, Visualization. **Tuğçe Ay**: Methodology, Data curation, Investigation. **Bihter Tandoğdu**: Methodology, Data curation, Investigation. **Yu Zhang**: Investigation, Validation, Formal analysis. **Hüseyin Mercan**: Investigation, Writing – review & editing.

Declaration of competing interest

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

Acknowledgements

This work is partially supported by Scientific and Technological Research Council of Turkey - TÜBİTAK (119Y176) and also is a part of the first author's dissertation. We acknowledge partial supports from the United States National Science Foundation (NSF) Partnerships for Innovation Program (2044704), the United States Agency for International Development (USAID) project (72038621CA00002), and the NASA Earth Surface Interior Program (80NSSC20K0494). We thank Dr. Shaoxing Mo from Nanjing University for helpful discussions. CSR RL06 Mascon solutions are available in http://www2.csr.utexas.edu/grace. Swarm Level-2 data products are downloaded from International Center for Global Earth Models (ICGEM - http://icgem.gfz-potsdam.de/series/02_COST-G/Swarm). ERA5-Land (ERA5L) datasets are available on European Centre for Medium-Range Weather Forecast website (ECMWF - https://cds.climate.copernicus.eu). GLDAS Noah Land Surface Model is available in https://disc.gsfc.nasa.gov/datasets/. Annual subsidence rate data are obtained from https://data.cnra.ca.gov/. The developed codes and predicted TWSA data set are available from corresponding author upon reasonable request. The handling editor and four anonymous reviewers are gratefully acknowledged for their constructive comments which significantly improved the manuscript.

Appendix A. Supplementary data

Additional figures (Figs. A1, A2 and A3) are the supplementary material related to this article. Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2022.154701.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A System for Large-Scale Machine Learning. 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), pp. 265–283. https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Agarwal, V., 2020. Machine Learning Applications for Downscaling Groundwater Storage Changes Integrating Satellite Gravimetry and Other Observations. The Ohio State University, Columbus, OH, USA Ph.D. dissertation.

Akanda, A.S., Palash, W., Hasan, M.A., Nusrat, F., 2017. Understanding the unusual 2017 monsoon and floods in South Asia. AGU Fall Meeting Abstracts. 2017 NH51D-01.

Alain, G., Bengio, Y., 2014. What regularized auto-encoders learn from the data-generating distribution the. J. Mach. Learn. Res. 15 (1), 3563–3593.

Bandikova, T., McCullough, C., Kruizinga, G.L., Save, H., Christophe, B., 2019. GRACE accelerometer data transplant. Adv. Space Res. 64 (3), 623–644.

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.

Chollet, F., et al., 2015. Keras. GitHub. https://github.com/fchollet/keras.

Chen, W., Luo, J., Ray, J., Yu, N., Li, J.C., 2019. Multiple-data-based monthly geopotential model set LDCmgm90. Nat. Sci. Data 6, 228. https://doi.org/10.1038/s41597-019-0239-7.

Cheng, M., Tapley, B.D., Ries, J.C., 2013. Deceleration in the Earth's oblateness. J. Geophys. Res. Solid Earth 118, 740–747.

Cun, Y., Guyon, I., Jackel, L., Henderson, D., Boser, B., Howard, R., Denker, J., Hubbard, W., Graf, H., 1989. Handwritten digit recognition: applications of neural network chips and automatic learning. IEEE Commun. Mag. 27 (11), 41–46.

Deggim, S., Eicker, A., Schawohl, L., Gerdener, H., Schulze, K., Engels, O., Kusche, J., Saraswati, A.T., van Dam, T., Ellenbeck, L., Dettmering, D., Schwatke, C., Mayr, S., Klein, I., Longuevergne, L., 2021. RECOG RL01: correcting GRACE total water storage estimates for global lakes/reservoirs and earthquakes. Earth Syst. Sci. Data 13, 2227–2244. https://doi.org/10.5194/essd-13-2227-2021.

Ditmar, P., 2018. Conversion of time-varying stokes coefficients into mass anomalies at the Earth's surface considering the Earth's oblateness. J. of Geod. 92, 1401–1412. https://doi.org/10.1007/s00190-018-1128-0.

Eicker, A., Jensen, L., Wöhnke, V., Doslaw, H., Kvas, A., Mayer-Gürr, T., Dill, R., 2020. Daily GRACE satellite data evaluate short-term hydro-meteorological fluxes from global atmospheric reanalyses. Sci. Rep. 10, 4504. https://doi.org/10.1038/s41598-020-61166-0.

Ekwueme, B.N., Agunwamba, J.C., 2020. Modeling the influence of meteorological variables on runoff in a tropical watershed. Civ. Eng. J. 6 (12), 2344–2351. https://doi.org/10.28991/cej-2020-03091621.

Encarnação, J.T., Arnold, D., Bezděk, A., Dahle, C., Doombos, E., Van Den Ijssel, J., Jäggi, A., Mayer-Gürr, T., Sebera, J., Visser, P., et al., 2016. Gravity field models derived from swarm gps data. Earth Planets Space 68 (1), 1–15.

Encarnação, J., Visser, P., Jaeggi, A., Bezděk, A., Mayer-Gürr, T., Shum, C., Arnold, D., Doornbos, E., Elmer, M., Guo, J., van den IJssel, J., Iorfida, E., Klokočník, J., Krauss, S., Mao, X., Meyer, U., Sebera, J., Zhang, C., Zhang, Y., Dahle, C., 2019. Multi-approach Gravity Field Models From Swarm gps Data. GFZ Data Services Accessed: 2020-05-03.

Encarnação, J., Visser, P., Arnold, D., Bezděk, A., Doornbos, E., Ellmer, M., Guo, J., van den IJssel, J., Iorfida, E., Jäggi, A., Klokočník, J., Krauss, S., Mao, X., Mayer-Gürr, T., Meyer, U., Sebera, J., Shum, C.K., Zhang, C., Zhang, Y., Dahle, C., 2020. Description of the multi-approach gravity field models from Swarm GPS data. Earth Syst. Sci. Data 12 (2), 1385–1417.

Feng, W., 2019. Gramat: a comprehensive matlab toolbox for estimating global mass variations from grace satellite data. Earth Sci. Inform. 12 (3), 389–404.

Flechtner, F., Neumayer, K.-H., Dahle, C., Dobslaw, H., Fagiolini, E., Raimondo, J.-C., Güntner, A., 2016. What can be expected from the grace-fo laser ranging interferometer for earth science applications? Surv. Geophys. 37 (2), 453–470.

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. https://doi.org/10.3390/rs12101639.

Gerken, T., Bromley, G.T., Ruddell, B.L., Williams, S., Stoy, P.C., 2018. Convective suppression before and during the United States northern Great Plains flash drought of 2017. Hydrol. Earth Syst. Sci. 22 (8), 4155–4163.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press. http://www.deeplearningbook.org.

Guo, J.Y., Shang, K., Jekeli, C., Shum, C.K., 2015. On the energy integral formulation of gravitational potential differences from satellite-to- satellite tracking. Celest. Mech. Dyn. Astr 121 (4), 415–429.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Hinton, G.E., Zemel, R.S., 1994. Autoencoders, minimum description length, and helmholtz free energy. Adv. Neural Inf. Process 6, 3–10.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Humprey, V., Gudmundsson, L., 2019. GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. Earth Syst. Sci. Data 11, 1153–1170. https://doi.org/10.5194/essd-11-1153-2019.

Jäggi, A., Dahle, C., Arnold, D., Bock, H., Meyer, U., Beutler, G., van den IJssel, J., 2016. Swarm kinematic orbits and gravity fields from 18 months of GPS data. Adv. Space Res. 57, 218–233.

Kamyshanska, H., Memisevic, R., 2014. The potential energy of an autoencoder. IEEE Trans. Pattern Anal. Mach. Intell. 37 (6), 1261–1273.

Keleş, M., Ay, T., Tandoğdu, B., Uz, M., Zhang, Y., Akyilmaz, O., Shum, C.K., Atman, K.G., 2021. Bridging the gap between GRACE and GRACE-FO by simulating GRACE-like terrestrial water storage anomalies using deep machine learning tools. IAG Scientific Assembly, Beijing 28 June-02 July 2021.

Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization.

Klees, R., Liu, X., Wittwer, T., Gunter, B.C., Revtova, E.A., Tenzer, R., Ditmar, P., Winsemius, H.C., Savenije, H.H.G., 2008. A comparison of global and regional GRACE models for land hydrology. Surv. Geophys. 29, 335–359. https://doi.org/10.1007/s10712-008-9049-8.

Kornfeld, R.P., Arnold, B.W., Gross, M.A., Dahya, N.T., Klipstein, W.M., Gath, P.F., Bettadpur, S., 2019. Grace-fo: the gravity recovery and climate experiment follow-on mission. J. Spacecr. Rocket. 56 (3), 931–951.

Klokočník, J., Wagner, C.A., Kostelecký, J., Bezděk, A., 2015. Ground track density considerations on the resolvability of gravity field harmonics in a repeat orbit. Adv. Space Res. 56 (6), 1146–1160.

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.

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, e2021GL093492. https://doi.org/10.1029/2021GL093492.

Li, X., He, M., Li, H., Shen, H., 2022. A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection. IEEE Geosci. Remote 19, 1–5.

Liu, Z., Liu, P.-W., Massoud, E., Farr, T.G., Lundgren, P., Famiglietti, J.S., 2019. Monitoring groundwater change in California's central valley using sentinel-1 and grace observations. Geosciences 9, 436. https://doi.org/10.3390/geosciences9100436.

Long, D., Chen, X., Scanlon, B., Wada, Y., Hong, Y., Singh, V.P., Chen, Y., Wang, C., Han, Z., Yang, W., 2016. Have GRACE satellites overestimated groundwater depletion in the Northwest India Aquifer? SciRep. 6, 24398. https://doi.org/10.1038/srep24398.

Löcher, A., Kusche, J., 2021. A hybrid approach for recovering high-resolution temporal gravity fields from satellite laser ranging. J. Geod. 95 (6), 1–15. https://doi.org/10.1007/s00190-020-01460-x.

Lück, C., Kusche, J., Rietbroek, R., Löcher, A., 2018. Time-variable gravity fields and ocean mass change from 37 months of kinematic swarm orbits. Solid Earth 9 (2), 323–339.

Mo, S., Zhong, Y., Forootan, E., Mehrnegar, N., Yin, X., Feng, W., Shi, X., Wu, J., 2021. Bayesian convolutional neural networks for predicting the terrestrial water storage anomalies during GRACE and GRACE-FO gap. J. Hydrol. (ISSN: 0022-1694), 127244. https://doi.org/10.1016/j.jhydrol.2021.127244.

Mukherjee, I., Tallur, S., 2021. Light-weight CNN enabled edge-based framework for machine health diagnosis. IEEE Access 9, 84375–84386.

Muñoz Sabater, J., 2019. Era5-land monthly averaged data from 1981 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10 (3), 282–290.

Ojha, C., Werth, S., Shirzaei, M., 2019. Groundwater loss and aquifer system compaction in San Joaquin Valley during 2012–2015 drought. J. Geophys. Res. Solid Earth 124 (3), 3127–3143.

Oki, T., Sud, Y.C., 1998. Design of total runoff integrating pathways (TRIP)—a global river channel network. Earth Interact. 2 (1), 1–37.

Olsen, N., Friis-Christensen, E., Floberghagen, R., Alken, P., Beggan, C.D., Chulliat, A., Doornbos, E., Da Encarnação, J.T., Hamilton, B., Hulot, G., 2013. The swarm satellite constellation application and research facility (scarf) and swarm data products. Earth Planets Space 65 (11), 1189–1200.

Oo, H.T., Zin, W.W., Kyi, C.C.T., 2020. Analysis of streamflow response to changing climate conditions using SWAT model. Civ. Eng. J. 6 (2), 194–209. https://doi.org/10.28991/cej-2020-03091464.

Palash, W., Akanda, A.S., Islam, S., 2020. The record 2017 flood in South Asia: state of prediction and performance of a data-driven requisitely simple forecast model. J. Hydrol. 589, 125190.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic Differentiation in Pytorch.

Peltier, W.R., Argus, D.F., Drummond, R., 2018. Comment on "an assessment of the ICE-6G_C (VM5a) glacial isostatic adjustment model" by Purcell et al. J. Geophys. Res. Solid Earth 123, 2018–2019. https://doi.org/10.1002/2016JB013844.

Philip, S., Sparrow, S., Kew, S.F., van der Wiel, K., Wanders, N., Singh, R., Hassan, A., Mohammed, K., Javid, H., Haustein, K., Otto, F.E.L., Hirpa, F., Rimi, R.H., Islam, A.K.M.S., Wallom, D.C.H., van Oldenborgh, G.J., 2019. Attributing the 2017 Bangladesh floods from meteorological and hydrological perspectives. Hydrol. Earth Syst. Sci. 23 (3), 1409–1429. https://doi.org/10.5194/hess-23-1409-2019.

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 (1), 1–14.

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.

Save, H., Bettadpur, S., Tapley, B.D., 2016. High-resolution csr grace rl05 mascons. J. Geophys. Res. Solid Earth 121 (10), 7547–7569.

Save, H., Tapley, B., Bettadpur, S., 2018. GRACE RL06 reprocessing and results from CSR. EGU General Assembly Conference Abstracts, p. 10697.

Save, H., 2020. Csr Grace and Grace-fo rl06 Mascon Solutions v02 Accessed: 2020-05-01.

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. Natl. Acad. Sci. U. S. A. 115 (6), E1080–E1089.

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 e2019WR026250 2019WR026250.

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. https://doi.org/10.1029/2020WR028666.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., Rippey, B., Tinker, R., Palecki, M., Stooksbury, D., Miskus, D., Stephens, S., 2002. The drought monitor. Bull. Am. Meteorol. Soc. 83 (8), 1181–1190.

Swenson, S., Chambers, D., Wahr, J., 2008. Estimating geocenter variations from a combination of GRACE and ocean model output. J. Geophys. Res. 113, B08410. https://doi.org/10.1029/2007JB005338.

Tapley, B.D., Bettadpur, S., Watkins, M., Reigber, C., 2004. The gravity recovery and climate experiment: mission overview and early results. Geophys. Res. Lett. 31 (9). https://doi.org/10.1029/2004GL019920.

Tapley, B., Watkins, M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., Sasgen, I., Famiglietti, J., Landerer, F., Chambers, D., Reager, J., Gardner, A., Save, H., Ivins, E., Swenson, S., Boening, C., Dahle, C., Wiese, D., Dobslaw, H., Tamisiea, M., Velicogna, I., 2019. Contributions of GRACE to understanding climate change. Nat. Clim. 2019. https://doi.org/10.1038/s41558-019-0436-2.

TRE Altamira Inc., 2021. March 2021InSAR Land Surveying and Mapping Services to DWR Supporting SGMA - 2020 Update, Technical Report, p. 22. https://data.cnra.ca.gov/.

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.-A., 2008. Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th Int. Conf. on Machine learning – ICML'08. ACM Press, New York, New York, USA, pp. 1096–1103 https://doi.org/10.1145/1390156.1390294.

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. 594, 125972. https://doi.org/10.1016/j.jhydrol.2021.125972.

Wu, P., Yin, Z., Yang, H., Wu, Y., Ma, X., 2019. Reconstructing geostationary satellite land surface temperature imagery based on a multiscale feature connected convolutional neural network. Remote Sens. 11 (3), 300.

Yi, S., Sneeuw, N., 2021. Filling the data gaps within grace missions using singular spectrum analysis. J. Geophys. Res: Solid Earth 126 (5), e2020JB021227 e2020JB021227 2020JB021227.

Yu, Q., Wang, S., He, H., Ma, L., 2021. Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model. Int. J. Appl. Earth Obs. Geoinf. 102, 102404.

Zhang, K., Zuo, W., Chen, Y., Meng, D., Zhang, L., 2017. Beyond a gaussian denoiser: residual learning of deep CNN for image denoising. IEEE Trans. Image Process. 26 (7), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206.

Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56 (8), 4274–4288.

Zhang, C., Shum, C.K., Bezděk, A., Bevis, M., Encarnação, J., Tapley, B., Zhang, Y., Su, X., Shen, Q., 2021. Rapid mass loss in West Antarctica revealed by swarm gravimetry in the absence of GRACE. Geophys. Res. Lett. https://doi.org/10.1029/2021GL095141.

Zhou, Y., Chellappa, R., 1988. Computation of optical flow using a neural network. IEEE 1988 International Conference on Neural Networks. 2, pp. 71–78.

Zhu, Y., Zabaras, N., 2018. Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. J. Comput. Phys. 366, 415–447.

Appendix A. Supplementary data

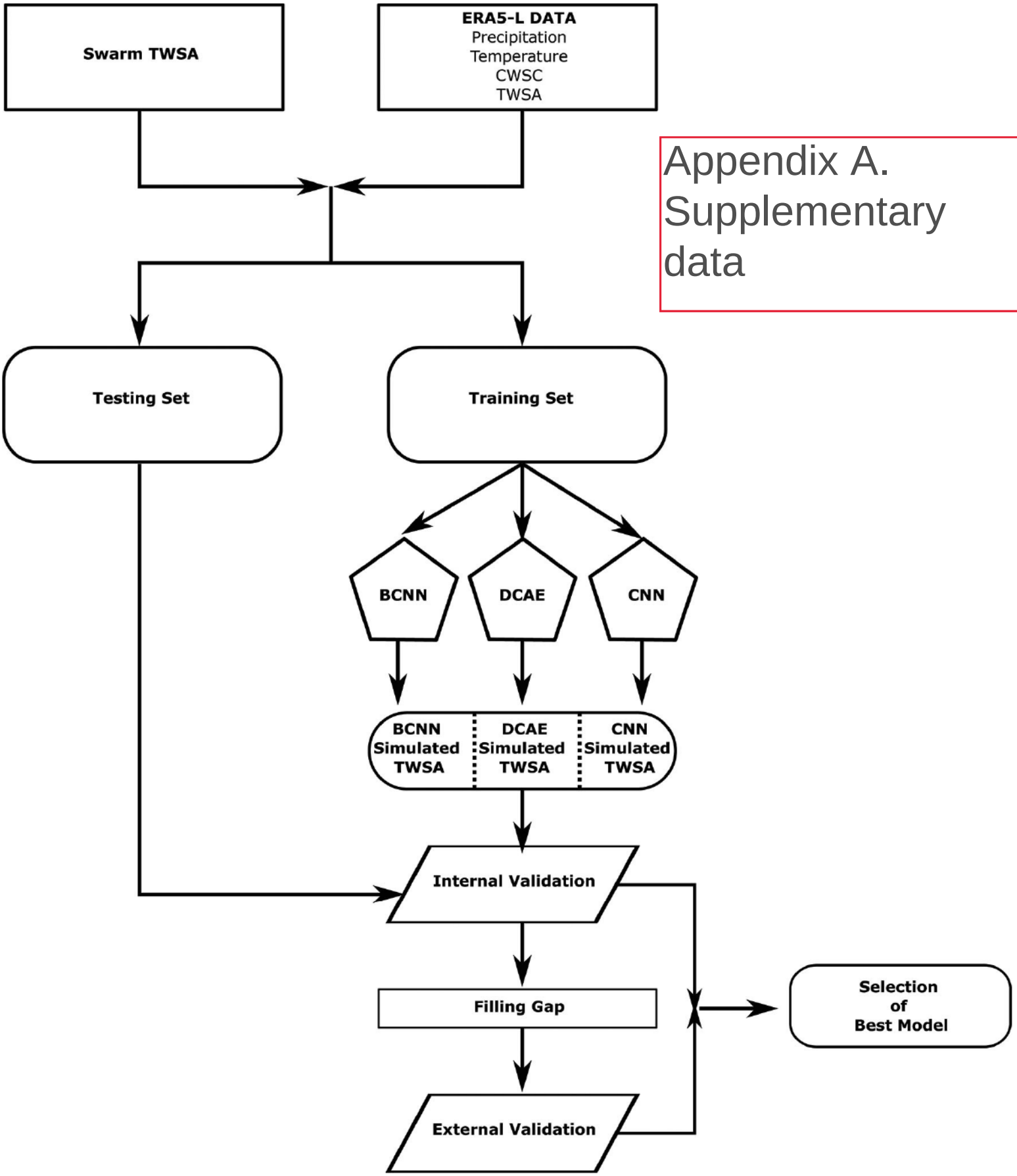


Fig. A1. The flowchart of the overall research methodology adopted in the study.

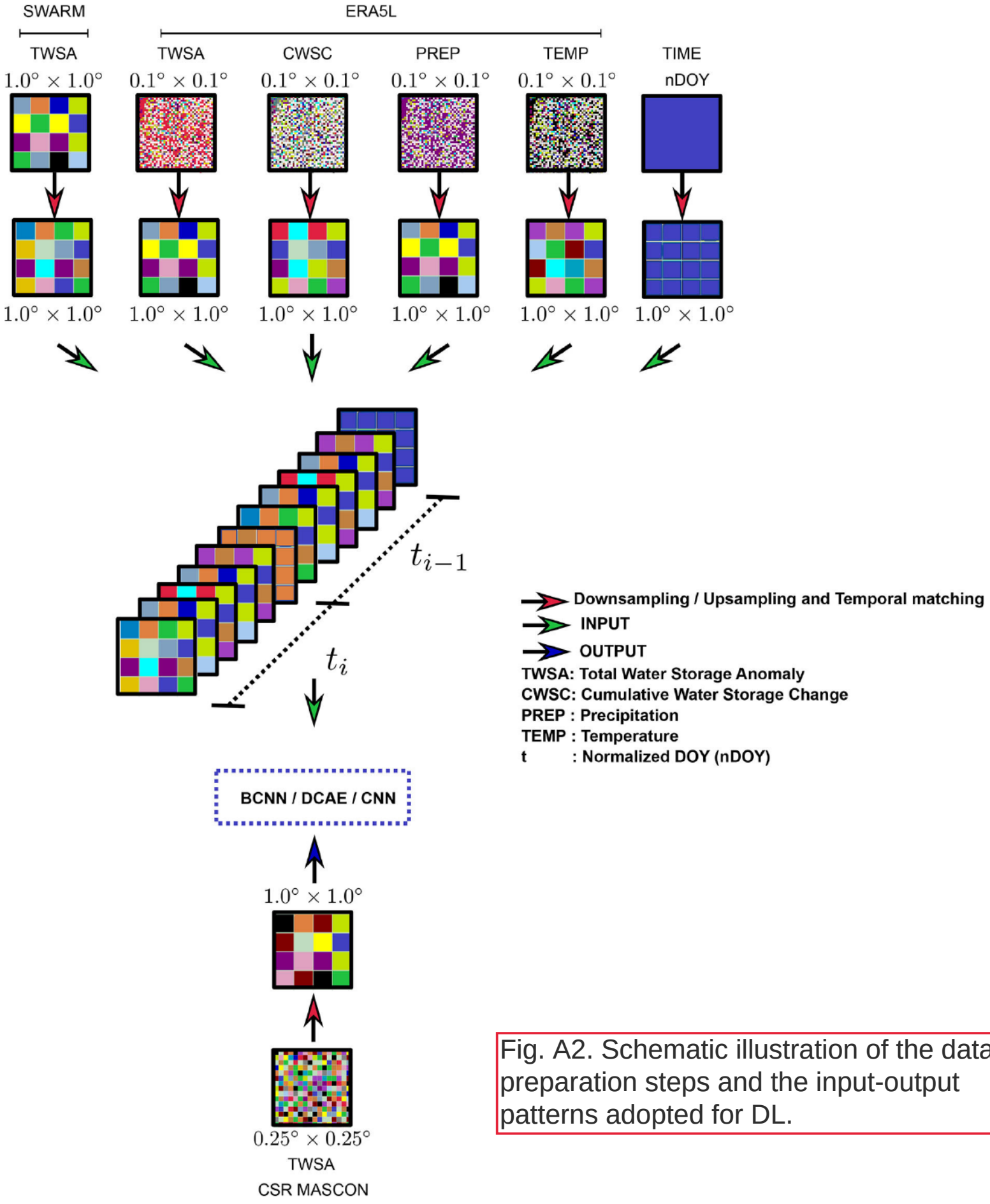


Fig. A2. Schematic illustration of the data preparation steps and the input-output patterns adopted for DL.

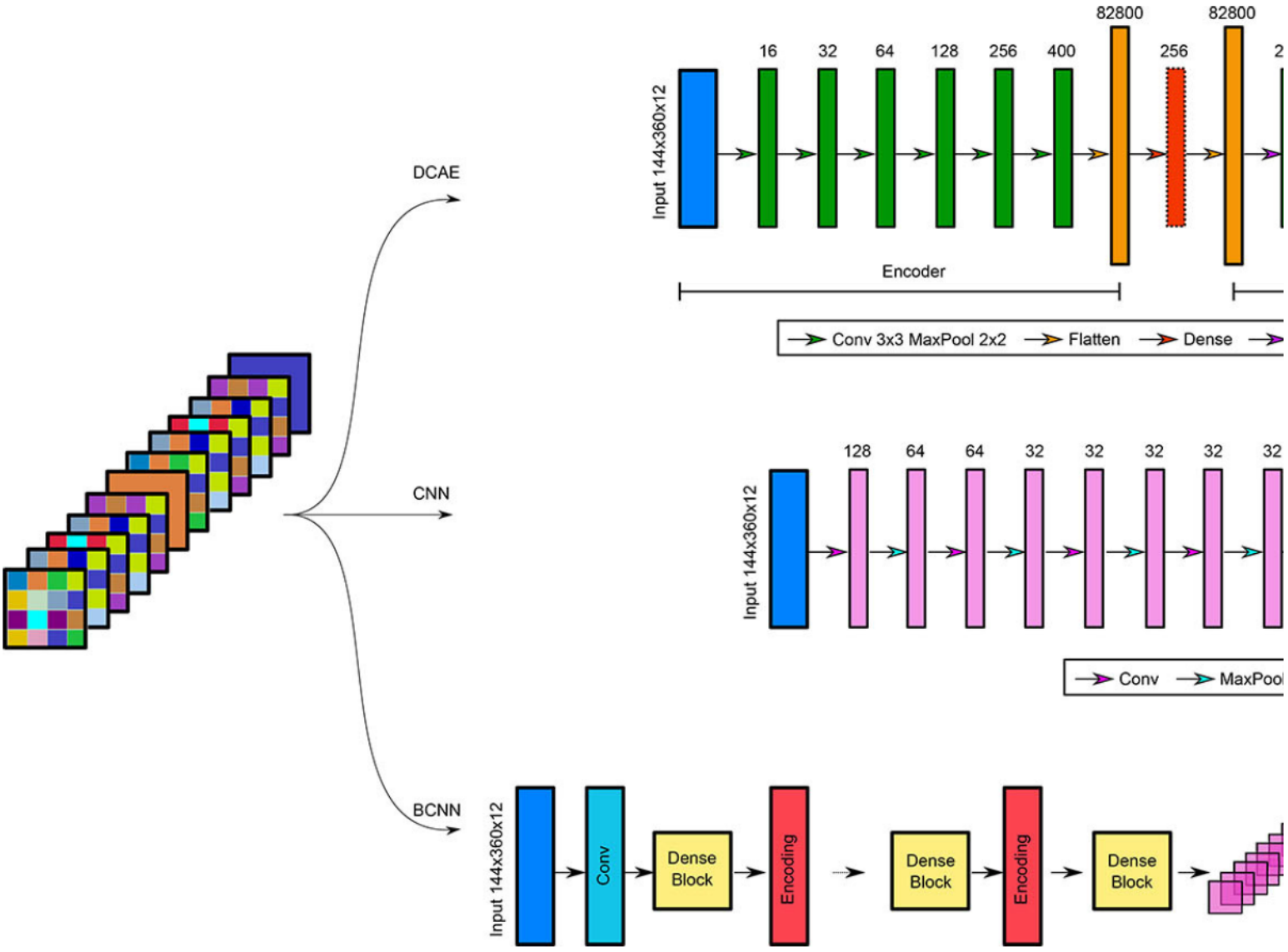


Fig. A3. Details of the network architectures of the three DL models adopted in this study.