Self-Validating Deep Learning for Recovering Terrestrial Water Storage From Gravity and Altimetry Measurements

Christopher Irrgang[1] [ID], Jan Saynisch-Wagner[1] [ID], Robert Dill[1] [ID], Eva Boergens[1] [ID], and Maik Thomas[1,2]

[1]Section 1.3: Earth System Modelling, Helmholtz Centre Potsdam – GFZ German Research Centre for Geosciences, Potsdam, Germany, [2]Institute of Meteorology, Freie Universität Berlin, Berlin, Germany

Abstract Quantifying and monitoring terrestrial water storage (TWS) is an essential task for understanding the Earth's hydrosphere cycle, its susceptibility to climate change, and concurrent impacts for ecosystems, agriculture, and water management. Changes in TWS manifest as anomalies in the Earth's gravity field, which are routinely observed from space. However, the complex underlying distribution of water masses in rivers, lakes, or groundwater basins remains elusive. We combine machine learning, numerical modeling, and satellite altimetry to build a downscaling neural network that recovers simulated TWS from synthetic space-borne gravity observations. A novel constrained training is introduced, allowing the neural network to validate its training progress with independent satellite altimetry records. We show that the neural network can accurately derive the TWS in 2019 after being trained over the years 2003 to 2018. Further, we demonstrate that the constrained neural network can outperform the numerical model in validated regions.

Plain Language Summary Continuous monitoring of the distribution and movement of continental water masses is essential for understanding the Earth's global water cycle, its susceptibility to climate change, and for risk assessments of ecosystems, agriculture, and water management. Changes of continental water masses are encoded as coarse blob-like patterns in satellite observations of the Earth's gravity field. Focusing on the South American continent, we introduce a self-validating artificial neural network to recover detailed and accurate spatiotemporal information of continental water masses from such gravity field observations.

1. Introduction

Deep learning (LeCun et al., 2015) is becoming an increasingly important tool in Earth system research, as it provides versatile concepts to complement and renew traditional methods to observe, simulate, or predict geophysical processes on Earth (Lary et al., 2016; Reichstein et al., 2019). Artificial neural networks consists of hierarchically arranged processing nodes (artificial neurons Rosenblatt, 1958) and work as non-linear function estimators through (un)supervised training. As such, neural networks can be trained to perform various different tasks, for instance, process simulation and prediction (e.g., Ham et al., 2019; Weyn et al., 2019), data inversion (e.g., Irrgang et al., 2019), data analysis and downscaling (e.g., Barnes et al., 2019; Bolton & Zanna, 2019; Irrgang et al., 2020; Kadow et al., 2020), or data assimilation (e.g., Cintra & Velho, 2014; Wahle et al., 2015).

The application of deep learning for quantifying and analyzing terrestrial water storage (TWS) is a currently emerging research branch, not only due to recent methodological advances but also for diverse prospects of aiding environment and water management (Sun & Scanlon, 2019). Most deep learning applications aim toward processing GRACE (Gravity Recovery and Climate Experiment) satellite observations of the Earth's gravity field and its anomalies due to spatiotemporal changes in TWS. As GRACE and GRACE-Follow On provide invaluable global gravity observations since 2002 (e.g., Dahle et al., 2019; Rodell et al., 2018; Tapley et al., 2019; Velicogna et al., 2020), large data streams are available to feed and train neural networks. Nevertheless, terrestrial water storage anomalies (TWSAs) extracted from GRACE measurements are blurred; that is, the spatial resolution of satellite gravimetry is limited to approximately 300 km, due to the signal's upward continuation to satellite altitude and the necessary data filtering. Thus, the underlying distribution

of water masses in rivers, lakes, aquifers, and ground water basins remains elusive (e.g., Boergens et al., 2014; Humphrey et al., 2016; Schrama et al., 2007). Satellite altimetry, in contrast, has been used to observe rivers for several years (e.g., Berry & Benveniste, 2010; Santos da Silva et al., 2010; Schwatke et al., 2015). Recently, the size of the observable rivers further decreased due to new satellites measuring with SAR altimetry and due to further improved classification and retracking algorithms. This allows now to observe whole river systems including smaller tributaries besides the main stream (Boergens et al., 2017, 2019; Michailovsky et al., 2012; Tourian et al., 2017). For quantitative comparisons to GRACE TWS, surface water volume variations of lakes and river flood plains can be estimated by combining altimetric water level observations with water surface data (Frappart et al., 2006, 2014, 2015).

Downscaling approaches try to recover or predict TWS components, for example, ground water level and storage (Chen et al., 2019; Seyoum et al., 2019; Seo & Lee, 2019) or watersheds (Ahmed et al., 2019), from GRACE observations and auxiliary data products, including precipitation, land surface temperatures, and vegetation cover. In related work, machine learning is also used for the removal of correlated errors in GRACE data (Piretzidis et al., 2018) and for reconstructing missing TWS observations between GRACE and GRACE-Follow On (Sun, Long, et al., 2020). Additionally, efforts are made to combine imperfect numerical hydrology models with neural networks, aiming to mitigate mismatches between simulated TWSAs and GRACE observations (Sun, Scanlon, Zhang, et al., 2019). First attempts have been made to use machine learning together with altimetry data; for example, Kim et al. (2019) applied ensemble learning (ELQ) to predict river discharge from altimetric water level time series.

In this study, we propose a different approach with the aim to recover multiscale TWSAs, ranging from large-scale total water storage to locally resolved river structures, just from GRACE-like observations. To achieve this task, we combine the respective advantages of space-borne gravity and altimetry observations with numerical hydrology modeling and deep learning. We have set up a model-based environment, in which a convolutional neural network (CNN) is trained with simulated TWSAs of the South American continent and corresponding GRACE-like observations, which are also generated from the hydrology model (see model description in section 2.1). A novel training approach is presented that allows the CNN to dynamically adapt and validate its learning process based on independent altimetry observations of surface water storage in the Amazonas region (see descriptions for the used altimetry data in sections 2.2 and for the deep learning in section 2.3). By unifying deep learning, numerical modeling, and space-borne observing systems, we highlight that the self-validating neural network can recover multiscale TWSAs from GRACE-like observations and can outperform the numerical model used in the training process (see results and discussion in section 3). Conclusions and next steps are provided in section 4.

2. Materials and Methods

2.1. Water Storage From Hydrology Model and Synthetic GRACE Fields

Global fields of TWS are taken from routinely simulations with the Land Surface Discharge Model (LSDM) (Dill, 2008). LSDM has been developed from earlier work at the Max-Planck-Institute for Meteorology in Hamburg, Germany (Hagemann & Dümenil, 1998; Hagemann & Gates, 2003). LSDM is defined on a regular global $0.5° \times 0.5°$ grid. The model is forced with precipitation, evaporation, and 2 m temperature from the operational simulations of the European Centre for Medium-range Weather forecast (ECMWF). After separating precipitation into rain and snow according to the temperature, snow accumulation, snowmelt, rainmelt, refreezing, throughfall, and drainage are simulated. Evaporation taken from ECMWF atmospheric forcing is raised up over open surface water areas to the potential evaporation given by Thornthwaite's equation. Actual evaporation is also adjusted to the available surface water. Subsequently, residual surface water is stored as soil moisture, groundwater, or in rivers and lakes. The water storage is discharged by a bucket-type runoff scheme through the river network and finally into the oceans, whereby rivers are represented by a generalized network of consecutive 0.5° grid cells. Glaciated regions are treated as permanent snow-covered regions with a constant ice sheet below. Snow is accumulated and melted in a seasonal cycle with no generation of new ice. LSDM is operationally used for the calculation of effective angular momentum functions describing TWS induced Earth rotation variations (Dill & Dobslaw, 2019a; Dill et al., 2019b) and for simulating elastic surface deformations caused by TWS loads (Dill et al., 2018; Neelmeijer et al., 2018; Zhang et al., 2017).

The model-based TWS variations are regarded as reference water distribution from which a pseudo GRACE-like observation is created. From the daily LSDM results we calculated monthly means at the GRACE epochs, and we applied a Gaussian filter ($R = 600$ km) in the spherical harmonic domain ($d/o = 180$) to obtain our synthetic GRACE observation (see Figure 1a). In spatial and temporal resolution these filtered TWS fields are comparable to typically processed GRACE Level 3 products. The pair of GRACE satellites detects the changes in water on Earth by measurement of variations of Earth's gravity field. The GRACE Science Data System (SDS), a service shared between the Jet Propulsion Laboratory (JPL), the University of Texas Center for Space Research (UTCSR), and the German Research Centre for Geosciences (GFZ), processes and archives the GRACE Levels 0 to 2 products. Level 2 products, commonly available in spherical harmonic coefficients, include the static and time-variable (monthly) gravity field and ancillary data products from the Level 2 processing such as GFZ's Level 1B short-term atmosphere and ocean dealiasing product (AOD1B). With nontidal atmospheric and oceanic short-term variations removed, the so-called Level 2 GSM fields represent the gravity changes due to TWS variations on land, residual ocean circulation not modeled by AOD1B, and seasonal and long-term sea level variations. To relate changes in gravity to changes in the mass distribution in a thin surface layer and to obtain user friendly gridded TWS estimations from GRACE, so-called Level 3 products, the Level 2 products need further processing, for example, replacement of SH coefficients of degree 1 (C10, C11, and S11), corrections for glacial isostatic adjustment (Peltier, 2004) and earthquakes, and spatial filtering. There exist many different filter approaches to remove the degree-dependent correlated noise, which manifests itself in north-south stripes, that makes it difficult to analyze small-scale regional mass variations (Swenson & Wahr, 2006). Common methods to mitigate this problem is the application of a Gaussian isotropic filter or anisotropic filters, such as the DDK filter (Kusche et al., 2009). Due to this necessary spatial filtering and the reduced observational spatial resolution from the altitude of the spacecrafts ($\approx$500 km), TWS fields from GRACE Level 3 products have a much lower resolution than modeled TWS fields. Applying a Gaussian filter with $R = 600$ km results in similar TWS fields for both modeled TWS from LSDM and observed TWS from GRACE gravity.

2.2. Surface Water Level Observations by Satellite Altimetry

Satellite altimetry data along the Amazon River and its tributaries are used twofold in this study: (1) as an additional constraint to enhance the neural network training (see section 2.3) and (2) as an independent source to validate the neural network downscaling of the GRACE-like TWSA observations (Figure 1). The data were provided by DAHITI (Schwatke et al., 2015) and consist of Virtual Stations (VSs) from the missions Jason-2, Jason-3, ERS-2, Envisat, SARAL, Sentinel-3 A, and Sentinel-3 B. Overall, 1,433 VS are available in the Amazon Basin with water level recordings between 1995 and 2020. In the further processing, we only use water level changes; thus, the long-term mean water level is reduced.

In order to combine the altimetry observations with GRACE data, the water level variation needs to be transferred to water volume change (SWSA—surface water storage anomaly). First, we assume that all water level time series inside the $0.5° \times 0.5°$ LSDM grid cells observe the same variations and thus can be combined to one time series. We assume that the water level time series observe the same variation if they have a correlation larger than 0.8 or in case of no temporal overlap if their annual amplitude fit together within 25% discrepancy. Only in five out of 390 grid cells none of the time series fitted together, in which case the longest time series is used. Second, we derive the water surface change from the Global Inundation Extent from Multi-Satellites (GIEMS) data set (Prigent et al., 2007). GIEMS provides the flooded area of an equal area grid (approx. 770 km², $0.25° \times 0.25°$ at the equator) at monthly resolution for the years 1993–2007. The flooded area is derived from active and passive microwave (visible reflectance and near-infrared reflectance) remote sensing satellite data. Before further processing, the water surface change data are aggregated to the $0.5° \times 0.5°$ grid cells, too.

In the final step, the two data sets are combined to SWSAs. For each grid cell a linear functional relationship between water level and water surface change is established in the overlapping time span. We found a linear function sufficient to describe the relationship between water level and area in the Amazon basin considering the data uncertainties. This relationship is then used to extrapolate the surface area change beyond 2007. Water level (h) and surface area (A) are combined to the water volume change between the epochs $i - 1$ and i with the truncated pyramid formula (Singh et al., 2015):

$$\Delta V(i) = \frac{1}{3} \cdot (h(i) - h(i-1)) \cdot (A(i) + A(i-1) + \sqrt{A(i) \cdot A(i-1)}). \tag{1}$$
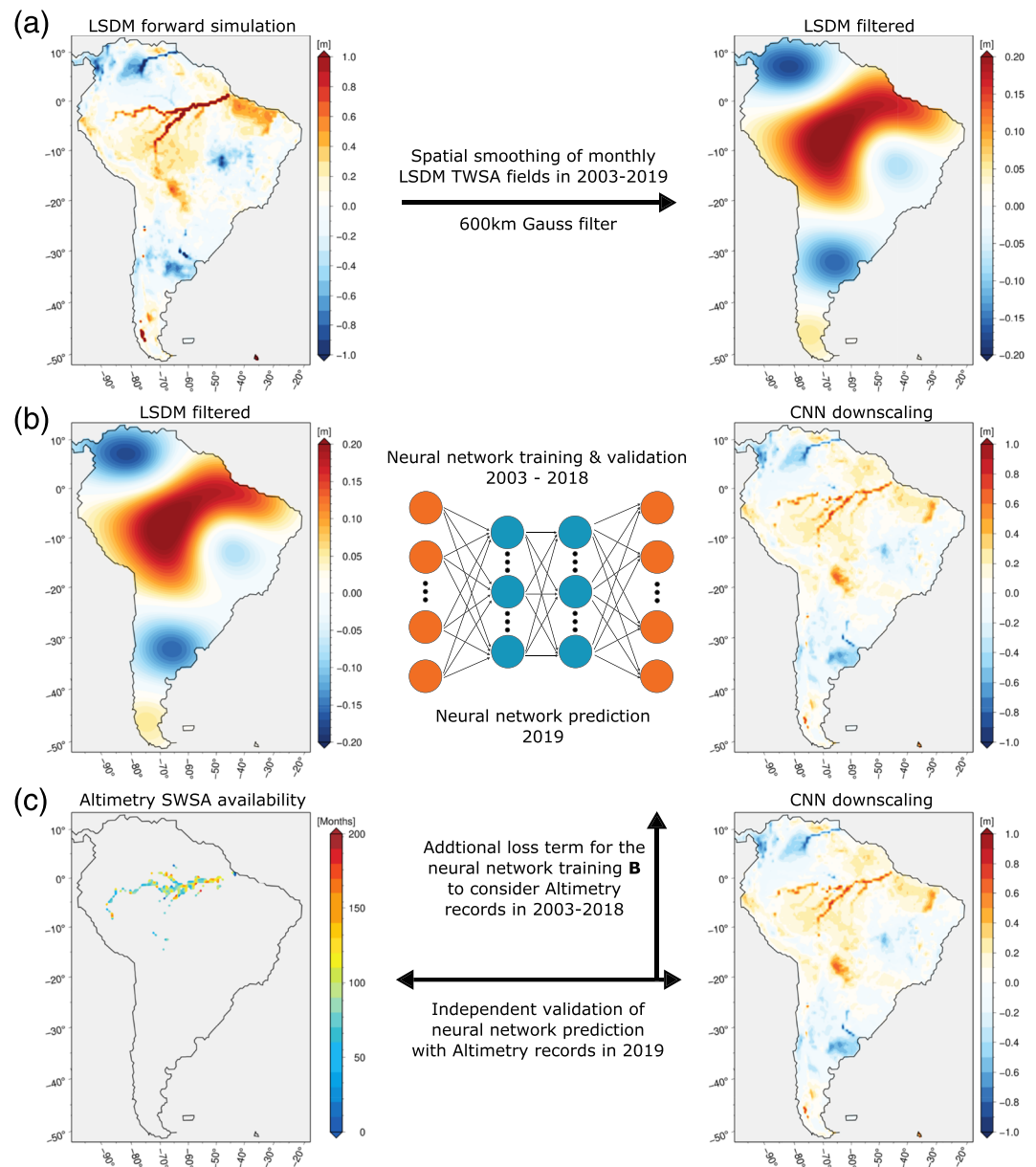
Figure 1. Methodology sketch of this study. (a) Monthly-mean terrestrial water storage anomalies (TWSAs) of the South American continent are calculated from numerical forward simulations of LSDM over the time period of 2003 to 2019. Corresponding monthly GRACE-like observations are derived by spatially smoothing the LSDM TWSA fields with a 600 km Gauss filter. (b) The convolutional neural network (CNN) is trained with pairs of synthetic GRACE observations as input and LSDM TWSAs as output for the time period of 2003 to 2018. The final year 2019 is used for the neural network prediction. (c) The neural network training is constrained by considering altimetry-based surface water storage anomalies (SWSAs) in the Amazonas region, where available in space and time during the training period 2003 to 2018. The accuracy of the CNN downscaling for 2019 is asserted with respective independent SWSAs.

To gain a volume change time series, these ΔV are accumulated and referenced to a common epoch for all grid cells. By scaling the volume change by the grid cell area, we get SWSAs.

2.3. Deep Learning

The downscaling is performed by a CNN, which is implemented in a Tensorflow (Abadi et al., 2016) and Keras (Chollet, 2015) environment. Similar to feed-forward neural networks, information is passed from the input layer to the output layer through a series of artificial neurons (Rosenblatt, 1958), which are linked through weighted neural connections. In contrast to feed-forward networks that process input data pixel

by pixel, CNNs contain convolutional and pooling layers that function to extract recurring spatial features from the input data. Finding an optimal set of connection weights between the neurons is the task of the neural network training. Here, the supervised training is based on pairs of synthetic GRACE observations as inputs and corresponding monthly averaged simulated TWSAs in South America as outputs (see Figure 1b and section 2.1). As such, the input and output layers of the CNN consist of 14,140 neurons, matching the rectangular 101×140 pixel cutout of South America on the $0.5° \times 0.5°$ LSDM grid. The hidden topology of the CNN consists of three convolutional layers with 64 neurons in each layer, followed by a dense layer with 128 neurons. Quadratic convolution kernels with edge lengths of 10, 5, and 3 pixels, respectively, are used in the convolutional layers. The second and third convolutional layers are each followed by a 2×2 max-pooling layer and a dropout layer with a ratio of 0.25 to prevent overfitting during the training. The hyperbolic tangent activation function is used throughout the network to gate information between the neurons.

The training procedure and subsequent application of the CNN are visualized in Figure 1. Coinciding with the GRACE satellite mission start, the training time period extends from 2003 to 2018, which corresponds to 192 monthly input-output pairs. Unseen data pairs of 2019, which were excluded from the training, are used for analyzing the downscaling and generalization abilities of the trained CNN. The training routine consists of 30 iterations, where in each iteration the 16 years of training data are passed through the network. During the training, the weights of the neural connections are successively adapted with the goal to minimize a loss function that measures the error between the current CNN downscaling and the simulated target values from LSDM.

To provide the CNN the ability to dynamically adapt its training progress based on the altimetry-based SWSA reference, we have designed a customized loss function L that incorporates the SWSAs in the Amazonas region by resembling a data assimilation scheme. As such, the neural network downscaling that is otherwise solely achieved through training with LSDM forward simulations is constrained by an additional observation reference. The loss function L is constructed in terms of mean squared error terms between the neural network downscaling, the LSDM target values, and the altimetry-based records, that is,

$$L(y_{\text{LSDM}}, y_{\text{CNN}}, y_{\text{altimetry}}) \tag{2}$$

$$= \ \alpha \cdot L_{\text{LSDM}}(y_{\text{LSDM}}, y_{\text{CNN}}) + (1 - \alpha) \cdot L_{\text{altimetry}}(y_{\text{altimetry}}, y_{\text{CNN}}) \tag{3}$$

$$= \ \frac{\alpha}{N} \sum_{i=1}^{N} (y_{\text{LSDM},i} - y_{\text{CNN},i})^2 + \frac{1-\alpha}{M} \sum_{j=1}^{M} (y_{\text{altimetry,j}} - \hat{y}_{\text{CNN,j}})^2, \tag{4}$$

where y and $\hat{y}$ are vectorized representations with length N and M $(M < N)$, respectively, of the monthly LSDM, CNN, and altimetry-based TWSA and SWSA maps. The hat denotes the spatial restriction of the neural network downscaling to those regions, where altimetry data are available (see Figure 1c). The scaling factor $0 < \alpha < 1$ allows a weighting between the two loss terms L_{LSDM} and $L_{\text{altimetry}}$. Through this modified training procedure, the CNN gains the ability to self-validate and adapt its training progress that is otherwise only based on the numerical forward simulation from LSDM. For the prediction year 2019, the altimetry-based SWSAs are used as an independent measure to quantify the performance of the CNN. In the style of data assimilation terminology, we designate the term "constrained training" to the CNN learning routine with the custom loss function. After moving α in the allowed range and comparing the corresponding neural network downscaling, we identified $\alpha = 0.25$ as a balanced weight between maintaining the integrity from LSDM and accounting for the synthetic GRACE observations. For the purpose of comparison, we will also show results from the same CNN but with unconstrained training where $\alpha = 1$ is set, that is, $L = L_{\text{LSDM}}$.

3. Results and Discussion

In the first part, we inspect the performance of the CNN by quantifying the ability to generalize the learned spatiotemporal dynamics of the South American continental hydrology, which were captured during the 15 years of training data (Figures 2 and 3). In the second part, we focus on the impact of the constrained neural network training through the custom loss function defined in Equation 4 (Figure 4).
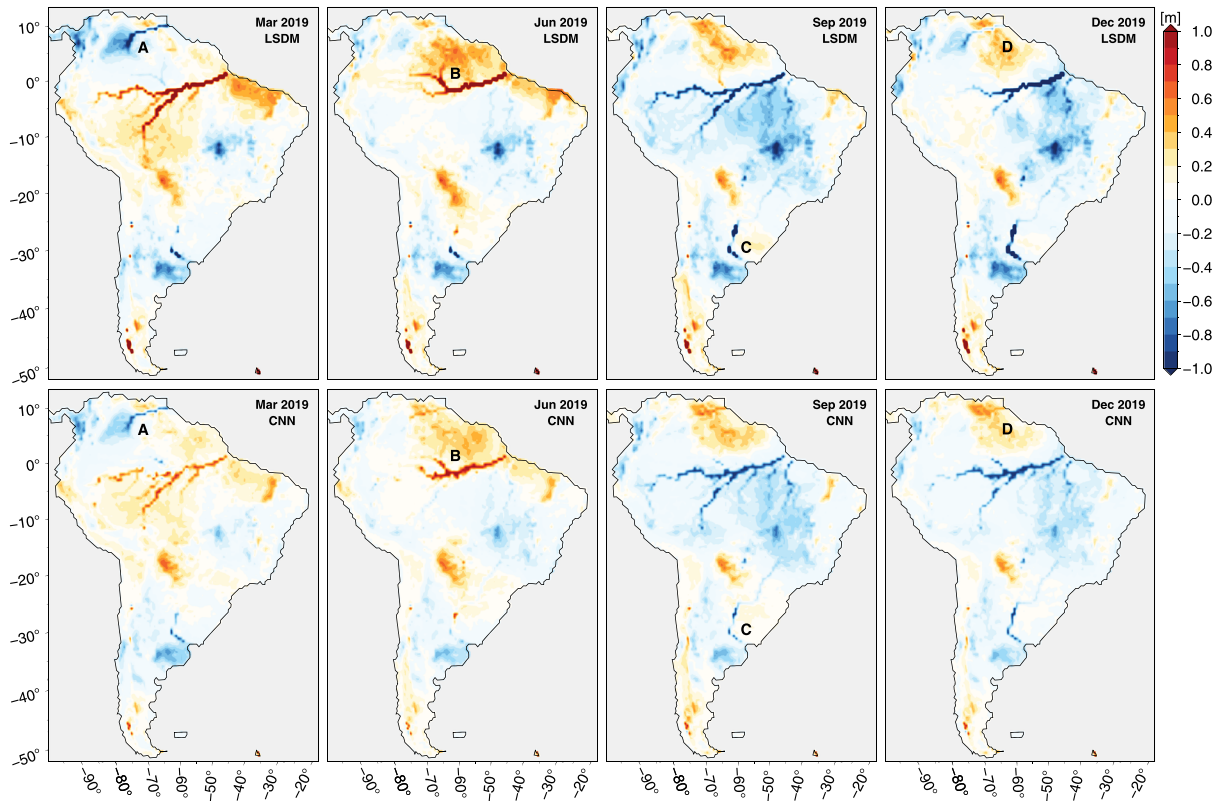
Figure 2. Comparison of monthly-mean TWSAs in selected months of the prediction year 2019. Top: target values simulated by LSDM. Bottom: TWSA downscaling by the convolutional neural network with constrained training. Labels a–d are used for referencing in the main text.

The constrained CNN TWSA downscaling for selected months of 2019 are shown in Figure 2 and compared to the simulated target values of LSDM. The monthly CNN predictions show a close resemblance to the target TWSA from small to large spatial scales. Prominent strong small-scale features and river distributions are accurately reproduced by the CNN, for example, the Orinoco (see Figure 2a), the Amazonas and connecting rivers (Figure 2b), or the Paraná (Figure 2c). Similarly, the CNN is able to correctly infer weaker but large-scale TWSA patterns from soil moisture, wetlands, aquifers, and groundwater distributions from the synthetic gravity fields observations (e.g., Figure 2d). In addition, the annual cycle of the South American TWS in 2019 is reproduced by the CNN throughout the continent. Regions with high TWSAs and high variances, particularly in the Amazons, tend to be underestimated by the CNN. This mismatch of accurately learning to predict values in the tails of the training data distribution is a common behavior of neural networks with a regression task. The respective TWSA downscaling of the unconstrained CNN, as well as difference plots between LSDM and the CNN downscaling, are shown in the supporting information (Figures S1 and S3 in the supporting information).

In addition to the comparison of TWSAs on a monthly basis, we show the annual correlation and explained variance between LSDM and the constrained CNN downscaling in Figure 3 (see Figure S2 for the unconstrained CNN downscaling). Correlation values of 0.8 and higher are found throughout the continent, particularly in tropical regions north of −20° latitude, where rain forests, extensive surface waters, and also large groundwater basins reside (left plot of Figure 3). Lower and negative correlation values are found, for instance, in southern parts of the continent with temperate climate around −30° latitude, or in arid and desert regions along the western coast. In terms of the neural network learning, the observed low correlation values in arid and desert regions can be explained with the locally only marginal TWS occurrence and variability, which translates into little relative contributions to the quadratic loss function used in the training. Inspecting the LSDM variance explained by the CNN shows a similar behavior (right plot of Figure 3). Explained variances of 0.6 and higher are found throughout the continent, with peak values of 0.8 and higher detected in the tropical northern parts of the continent with high total water storage and in the groundwater-rich western part of the southern tip.
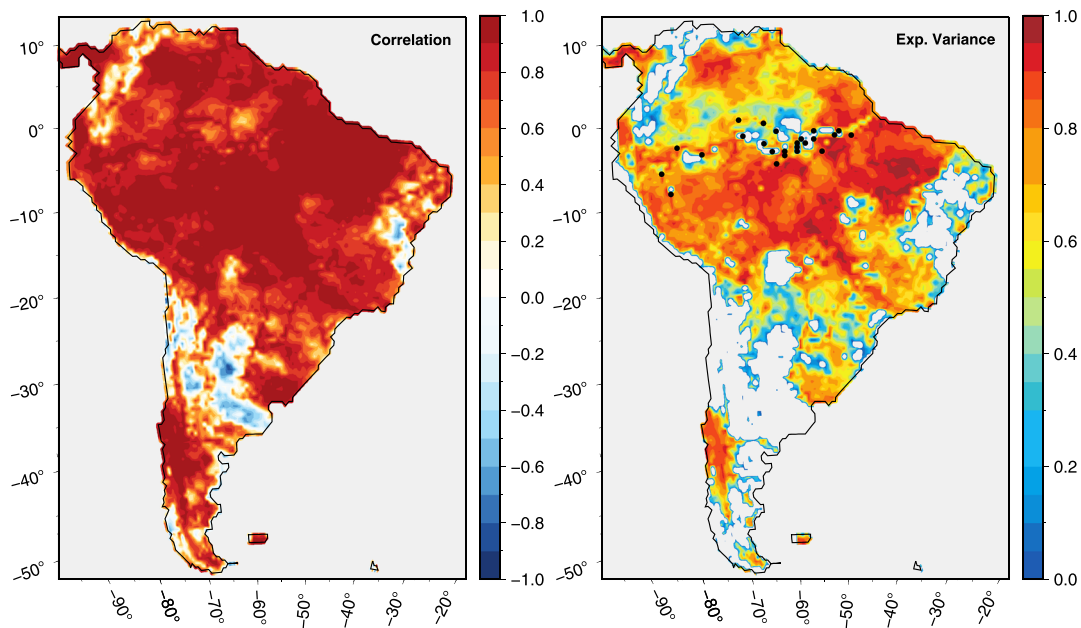
Figure 3. Correlation (left) and explained variance (right) between the LSDM simulation and the CNN downscaling (with constrained training) for 2019. Black dots on the right map indicate the position of validation grid points used for results shown in Figure 4.

The CNN with constrained training is based on a custom loss function to adapt its training progress by considering independent altimetry-based SWSAs (see section 2.3). As such, we utilize the altimetry-based SWSAs as spatially and temporally sparse observation reference during the 2003–2018 training time period, and as an independent validation for the CNN downscaling in 2019. Figure 4a shows monthly RMS errors (RMSE) between the altimetry-based SWSAs and the LSDM forward simulation, the CNN downscaling after constrained training, and the CNN downscaling after unconstrained training (i.e, $L = L_{LSDM}$; see Equation 4). The RMSE trajectories indicate that the CNN with unconstrained training performs worse than the training model LSDM in most months of 2019, when validated against the altimetry-based SWSAs (see orange and black lines in Figure 4a). The large RMS errors between the unconstrained CNN downscaling and the altimetry reference can be explained by two additive error terms: (1) the expected nonperfect replication of the LSDM forward simulation (see also Figure S3) and (2) implicitly learned differences between LSDM and the altimetry records resulting from the missing knowledge about the altimetry records during the training process. Constraining the CNN training mitigates error source (2) significantly (blue line in Figure 4a). As a consequence, the CNN with the constrained training improves the downscaling ability compared to the CNN with standard training considerably. Throughout the year, consistent RMSE reductions between 10% and 40% are found (compare blue and orange dots in Figure 4b). More importantly, the CNN with constrained training outperforms its trainer model LSDM by achieving around 5–45% lower RMSEs in almost all months of 2019 (see blue dots and black reference line in Figure 4b).

The performance of the CNN downscaling is further assessed by its ability to explain the variance of the unseen altimetry-based SWSAs in 2019. For this test, we have selected grid points in the Amazonas region, where the full 2019 time series is available (see locations on the right map of Figures 3 and 4c). The constrained show the overall best performance to explain the variance of the altimetry-based SWSAs. At most locations, explained variance values are consistently higher compared to the forward simulation from LSDM and to the unconstrained CNN. Note that none of the three have knowledge about the altimetry-based ground truth in 2019. However, the constrained CNN is able to generalize its training knowledge from the years 2003–2018, which depends on the respective altimetry-based measurements, to now derive down-scaled water storage values for a new time period that match well with the altimetry reference. At some locations, a degraded performance of the CNN is evident when compared to LSDM, that is, at grid points 1, 11, 15, and 24. This degradation can originate from large error-based deviations between LSDM and the altimetry data, which are transferred to the CNN in the training process. Further, this behavior can be either explained by an insufficient representation of altimetry-based SWSAs at these locations in
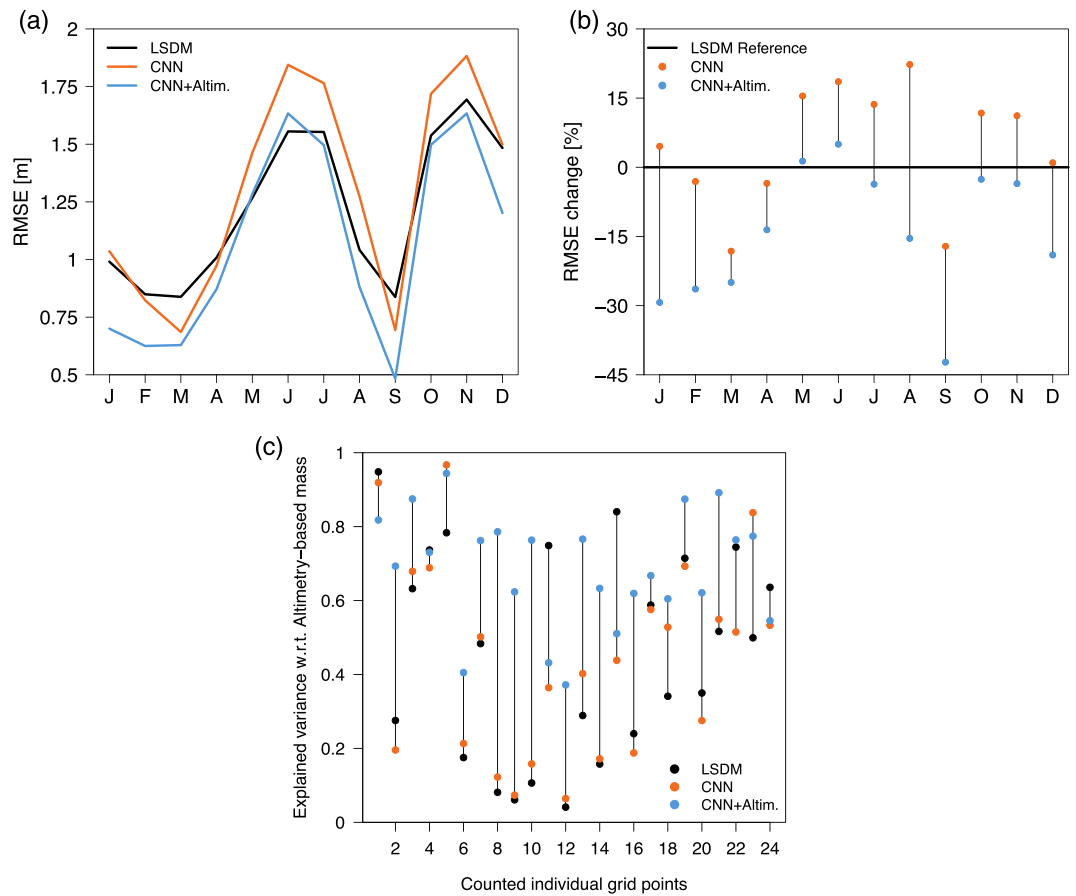
Figure 4. (a) Month-wise RMS error in 2019 between altimetry-based SWSAs (where available in the respective month) and LSDM (black), the CNN with unconstrained training (orange), and the CNN with constrained training (blue). (b) Relative differences between the CNNs RMSE and the LSDM RMSE. Positive values indicate an increase of the CNN RMSE compared to LSDM; negative values indicate a decrease of RMSE. (cx) Explained variance in 2019 between altimetry-based SWSAs and LSDM and the (un)constrained CNN downscaling. Explained variance values are shown for grid points in the Amazonas region, where the full 2019 altimetry-based SWSA time series is available (see positions on the right map of Figure 3).

the sparse training data, or by anomalous events (e.g., large TWSAs or phase changes of annual variations) that cannot be explained through the dynamics learned from the training time period. While constraining the downscaling is in principle constructed to be nonlocal, that is, TWSA values at all grid points (all output neurons) can be modified based on the altimetry records, the dominant changes are found at the locations as described above (see also Figure S3). As such, also, the downscaling from the unconstrained CNN could be a valuable addition to numerical hydrology models, when applied to novel satellite observations.

In summary, we conclude that a CNN with the presented constrained training scheme can combine the advantages of large-scale numerical hydrological models and local real-world observations to perform an accurate downscaling from coarse-structure satellite-based gravity field observations to fine-structure TWS.

4. Conclusions

We combined deep learning, numerical modeling of continental hydrology, and space-borne observation systems to explore whether a trained artificial neural network can derive continental-scale TWS only from GRACE-like gravity field observations as input. For this downscaling task, we have trained a CNN to recover simulated TWSAs in South America from synthetic space-borne gravity field observations. All training data were derived from the same hydrological model to ensure a consistent test environment. Additionally, we introduced a constrained training approach, in which the neural network is provided with independent

altimetry observations of surface water storage to adapt and validate its training progress by constraining its downscaling. The highlights of this study are as follows:

(1) The CNN is able to learn the prevalent multiscale hydrological features in South America from the hydrological model, ranging from local river networks to large-scale total water storage, including ground water basins, soil moisture, and other surface water components.
(2) The CNN accurately performs the downscaling task for novel, that is, unseen, synthetic space-borne gravity field observations over South America.
(3) Validation with independent altimetry-based surface water storage values in the Amazonas region shows that the proposed constrained training scheme allows the CNN to outperform the numerical hydrological model that provides the training data.

In future studies, the CNN training can be adapted toward increased downscaling resolution by either utilizing high-resolution (higher than $0.5° \times 0.5°$) numerical model data, for example the 0.125° solution from Dill et al. (2018), or by including additional constraints, for example, precise river locations such as HydroRIVERS from World Wildlife Fund's HydroSHEDS data (Lehner et al., 2008; Lehner & Grill, 2013) with 15 arc-seconds grid resolution to concentrate observed short-term TWS variations in the river channels. Furthermore, CNN is not confined to a single continent and can also be trained to derive a global downscaling of TWSAs if computational resources allow. We could also include more hydrological variables (different water storage compartments) into the CNN training.

Given the availability of altimetry observations, the usage of constrained machine learning provides a novel methodology for a global multiscale downscaling of GRACE observations. Subsequent applications of our approach can be beneficial for several related topics, including assessments of climate change impacts on continental hydrology, identification, and analyses of ecosystem stresses, or aiding water management in agricultural and metropolitan areas. Additionally, TWS downscaling products can be applied to further improve numerical hydrology models by parameter optimization or data assimilation. Finally, TWS downscaling can be linked to other components of the Earth's hydrosphere, for instance, to provide continental fresh-water forcing for ocean models.

Data Availability Statement

The full Amazon basin altimetry time series were provided by Christian Schwatke on personal request via DAHITI (dahit.dgfi.tum.de). The GIEMS data set is available online (at https://lerma.obspm.fr/spip.php?article91&lang=en). Tensorflow and Keras libraries and codes can be downloaded online (via https://www.tensorflow.org/ and https://keras.io/). The generated data within this study will be made available in a public repository of the German Research Centre for Geosciences (GFZ).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In *12th symposium on operating systems design and implementation*, pp. 265–283.

Ahmed, M., Sultan, M., Elbayoumi, T., & Tissot, P. (2019). Forecasting grace data over the African watersheds using artificial neural networks. *Remote Sensing*, *11*(15), 1769. https://doi.org/10.3390/rs11151769

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing forced climate patterns through an AI Lens. *Geophysical Research Letters*, *46*, 13,389–13,398. https://doi.org/10.1029/2019GL084944

Berry, P. A. M., & Benveniste, J. (2010). Measurement of inland surface water from multi-mission satellite radar altimetry: Sustained global monitoring for climate change. In S. P. Mertikas (Ed.), *Gravity, Geoid and Earth Observation* (pp. 221–229), International Association of Geodesy Symposia. Berlin, Heidelberg: Springer.

Boergens, E., Dettmering, D., & Seitz, F. (2019). Observing water level extremes in the mekong river basin: The benefit of long-repeat orbit missions in a multi-mission satellite altimetry approach. *Journal of Hydrology*, *570*, 463–472. https://doi.org/10.1016/j.jhydrol.2018.12.041

Boergens, E., Nielsen, K., Andersen, O. B., Dettmering, D., & Seitz, F. (2017). River levels derived with CryoSat-2 SAR data classification—A case study in the Mekong River Basin. *Remote Sensing*, *9*(12), 1238. https://doi.org/10.3390/rs9121238

Boergens, E., Rangelova, E., Sideris, M. G., & Kusche, J. (2014). Assessment of the capabilities of the temporal and spatiotemporal ICA method for geophysical signal separation in GRACE data. *Journal of Geophysical Research: Solid Earth*, *119*, 4429–4447. https://doi.org/10.1002/2013JB010452

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*, 376–399. https://doi.org/10.1029/2018MS001472

Chen, L., He, Q., Liu, K., Li, J., & Jing, C. (2019). Downscaling of grace-derived groundwater storage based on the random forest model. *Remote Sensing*, *11*(24), 2979. https://doi.org/10.3390/rs11242979

Chollet, F. (2015). Keras. GitHub, https://github.com/fchollet/keras

Cintra, R. S., & Velho, H. F. C. (2014). Data assimilation by artificial neural networks for an atmospheric general circulation model: Conventional observation. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.5772/intechopen.70791

Dahle, C., Murbck, M., Flechtner, F., Dobslaw, H., Michalak, G., Neumayer, K. H., et al. (2019). The GFZ GRACE RL06 monthly gravity field time series: Processing details and quality assessment. *Remote Sensing*, *11*(18), 2116. https://doi.org/10.3390/rs11182116

Dill, R. (2008). Hydrological model LSDM for operational Earth rotation and gravity field variations: Helmholtz-Zentrum Potsdam Deutsches GeoForschungsZentrum.

Dill, R., & Dobslaw, H. (2019a). Seasonal variations in global mean sea level and consequences on the excitation of length-of-day changes. *Geophysical Journal International*, *218*(2), 801–816. https://doi.org/10.1093/gji/ggz201

Dill, R., Dobslaw, H., & Thomas, M. (2019b). Improved 90-day Earth orientation predictions from angular momentum forecasts of atmosphere, ocean, and terrestrial hydrosphere. *Journal of Geodesy*, *93*(3), 287–295. https://doi.org/10.1007/s00190-018-1158-7

Dill, R., Klemann, V., & Dobslaw, H. (2018). Relocation of river storage from global hydrological models to georeferenced river channels for improved load-induced surface displacements. *Journal of Geophysical Research: Solid Earth*, *123*, 7151–7164. https://doi.org/10.1029/2018JB016141

Frappart, F., Do Minh, K., L'Hermitte, J., Cazenave, A., Ramillien, G., Le Toan, T., & Mognard-Campbell, N. (2006). Water volume change in the lower Mekong from satellite altimetry and imagery data. *Geophysical Journal International*, *167*(2), 570–584. https://doi.org/10.1111/j.1365-246X.2006.03184.x

Frappart, F., Papa, F., Malbeteau, Y., Len, J., Ramillien, G., Prigent, C., et al. (2014). Surface freshwater storage variations in the Orinoco floodplains using multi-satellite observations. *Remote Sensing*, *7*(September 2014), 89–110. (tex.mendeley-tags: Inland) https://doi.org/10.3390/rs70100089

Hagemann, S., & Dümenil, L. (1998). A parametrization of the lateral waterflow for the global scale. *Climate Dynamics*, *14*(1), 17–31. https://doi.org/10.1007/s003820050205

Hagemann, S., & Gates, L. D. (2003). Improving a subgrid runoff parameterization scheme for climate models by the use of high resolution data derived from satellite observations. *Climate Dynamics*, *21*(3-4), 349–359. https://doi.org/10.1007/s00382-003-0349-x

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. *Nature*, *573*, 568–572. https://doi.org/10.1038/s41586-019-1559-7

Humphrey, V., Gudmundsson, L., & Seneviratne, S. I. (2016). Assessing global water storage variability from GRACE: Trends, seasonal cycle, subseasonal anomalies and extremes. *Surveys in Geophysics*, *37*, 357–395. https://doi.org/10.1007/s10712-016-9367-1

Irrgang, C., Saynisch, J., & Thomas, M. (2019). Estimating global ocean heat content from tidal magnetic satellite observations. *Scientific Reports*, *9*, 7893. https://doi.org/10.1038/s41598-019-44397-8

Irrgang, C., Saynisch-Wagner, J., & Thomas, M. (2020). Machine learning-based prediction of spatiotemporal uncertainties in global wind velocity reanalyses. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001876. https://doi.org/10.1029/2019MS001876

Kadow, C., Hall, D. M., & Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*, *13*, 408–413.

Kim, D., Yu, H., Lee, H., Beighley, E., Durand, M., Alsdorf, D. E., & Hwang, E. (2019). Ensemble learning regression for estimating river discharges using satellite altimetry data: Central Congo River as a test-bed. *Remote Sensing of Environment*, *221*, 741–755. https://doi.org/10.1016/j.rse.2018.12.010

Kusche, J., Schmidt, R., Petrovic, S., & Rietbroek, R. (2009). Decorrelated grace time variable gravity solutions by GFZ, and their validation using a hydrological model. *Journal of Geodesy*, *83*(10), 903–913.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3–10. https://doi.org/10.1016/j.gsf.2015.07.003

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, *27*(15), 2171–2186.

Lehner, B., Verdin, K., Jarvis, A., & Grill, G. (2008). New global hydrography derived from spaceborne elevation data. *Eos, Transactions, AGU*, *89*(10), 93–94.

Michailovsky, C. I., McEnnis, S., Berry, P. A. M., Smith, R., & Bauer-Gottwein, P. (2012). River monitoring from satellite radar altimetry in the Zambezi River basin. *Hydrology and Earth System Sciences*, *16*(7), 2181–2192. https://doi.org/10.5194/hess-16-2181-2012

Neelmeijer, J., Schöne, T., Dill, R., Klemann, V., & Motagh, M. (2018). Ground deformations around the Toktogul Reservoir, Kyrgyzstan, from Envisat ASAR and Sentinel-1 data: A case study about the impact of atmospheric corrections on InSAR time series. *Remote Sensing*, *10*(3), 462. https://doi.org/10.3390/rs10030462

Peltier, W. R. (2004). Global glacial isostasy and the surface of the ice-age Earth: The ICE-5G (VM2) model and GRACE. *Annual Review of Earth and Planetary Sciences*, *32*, 111–149.

Piretzidis, D., Sra, G., Karantaidis, G., Sideris, M. G., & Kabirzadeh, H. (2018). Identifying presence of correlated errors using machine learning algorithms for the selective de-correlation of GRACE harmonic coefficients. *Geophysical Journal International*, *215*, 375–388. https://doi.org/10.1093/gji/ggy272

Prigent, C., Papa, F., Aires, F., Rossow, W. B., & Matthews, E. (2007). Global inundation dynamics inferred from multiple satellite observations, 1993–2000. *Journal of Geophysical Research*, *112*, D12107. https://doi.org/10.1029/2006JD007847

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., & Lo, M. H. (2018). Emerging trends in global freshwater availability. *Nature*, *557*(7707), 651–659. https://doi.org/10.1038/s41586-018-0123-1

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Santos da Silva, J., Calmant, S., Seyler, F., Rotunno Filho, O. C., Cochonneau, G., & Mansur, W. J. (2010). Water levels in the Amazon basin derived from the ERS-2 and ENVISAT radar altimetry missions. *Remote Sensing of Environment*, *114*(10), 2160–2181. https://doi.org/10.1016/j.rse.2010.04.020

Schrama, E. J. O., Wouters, B., & Lavalle, D. A. (2007). Signal and noise in gravity recovery and climate experiment (GRACE) observed surface mass variations. *Journal of Geophysical Research*, *112*, B08407. https://doi.org/10.1029/2006JB004882

Schwatke, C., Dettmering, D., Bosch, W., & Seitz, F. (2015). Dahiti—An innovative approach for estimating water level time series over inland waters using multi-mission satellite altimetry. *Hydrology and Earth System Sciences*, *19*(10), 4345–4364.

Seo, J. Y., & Lee, S.-I. (2019). Spatio-temporal groundwater drought monitoring using multi-satellite data based on an artificial neural network. *Water*, *11*(9), 1953. https://doi.org/10.3390/w11091953

Seyoum, W., Kwon, D., & Milewski, A. (2019). Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system. *Remote Sensing*, *11*(7), 824. https://doi.org/10.3390/rs11070824

Singh, A., Kumar, U., & Seitz, F. (2015). Remote sensing of storage fluctuations of poorly gauged reservoirs and state space model (SSM)-based estimation. *Remote Sensing*, *7*(12), 17,113–17,134.

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 Resources Research*, *56*, e2019WR026250. https://doi.org/10.1029/2019WR026250

Sun, A. Y., & Scanlon, B. R. (2019). How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters*, *14*(7), 073,001. https://doi.org/10.1088/1748-9326/ab1b7d

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 Resources Research*, *55*, 1179–1195. https://doi.org/10.1029/2018WR023333

Swenson, S. C., & Wahr, J. (2006). Post-processing removal of correlated errors in GRACE data. *Geophysical Research Letters*, *33*, L08402. https://doi.org/10.1029/2005GL025285

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., et al. (2019). Contributions of grace to understanding climate change. *Nature Climate Change*, *9*, 358–369. https://doi.org/10.1038/s41558-019-0456-2

Tourian, M. J., Schwatke, C., & Sneeuw, N. (2017). River discharge estimation at daily resolution from satellite altimetry over an entire river basin. *Journal of Hydrology*, *546*, 230–247. https://doi.org/10.1016/j.jhydrol.2017.01.009

Velicogna, I., Mohajerani, Y., Geruo, A., Landerer, F., Mouginot, J., Noel, B., et al. (2020). Continuity of ice sheet mass loss in Greenland and Antarctica from the GRACE and GRACE Follow-On missions. *Geophysical Research Letters*, *47*, e2020GL087291. https://doi.org/10.1029/2020GL087291

Wahle, K., Staneva, J., & Guenther, H. (2015). Data assimilation of ocean wind waves using neural networks. A case study for the German Bight. *Ocean Modelling*, *96*, 117–125. https://doi.org/10.1016/j.ocemod.2015.07.007

Weyn, J. A., Durran, D. R., & Caruana, R. (2019). Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. *Journal of Advances in Modeling Earth Systems*, *11*, 2680–2693. https://doi.org/10.1029/2019MS001705

Zhang, L., Dobslaw, H., Stacke, T., Güntner, A., Dill, R., & Thomas, M. (2017). Validation of terrestrial water storage variations as simulated by different global numerical models with GRACE satellite observations. *Hydrology and Earth System Sciences*, *21*(2), 821–837. https://doi.org/10.5194/hess-21-821-2017