https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Hierarchical Graph Networks for Seasonal Forecasts of Terrestrial Water Storage Anomalies

Viola Steidl[1,2], Jürgen Kusche[3], Fupeng Li[3], and Xiao Xiang Zhu[1,2]

[1]Chair of Data Science in Earth Observation, Technical University of Munich, Munich, Germany
[2]Munich Center for Machine Learning (MCML), Munich, Germany
[3]Institute of Geodesy and Geoinformation, University of Bonn, Bonn, Germany

Correspondence: Viola Steidl (viola.steidl@tum.de)

Abstract. Fresh water availability is critical for ecosystems, agriculture, industry, and human communities. Anticipating drought conditions benefits from forecasting changes in terrestrial water storage (TWS), the total water stored on land across all compartments, including groundwater, rivers, glaciers, and soil moisture. While individual compartments, such as groundwater, are difficult to observe directly at large scales, TWS integrates their combined changes and can be measured globally through satellite gravimetry. Since 2002, the GRACE and GRACE Follow-On (GRACE-FO) missions have delivered monthly, global estimates of terrestrial water storage anomalies (TWSA), deviations from a long-term mean, making TWSA the most accessible large-scale indicator of hydrological change. Predicting TWSA is nonetheless challenging as it reflects processes operating at vastly different temporal and spatial scales. We present HiGNN-LSTM, a hierarchical graph neural network that represents the Earth across two spatial scales, coupled with an LSTM module to forecast global TWSA for up to six months ahead. As a proof of concept, we show that the hierarchical graph neural network can automatically generate meaningful input features for TWSA forecasting from ERA5 climate variables, without requiring manual predictor selection or lag-correlation analysis. Trained on a GRACE-like reconstruction of TWSA extending from 1979 to 2020, the model substantially reduces the one-month-lead RMSE relative to a seasonal climatology baseline (1.83 cm vs 3.70 cm) and consistently outperforms a ConvLSTM across the full six-month horizon. Skill over climatology shrinks at longer leads and is lost by six months, indicating that most of the gain concentrates at short leads. Evaluation against GRACE- and GRACE-FO-derived TWSA highlights the difficulty of transferring a model trained on reconstructed TWSA to satellite-derived observations.

1 Introduction

The total amount of water stored on land is referred to as terrestrial water storage (TWS) (Humphrey et al., 2023). It includes all natural or artificial water reservoirs, like groundwater, rivers, glaciers, water stored in biomass, or human-made aquifers. Therefore, TWS is an essential climate variable for the land hydrology (World Meteorological Organization (WMO) et al., 2022). Changes in TWS occur as a response to any of the main water fluxes: precipitation (P), evapotranspiration (ET), and runoff (R)

$$\frac{dTWS}{dt} = P - ET - R. \tag{1}$$

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

However, this view of the system overlooks anthropogenic effects, such as water withdrawals for irrigation or industry. Since 2002, the Gravity Recovery and Climate Experiment (GRACE) and its successor mission GRACE Follow-On (GRACE-FO) have measured changes in the Earth's gravity field globally at a monthly resolution. Terrestrial water storage anomalies (TWSA) are then derived by comparing to a long-term mean of TWS (Tapley et al., 2019).

In recent years, many works have focused on extending the time series of GRACE-like TWSA back to the pre-GRACE era (Humphrey and Gudmundsson, 2019; Yu et al., 2021; Sun et al., 2021; Li et al., 2021; Yin et al., 2023; Palazzoli et al., 2025). These historical reconstructions are used in hydrological modelling and benchmarking, for sea level budget studies, or long-term assessment of changes in the frequency of droughts (Humphrey and Gudmundsson, 2019). However, global forecasting of GRACE-derived TWSA from past observations remains underexplored (Li et al., 2024), although Li et al. (2025) has shown that assimilating forecasted GRACE-derived TWSA into physics-based land surface models can increase their predictive power.

Early work by Forootan et al. (2014) demonstrated that autoregressive models with exogenous inputs could outperform physics-based hydrological models in hindcast experiments over West Africa. They demonstrate that the strong autocorrelation of TWSA can be exploited by a data-driven model and that ocean-atmosphere teleconnections carry meaningful predictive information for forecasting TWSA. Later, Ahmed et al. (2019) used a nonlinear autoregressive with exogenous input (NARX) model to forecast GRACE-derived TWSA over 10 major African basins and found that their model outperformed ARX and a multilinear regression (MLR) model. Wang and Chen (2022) also focus on NARX models to forecast GRACE-derived TWSA. For 11 basins, they identify optimal hydrological variables and their combinations as exogenous inputs for each basin's model. Additionally, they analysed which time delay of these variables leads to the best performance and thus achieve high correlation coefficients for their forecasts.

Li et al. (2024) explore how data-driven methods, including Long Short-Term Memory (LSTM) models, trained on a specific set of lagged climate and hydrological variables, perform when forecasting GRACE-derived TWSA. For each lead time in their forecasting horizon, they identify hydrometeorological variables with optimal lag times and strong correlations with observed TWSA as predictors. Then, they create a separate model for each lead time trained on the most correlated set of predictors, thereby generating a global forecast of gridded TWSA data. In Li and Kusche (2026), they advance the method to produce a semi-operational global forecast of TWSA for up to 12 months.

All the above-mentioned works focus heavily on identifying optimal predictors of GRACE-derived TWSA, either from teleconnections, time-delayed correlations, or both. In addition, they rely solely on correlation as the criterion for identifying optimal predictors, which may lead to biased or suboptimal results, especially when relationships are complex and nonlinear. Graph-based learning methods can learn large-scale interactions over long timescales directly from data. For example, Cachay et al. (2021) employed a Graph Neural Network (GNN) for forecasting the Oceanic Niño Index (ONI) six months ahead. By establishing a novel graph learning module that learns the graph structure from data, the method outperforms other deep learning approaches. However, their study region is limited to the ocean at a resolution of $5°$, yielding 1345 nodes in their graph. While this approach shows promising results for forecasting ONI, a single-valued index, global forecasting of TWSA at a $1°$ resolution would quickly become computationally intensive.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

In this work, we employ a hierarchical graph representing the world at two spatial scales, the grid level and the hydrological basin or oceanic region level. This approach guarantees that the computation remains tractable while also including hydrological domain knowledge. We use a GNN to encode processes at different spatial and temporal scales and generate sequences of latent features. These latent feature sequences will serve as input features to a recurrent neural network that predicts monthly TWSA over the next six months. This means that, given any fixed input set, the optimal predictors and their lag correlations do not need to be identified manually, but the GNN learns them directly from data.

We build and train this model on data from a TWSA reconstruction product (Li et al., 2021) and ERA5 climate variables. First, we analyse how the GNN latent features contribute to the forecasts from the recurrent neural network. Then, we investigate how well our model can forecast TWSA by evaluating on a held-out test set from the TWSA reconstruction. Lastly, we measure how far the trained model transfers to GRACE-derived TWSA observations and compare the results to a TWSA forecast from a model trained directly on GRACE-derived TWSA.

We frame this study as a proof of concept: the aim is to demonstrate that a hierarchical graph architecture can learn useful predictors of TWSA end-to-end, not to establish state-of-the-art operational skill.

2 Hierarchical Graph Neural Network with LSTM

Our model uncovers spatio-temporal relationships in a fixed set of climate inputs and uses them to forecast TWSA without requiring manual predictor selection or lag-correlation analysis. We combine a GNN with a Long Short-Term Memory-based (LSTM) module. The GNN is employed to create four sequences of artificial latent-space features for each lead time in the forecast horizon. These artificial features encode the spatio-temporal information of the original input features. The temporal component of the model, the LSTM module, takes these features, along with the last 12 months of TWSA and an embedding of the lead time in months, to produce TWSA forecasts up to six months ahead. A preliminary version of this architecture is presented in Steidl and Zhu (2025). Figure 1 shows the complete model architecture. We refer to our model as HiGNN-LSTM (Hierarchical Graph Neural Network with LSTM) in the following. All model architecture and hierarchical graph parameters were selected via manual tuning on the validation set. Ablations justifying the key choices are reported in App. B.

2.1 Graph construction

TWSA at a given location responds not only to local conditions but also to processes elsewhere on the globe, expressed as teleconnections and lagged correlations among climate variables. Because such connections span multiple spatial scales, we employ a hierarchical graph to uncover them from the input data. The hierarchical graph consists of a high-resolution grid level and a coarser mesh level.

The graph's grid-level nodes correspond to the land grid cells, holding a time series of variables described in Sect. 3 as node features. To keep this graph level compact, the ocean is not represented cell by cell. Instead, ocean cells are grouped into 20 clusters by geographic location (Figure A1). Each ocean node carries the mean time series of its member cells' variables as node features.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

We refer to the nodes of the first graph level as *grid nodes* in the following. At a grid resolution of $1° \times 1°$, we get 20370 grid nodes in total.

At the graph's mesh level, we have nodes corresponding to the hydrological basins defined by Lehner and Grill (2013). Additionally, the 20 ocean nodes are carried over to the mesh level. We refer to this set of nodes as the *mesh nodes*. Their features are initialised from each node's longitude and latitude after a sine-cosine embedding.

The graph's connectivity is defined by three sets of edges connecting the nodes of the two levels. Nodes of the grid level link to the nodes of their hydrological basin or ocean cluster. At the mesh level, nodes connect to their ten spatially nearest neighbours. Every ocean-cluster node additionally links to all basin nodes, reflecting our assumption of a far-reaching oceanic influence on TWSA. The third set of edges links mesh-level nodes back to the grid level, where each grid node is linked to three nearest mesh neighbours.

The number of ocean clusters, the number of edges per node at the second graph level, and the number of edges from the second graph level to each node of the first level were set to 20, 10, and 3, respectively. These values balance graph connectivity against computational cost. In App. B, we perform an ablation study, varying each parameter on the validation set, to confirm that model performance is robust around the chosen configuration.

2.2 Grid–Mesh–Grid spatial encoding

The information is passed along the different graph levels by a pipeline of three separate GNNs: Grid2Mesh, Mesh2Mesh, and Mesh2Grid. These transform regular gridded Earth observation products into sparse graph-based representations and back to a gridded space of latent features. Following GraphCast (Lam et al., 2023), every stage is built from simple multi-layer perceptrons (two linear layers with a Batch Normalization layer between them and nonlinear activations). Our implementation builds on Dufourg et al. (2024).

Before feeding any information to the GNNs, we embed node and edge features into a shared 64-dimensional space. Each type of node and the edge is embedded by a separate MLP. Then the Grid2Mesh GNN propagates the embedded information from grid nodes to mesh nodes along their edges, updating both the destination node features and the edge attributes.

After that, the information is processed at the mesh level. Two sequential Mesh2Mesh GNNs pass information along edges to connected nodes and update their features. These updates are modulated on the lead-month index i through Feature-wise Linear Modulation (FiLM) as described in Perez et al. (2017) to generate distinct encodings for each lead time.

Finally, the Mesh2Grid GNN collects information at the mesh level and passes it back to the grid level nodes to update their feature sequences. These updates are also modulated by FiLM depending on the lead-month index i.

An output layer collapses the spatio-temporal information from the shared 64-dimensional space into a 4-dimensional sequence of latent features for every land grid node. This is the spatially encoded input to the following temporal block.

The ablation in App. B confirms that the full Grid2Mesh $\longrightarrow$ Mesh2Mesh $\longrightarrow$ Mesh2Grid pipeline contributes meaningfully to forecast skill relative to an embedder-only variant that bypasses all message passing GNN layers.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

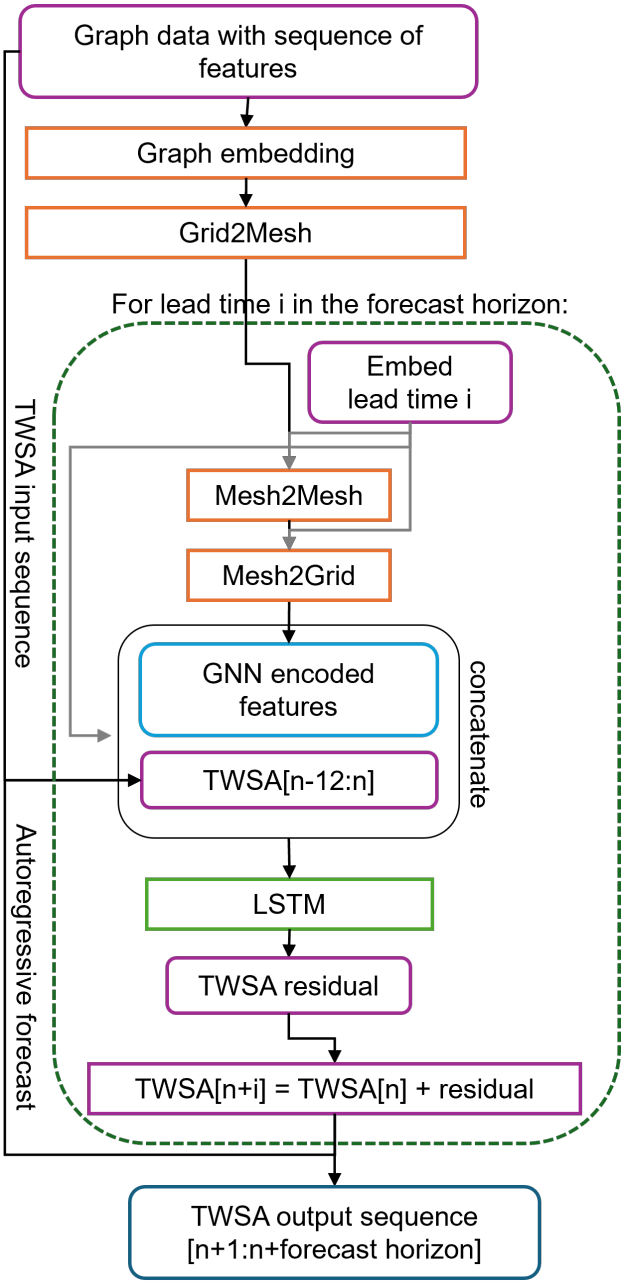

Figure 1. Scheme of the model architecture.

2.3 Spatio-temporal forecasting

The temporal prediction is handled by a small recurrent module consisting of two LSTM layers with 32 neurons, and a linear output layer. The temporal module produces a six-month forecast of TWSA. At each lead time i, the module receives three

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

inputs: the preceding 12 months of TWSA, the lead time specific latent-feature sequence from the Grid–Mesh–Grid encoder, and an embedding of the lead-time index i. Rather than predicting the field directly, it outputs a residual that is added to the most recent TWSA observation to give the following month's global field.

3 Data

3.1 TWSA from GRACE-reconstruction

Measurements from the GRACE and GRACE-FO missions date back only to 2002 and are available at a monthly resolution, with an 11-month gap between the two missions. This poses a significant hurdle for applying deep learning approaches, which rely on having a vast amount of training data points to learn from. Therefore, we develop our method on a reconstruction of GRACE-derived TWSA from Li et al. (2021), spanning 1979 to 2020 at monthly $0.5° \times 0.5°$ grid resolution. This dataset provides over twice as many time steps as would be available from only GRACE/GRACE-FO measurements.

Li et al. (2021) created their data product by combining three different data-driven techniques. For each region, they combine one of two statistical methods, one of three time-series decomposition methods, and one of three machine learning models. On top of that, they identify the optimal predictors from a set of climatic and hydrological variables for each region. The RL06 GRACE mascon solutions (Save, 2019; Save et al., 2016) from April 2002 to June 2017 are used for model training, while data from the GRACE-FO mission between June 2018 and June 2020 are used for validation. Additionally, the TWSA reconstruction is validated on temporal gravity field models derived from the satellite laser ranging (SLR) technique (Löcher and Kusche, 2020) for the pre-GRACE time period from 1992 to 2002. Li et al. (2021) compare their product to the TWSA reconstruction from Humphrey and Gudmundsson (2019). Both at the grid and basin scale, the TWSA reconstruction from Li et al. (2021) fits better with SLR TWSA.

For our study, we use the detrended version of the data product, where they removed linear long-term trends, but kept the seasonal signal (Li et al., 2021). Antarctica is not included in the dataset.

3.2 Climatic and hydrological variables

To exploit the full 1979–2020 training period offered by the TWSA reconstruction, we use readily available variables from ERA5 (Hersbach et al., 2023) instead of direct observations. We collect 8 variables (see Table 1) from ERA5 monthly-averaged single-level data (Hersbach et al., 2023). These variables are either directly linked to TWSA through the terrestrial water budget (precipitation, runoff, and evaporation) or have a more subtle correlation with TWSA. It is important to note that ERA5 data was created by a land-atmosphere model constrained by observations through data assimilation. Therefore, our predictor variables exhibit systematic biases, particularly in regions with sparse observations and where model physics dominate (Hersbach et al., 2020). Additionally, we provide the model with the month of the year as input.

All inputs are coarsened to a $1° \times 1°$ resolution by latitude-weighted averaging over the points in a grid cell. The linear trends are removed using `scipy.signal.detrend` from the SciPy library.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Table 1. Input features.

Feature	Source	Original resolution
TWSA	Li et al. (2021)	$0.5° \times 0.5°$
Total precipitation	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Surface pressure	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
2m temperature	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
10m wind speed	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Evaporation	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Potential evaporation	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Runoff	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Sea surface temperature	ERA5 (Hersbach et al., 2023)	$0.25° \times 0.25°$
Month of the year	-	$1° \times 1°$

3.3 GRACE-derived TWSA

We want to measure how well the model trained on a reconstruction of GRACE-derived TWSA transfers to predicting TWSA derived directly from measurements. Therefore, we collect RL06.03 monthly GRACE/-FO mascon solutions from (Save, 2025; Save et al., 2016) derived by the Center for Space Research (CSR) for the time period from March 2008 to December 2011 and November 2018 to November 2023. These two time periods have uninterrupted monthly data, so we do not have to impute any missing measurements. RL06.03 GRACE/-FO derived TWSA comes at a $0.25° \times 0.25°$ resolution, which we coarsen to a $1° \times 1°$ in the same way we processed the reconstructed TWSA data. Since we are training the model to forecast detrended TWSA, we also detrend this data for fair comparison using `scipy.signal.detrend`.

4 Results

The model trains on the reconstructed GRACE-like TWSA data product, which we split into training (July 1979 - March 2008), validation (April 2008 - May 2016), and test (June 2016 - June 2020) time series. We employ pixel-wise normalization along the temporal dimension. Validation and test datasets are normalized using the mean and standard deviation of the training set. As TWSA for Antarctica is not present in our training dataset, we exclude predictions for Antarctica from the model training, i.e., the loss calculation. Also, we apply a latitude-aware weighting to the loss calculation to account for the higher grid-cell density at high latitudes. The model is trained with the AdamW optimizer (Loshchilov and Hutter, 2019) (learning rate $= 5e^{-4}$, weight decay$= 1e^{-2}$, batch size $= 16$) optimizing a latitude-weighted Mean Squared Error (MSE) for 100 epochs. To prevent overfitting, we save the model checkpoints of the best validation score and proceed with our analysis with those. All experiments were conducted on a single GPU (NVIDIA A40).

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

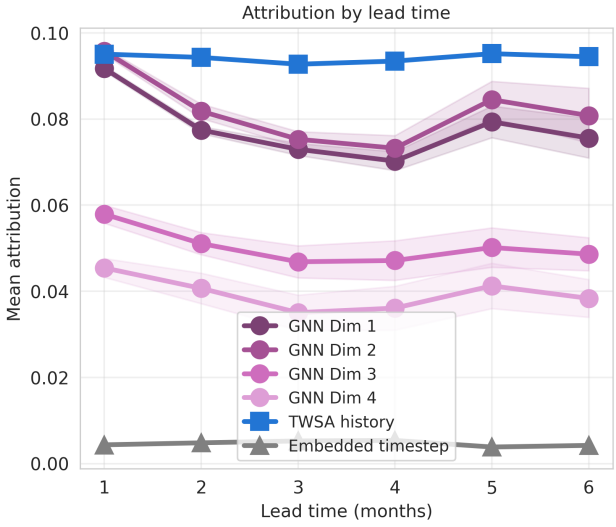

Figure 2. SmoothGrad attributions for the input to the LSTM component of the model for each lead time.

4.1 Gradient analysis of GNN latent features

First, we analyze how the latent features created from the GNN block are used in the subsequent temporal block featuring the LSTM module. The GNN encodes the input features of all grid nodes and every input time step into four latent feature sequences of 12 time steps per node and per lead time within the forecast horizon. To assess how the LSTM component of our model utilises the information encoded by the GNN, we perform gradient-based saliency analysis on the trained model. Specifically, we employ SmoothGrad (Smilkov et al., 2017) implemented via the Captum library (Kokhlikyan et al., 2020) using `Saliency` and `NoiseTunnel` functions. The `Saliency` approach returns the gradients of a model output with respect to the inputs. The `NoiseTunnel` perturbs the inputs by adding Gaussian noise, making the gradient attributions more robust.

The three input streams to the LSTM at each lead time are the GNN latent encoding, the TWSA history, and a sinusoidal embedding of the lead time index. The attributions are computed with respect to the LSTM's TWSA prediction per lead time. They are aggregated as mean absolute attribution values across all nodes, input sequence positions, and batches in the validation set, with standard deviations reported to reflect spatial and temporal variability across samples. The results, shown in Figure 2, confirm that the GNN latent embedding serves as a meaningful input to the LSTM throughout the forecast horizon. The TWSA history has the highest overall attribution, which is not surprising, as we use the last TWSA time step as an offset for the residual prediction. The first two latent features have attributions of comparable magnitude to those of the TWSA history, especially at lead time 1. This shows how important the GNN encoding is to the subsequent temporal block.

All four GNN latent dimensions receive non-negligible attribution, confirming that the embedding is not degenerate. This indicates that the LSTM exploits all four latent dimensions jointly, rather than relying primarily on a single dominant one. We

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Table 2. Globally averaged RMSE [cm] on test set (June 2017 - June 2020) per lead time.

Lead time (months)	HiGNN-LSTM	ConvLSTM	Climatology
1	1.83	2.44	3.70
2	2.66	3.75	3.67
3	3.12	4.36	3.67
4	3.41	4.58	3.67
5	3.62	4.70	3.68
6	3.77	4.83	3.69

want to stress that these dimensions do not carry direct physical interpretations, as they represent a learned compression of the spatio-temporal graph structure rather than explicitly engineered features.

Lastly, the embedding of the lead time index contributes substantially less across all lead times, suggesting the model relies primarily on the data-driven signals rather than explicit temporal positional encoding. This might be because the GNN latent space already encodes sufficient temporal context for the LSTM to produce accurate forecasts, so the lead time embedding adds little additional information.

4.2 Evaluation on the test set

Having trained exclusively on reconstructed TWSA, we first evaluate performance on the held-out test set before evaluating on TWSA from GRACE/-FO observations. As described, we split our dataset into three time periods for training, validation, and testing. The model has never seen the latter during training or hyperparameter tuning. On this test set, the model achieves a globally averaged root mean squared error (RMSE) of 1.83 cm for the lead time of one month. The error rises abruptly by 45% over a two-month lead time. After that, RMSE rises steadily with smaller increases (see Table 2). This type of error progression is expected as the first month in the forecasting horizon is strongly correlated with the last time step of the TWSA observations, whereas longer lead times require the model to rely on more subtle signals in the input features. Figure 3 shows the six-month forecast for a sequence in the test set. The spatial resolution is the same as the input resolution ($1^\circ \times 1^\circ$). Prominent spatial patterns are generally well matched. The Mean Absolute Difference (MAD) rises from 1.20 cm to 2.08 cm.

We compare the HiGNN-LSTM performance on the test set to the performance of a sequence-to-sequence ConvLSTM (Shi et al., 2015) and to the RMSE of a long-term mean for each month (climatology) derived from the training TWSA time series.

A ConvLSTM (Shi et al., 2015) replaces the linear transformations in a standard LSTM with local convolutional filters, capturing not only temporal but also spatial patterns within a limited receptive field. With a comparable number of parameters, it serves as a direct baseline for what local spatial aggregation can achieve, in contrast to the features encoded by the GNN. Implementation details of the ConvLSTM architecture are in App. C. The consistent margin over the ConvLSTM suggests that GNN-encoded features carry more predictive information than spatially local convolutions of comparable parameter count. We compute the climatology from the training time series as the mean TWSA for each calendar month and each grid cell.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

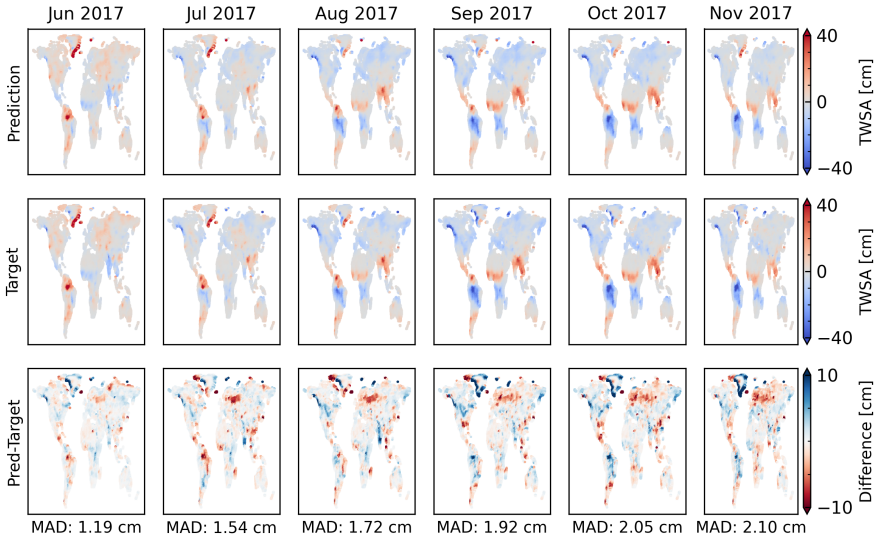

Figure 3. Six-month forecast (June 2017 - November 2017) from input time series June 2016 to May 2017. Rows show (top) model prediction, (middle) target TWSA from the reconstruction, and (bottom) their difference, with the Mean Absolute Difference (MAD) indicated below each difference map. The resolution is $1^\circ \times 1^\circ$.

It represents the average seasonal cycle but cannot respond to interannual variability. Its RMSE, therefore, equals the RMS amplitude of interannual anomalies in the test period. The margin between HiGNN-LSTM RMSE and climatology RMSE in Table 2 can thus be read as the model's skill on the anomaly component alone. This margin is large at lead 1 but shrinks monotonically with lead time, becomes negligible by lead 5, and marginally reverses at lead 6 (HiGNN-LSTM 3.77 cm vs climatology 3.69 cm). The model captures temporal patterns beyond the seasonal cycle at short leads, but it reverts to the seasonal average at longer horizons.

We evaluate our model against the basin-wide mean of 16 major river basins and investigate the temporal performance. Figure 4 shows the mean predicted TWSA in six-month sequences with the start date shifting by one month, and thus covering the whole test period from June 2017 to June 2020. The black curves show the target TWSA from the reconstruction dataset.

A basin-wide evaluation reveals that our model effectively captures seasonality in basins with a strong seasonal component, as evidenced by the almost perfect match with the TWSA of the Amazon or Mackenzie basins, for example. Whenever the TWSA exhibits stronger deviations from the mean seasonal signal, the model struggles to predict these perfectly, but it is still able to follow the trend. The predictions in the basins of Amur, Orange, or Murray show this behaviour. This demonstrates that the model can not yet accurately predict deviations from seasonal TWSA.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

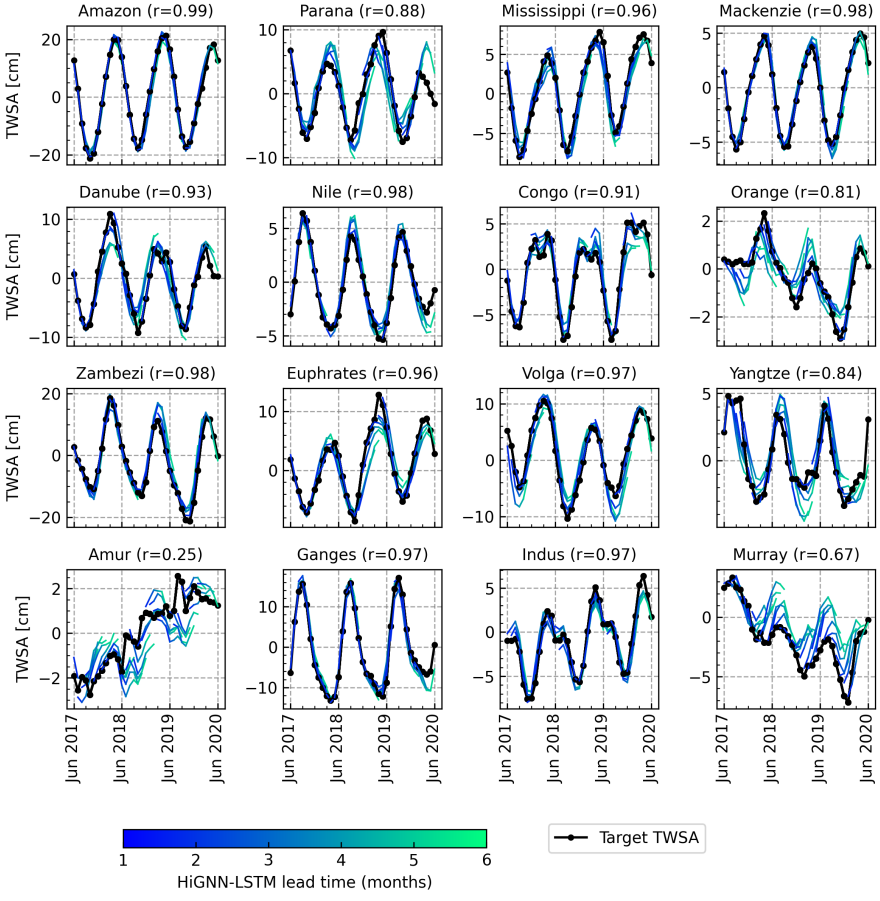

Figure 4. Predicted TWSA (blue) vs target TWSA (black) in 16 test basins together with Pearson's correlation coefficient (r) averaged over all forecast sequences for the given basin from June 2017 until June 2020.

Table 3. RMSE [cm] between GRACE-derived TWSA and HiGNN-LSTM predictions or climatology.

Lead time (months)	HiGNN-LSTM	Climatology
1	5.59	7.98
2	6.63	7.94
3	6.98	7.94
4	6.98	7.96
5	6.89	7.96
6	6.96	7.96

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

4.3 Evaluation on GRACE-derived TWSA

We evaluate to what extent our model generalizes to forecasting GRACE-derived TWSA despite being trained on a reconstruction dataset. Our test period runs from March 2008 to November 2023, with predictions starting in March 2009. This ensures that the test period does not overlap with the training period. Since we do not want to impute missing TWSA data but measure the performance only on GRACE-derived TWSA, we have to skip all periods with less than 18 months of consecutive data. Therefore, we have one test period running from March 2009 to July 2010 and a second one running from November 2019 to November 2023.

Table 3 shows how the performance drops substantially compared to the performance on the test set from the TWSA-reconstruction. This is to be expected since the model must bridge the domain gap between the spatially smoothed TWSA reconstruction and the noisier GRACE-derived TWSA. Figure 5 shows the HiGNN-LSTM predictions against the GRACE-derived TWSA for 16 test basins for the later time period (November 2019 to November 2023). Pearson's correlation coefficients between the GRACE-derived TWSA and HiGNN-LSTM predictions are similar to those from the reconstruction test set. Only for the basins Orange, Yangtze, Indus, and Murray they shrink dramatically.

4.4 Comparison to existing methods

We want to put the model's performance on the GRACE-derived TWSA into perspective by comparing it with how another approach would perform in forecasting TWSA for our 16 test basins. We compare against Li and Kusche (2026), whose design choices differ fundamentally from ours. They train one independent LSTM per forecast lead time directly on GRACE-derived TWSA, with input predictors selected separately for each basin and each lead time through lag-correlation and teleconnection analysis. Prior to training, the seasonal cycle and linear trends are removed, so each model need only predict the residual anomaly relative to this deterministic component. HiGNN-LSTM operates under substantially more constrained conditions: a single model, no per-basin or per-lead-time predictor tuning, trained on a TWSA reconstruction rather than observations, and forecasting TWSA, including the seasonal signal, directly from ERA5 inputs. Li and Kusche (2026) provide a global GRACE-like TWSA forecast (GRACE-FCast) from 2010 until 2024 based on GRACE/-FO derived TWSA from different providers. We compare the predictions from our HiGNN-LSTM model to their predictions based on the CSR RL06.2 mascons (Save et al., 2016) for the period from November 2019 to November 2023. For the comparison, we also removed the linear trends from GRACE-FCast by subtracting the linear trend they extrapolated to create the forecast.

Figure 6 shows the basin-wise forecasts of GRACE-FCast and the forecasts from the HiGNN-LSTM versus the GRACE-FO TWSA. Across most basins, HiGNN-LSTM achieves lower forecast skill than Li and Kusche (2026) in both RMSE and correlation. This can be attributed to two factors. First, Li and Kusche (2026) train directly on GRACE-derived TWSA observations while our model trains on a reconstruction, introducing a distribution shift at inference time. Second, predicting total TWSA including the seasonal cycle is a substantially harder target than the anomaly-only prediction of Li and Kusche (2026), given that the seasonal cycle accounts for the majority of TWSA variance in most basins. Consistent with this, the performance gap is smallest in basins where the seasonal signal dominates and largest where TWSA deviates substantially from the seasonal

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

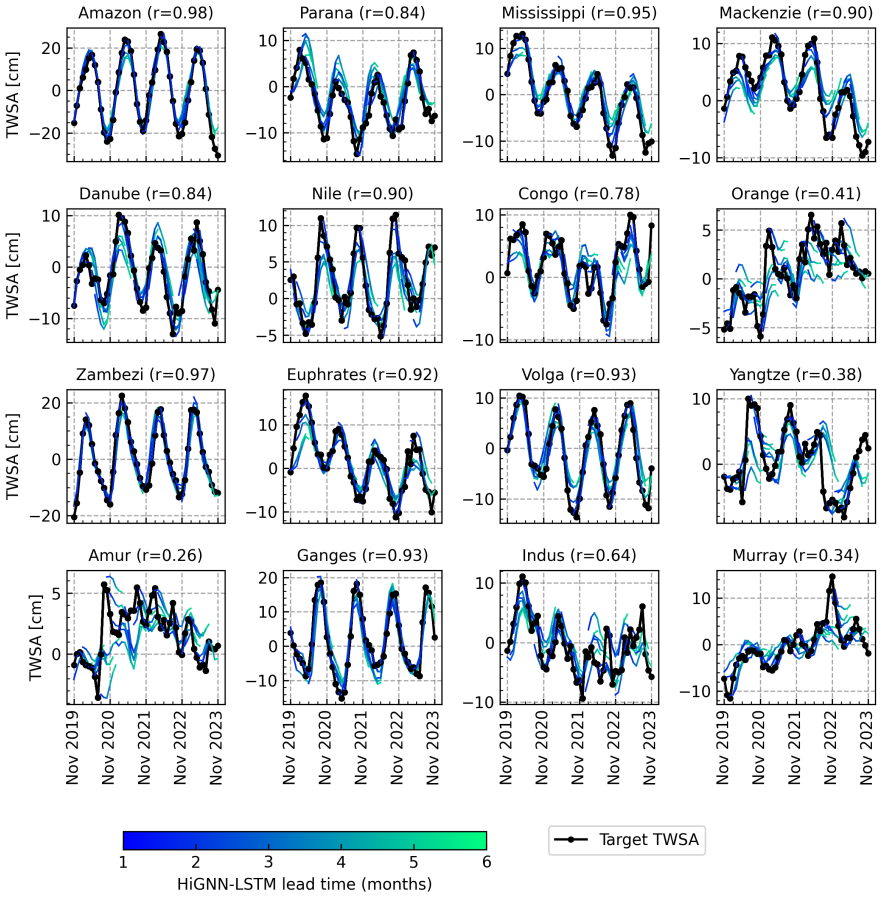

Figure 5. Predicted TWSA (blue) vs target GRACE-derived TWSA (black) in 16 test basins together with Pearson's correlation coefficient (r) averaged over all forecast sequences for the given basin from November 2019 until November 2023.

mean. For example, the gap is small in the Amazon and Mackenzie basins, where the seasonal signal dominates, and largest in the Orange and Murray basins, where anomalies drive most of the variance. Together, these factors explain the gap more than any limitation of the graph-based architecture itself, suggesting that automatic feature learning from ERA5 inputs is a viable approach to TWSA forecasting.

5 Limitations and future work

After investigating the performance of our model, we identify several challenges that must be addressed before forecasting skill can be further improved.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

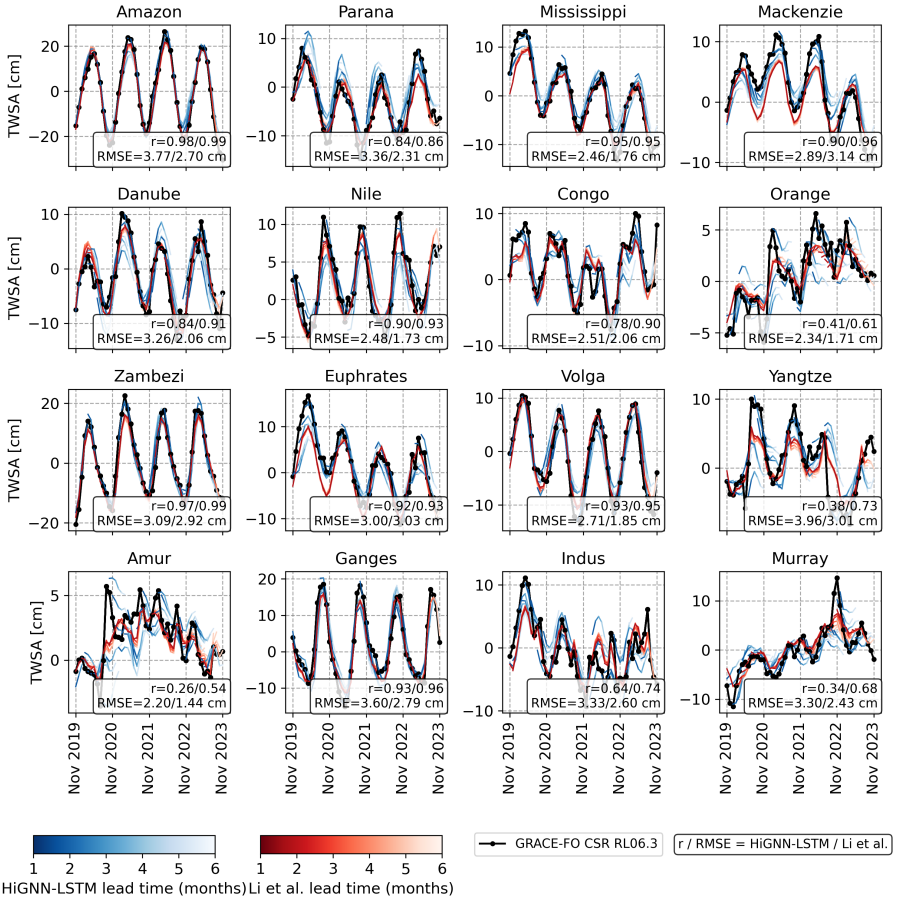

Figure 6. HiGNN-LSTM forecast (blue) and Li et al. TWSA forecast GRACE-FCast (red) vs. GRACE-derived TWSA (black) in the 16 test basins. The lighter the line color of the forecasted TWSA, the longer the lead time. Pearson's Correlation coefficient (r) and Root Mean Squared Error (RMSE) for the basins are given for the HiGNN-LSTM architecture/Li and Kusche (2026).

5.1 Decreased GNN feature importance at longer lead times

Our model relies purely on data from the past year before the forecasting time period and has no access to predictions of future climate variables. Since the correlations between the input features and TWSA are weaker for longer horizons, the GNN encodings will inevitably be less informative for longer lead times. To counteract this, we already train the GNN to generate lead-time-specific spatio-temporal embeddings of the input features by modulating it with the lead time index i. However, the latent space analysis shows that the GNN-encoded features could be improved to produce more informative features for longer lead times. A more explicit temporal encoding at the input stage, for example, a temporal convolution or attention module before the embedder module, could in principle help the GNN generate more informative features for later lead times. Currently, we feed all input features as a flattened vector of all time steps to the model. The GNN has to implicitly identify

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

which time steps are most relevant for each lead time. However, integrating such a module is non-trivial in this architecture: it changes the dimensionality of node features and substantially increases memory and training cost. Whether it would close the longer-lead gap in practice is therefore an open question rather than a straightforward extension.

5.2 Spatial variability

A spatial examination of the predictions reveals high variability in predictive performance (see Figure 3). Basin-level evaluation in Sect. 4.2 shows that the model performs well in basins where the seasonal signal dominates, such as the Amazon or Mackenzie, but struggles in basins where TWSA deviates more substantially from the seasonal mean, such as Orange, Murray, or Amur. Improving performance in anomaly-driven basins likely requires either a decomposition of the seasonal signal before training, as done in Li and Kusche (2026), or a graph construction that better accounts for the hydrological conditions of individual regions. For example, connecting graph nodes based on known spatio-temporal correlations between ERA5 variables and TWSA, rather than purely on spatial proximity, could improve the model's ability to distinguish regional signals. To improve accuracy across basins, it may be beneficial to weight basins according to their size, such that smaller basins receive the same attention as larger ones with more grid cells. Although this may not improve overall accuracy, it may help the model focus more on deviations from the climatology, rather than merely learning the seasonal signal.

5.3 Dataset

As noted in Sect. 3.2, our predictors come entirely from ERA5 and therefore inherit the reanalysis's biases, particularly in sparsely observed regions where model physics dominates. These biases likely already constrain the informativeness of latent features for shorter horizons.

Additionally, the training dataset consists of fewer than 400 input-target sequence pairs. This limits the model complexity that can be justified without risking overfitting, and means that each architectural component needs to be well motivated.

6 Conclusions

Forecasting TWSA is inherently difficult because it reflects the combined dynamics of multiple water compartments operating on different temporal and spatial scales. We have demonstrated, as a proof of concept, that a hierarchical GNN can automatically generate spatially meaningful latent features for TWSA forecasting from ERA5 climate variables. Gradient-based analysis confirms that the GNN encoding is actively used by the LSTM and is not degenerate, validating the use of a graph-based architecture. On reconstructed TWSA, the resulting model outperforms both a climatology baseline and a ConvLSTM, showing that automatically learned spatial teleconnections carry more predictive information than a temporal mean or a locally convolving method. Basin-level evaluation shows strong performance where the seasonal signal dominates but reveals limitations in basins where TWSA deviates substantially from the seasonal mean.

Transferability to GRACE/-FO-derived TWSA remains an open challenge. When evaluated directly against observations, model skill drops substantially compared to the test set. We partly attribute this to the smoothness of the reconstructed GRACE-

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

like TWSA product that may not fully reflect the spatial variability of the observation-derived TWSA. Future progress will therefore likely come from two directions: incorporating more domain knowledge into the graph and mesh latent space to strengthen the informativeness of the latent feature sequence, and training on multiple reconstruction datasets produced with different methodologies to reduce dependence on patterns specific to any single product.

315 *Code and data availability.* The code used in this study is available at https://github.com/viola1593/HiGNN-LSTM. The processed data to train and evaluate the model can be downloaded from https://doi.org/10.5281/zenodo.19664592.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

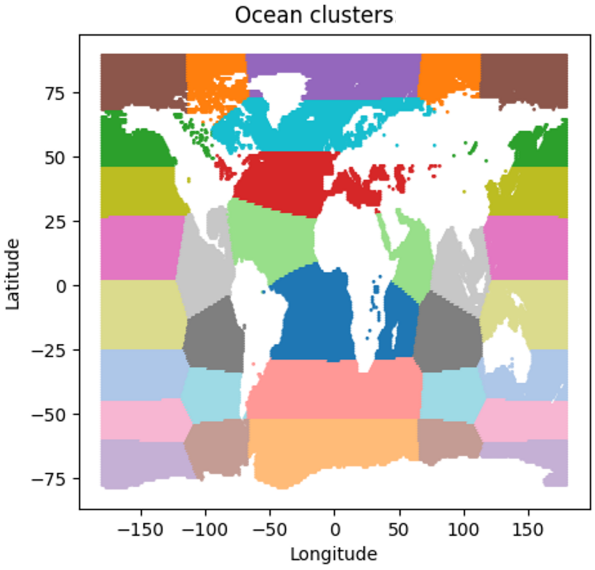

Figure A1. The colours indicate the ocean clusters formed by K-means clustering of each ocean grid-cell position at $1^\circ \times 1^\circ$ grid resolution.

Table A1. Summary of graph nodes and edges.

Component	Count
grid nodes	20 370
ocean nodes	20
mesh nodes	248
total nodes	20 638
grid to mesh edges	20 370
ocean to mesh edges	20
mesh to mesh edges	7440
mesh to grid edges	61 110
total edges	88 940

Appendix A: Graph Construction

Table A1 lists the number of nodes and edges in the hierarchical graph. Figure A1 shows which ocean grid cells belong to the same cluster in the ocean. The clusters are built from the geographical locations of the grid cells using `sklearn.cluster.KMeans` from the scikit learn python library. This approach enables fast clustering of grid cells that lie close together. We chose 20 clusters after preliminary experiments on the training set, and this configuration performed best.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Table B1. Globally averaged RMSE [cm] on validation set (April 2008 – May 2016) per lead time under various design choices. Default values for the base model are 20 ocean clusters, 10 M2M edges, and 3 M2G edges.

	Base model		# ocean clusters			# M2M edges		# M2G edges	
Lead time	default values	No message passing	10	30	50	5	20	1	5
1	1.87	1.92	1.86	1.86	1.88	1.85	1.86	1.86	1.88
2	2.83	2.89	2.80	2.80	2.82	2.79	2.82	2.80	2.82
3	3.30	3.39	3.29	3.27	3.29	3.29	3.31	3.29	3.30
4	3.65	3.77	3.66	3.64	3.63	3.67	3.68	3.66	3.68
5	3.87	3.95	3.89	3.86	3.88	3.88	3.90	3.92	3.90
6	4.02	4.10	4.04	4.04	4.04	4.03	4.07	4.08	4.05

Appendix B: Ablation studies

The hierarchical graph introduced in Sect. 2.1 and the Grid–Mesh–Grid pipeline of Sect. 2.2 together define the model's spatial encoder. We assess their contribution with two complementary ablations on the validation set (April 2008 – May 2016), retraining HiGNN-LSTM under each variant with all other architectural and training settings identical to those described in Sect. 4. First, to test whether the hierarchical graph implementation itself is beneficial, we replace the Grid2Mesh $\longrightarrow$ Mesh2Mesh $\longrightarrow$ Mesh2Grid pipeline with an embedder-only variant in which only the embedder MLP and the output layer are retained, so the latent features handed to the LSTM are produced per grid node without any spatial information exchange. Second, to test how sensitive the model is to the specific connectivity choices, we run a one-at-a-time sweep over the three discrete graph hyperparameters: the number of ocean clusters, mesh-to-mesh (M2M) edges per node at the second graph level, and mesh-to-grid (M2G) edges from the second level back to each grid node. Each is varied in turn while holding the other two at their default values as reported in Sect. 2.1 (20 ocean clusters, 10 M2M edges, and 3 M2G edges). Performance is reported as the globally averaged RMSE per lead time, consistent with the evaluation in Sect. 4.2. Results are summarised in Table B1.

Bypassing the Grid–Mesh–Grid pipeline increases RMSE by 0.05–0.12 cm consistently across all six lead times, confirming that the message-passing pipeline is the principal source of the model's spatial skill. In contrast, varying the number of ocean clusters (10–50), mesh-to-mesh edges (5–20), and mesh-to-grid edges (1–5) changes RMSE by at most 0.05 cm and without a consistent direction across lead times, indicating that the model is robust to the precise connectivity hyperparameters within the ranges explored.

Appendix C: ConvLSTM

The implementation of the Convolutional LSTM is adapted from Palazzi et al. (2017), which is a PyTorch implementation of Shi et al. (2015). We employ a two-layer ConvLSTM with 32 nodes in each layer to generate the results for comparison. The training was stopped after reaching the minimum validation score at 262 epochs. We used AdamW as optimizer with learning rate $= 1e^{-3}$, weight decay$= 1e^{-2}$, and batch size $= 8$ and chose Mean Squared Error (MSE) as the loss function.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Table D1. Major river basins used for basin-wise evaluation. Humid (AI > 0.65), subhumid (0.65>AI >0.5), semiarid (0.5>AI >0.2); AI, aridity index.

ID	Basin name	Continent	Area (km^2)	Climate
1	Mackenzie	North America	1,795,000	Humid
2	Mississippi	North America	3,240,000	Subhumid
3	Amazon	South America	5,913,000	Humid
4	Parana	South America	2,646,000	Humid
5	Nile	Africa	3,057,000	Semiarid
6	Congo	Africa	3,705,000	Humid
7	Orange	Africa	977,000	Semiarid
8	Zambezi	Africa	1,378,000	Subhumid
9	Danube	Europe	795,000	Humid
10	Euphrates	Europe	935,000	Semiarid
11	Volga	Europe	1,404,000	Humid
12	Amur	Asia	2,238,000	Humid
13	Yangtze	Asia	1,924,000	Humid
14	Ganges	Asia	1,584,000	Humid
15	Indus	Asia	864,000	Humid
16	Murray	Australia	1,055,000	Semiarid

Appendix D: River basins for evaluation

Table D1 summarizes in which continent the basins lie, their size, and which type of climate they represent.

Author contributions. VS conceived the study, conducted the analysis, developed the machine learning framework, and wrote the manuscript. FL contributed to the conceptualization of the research question. JK provided advice on the evaluation of the model and the interpretation of the results. XZ contributed to the conceptualization and interpretation of the results.

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

Acknowledgements. The project is funded by the German Federal Ministry for Economic Affairs and Energy under grant number 50EE2201C. The authors are responsible for the content of this publication. Viola Steidl is jointly supported by the Helmholtz Association under the joint research school "Munich School for Data Science – MUDS" and the Munich Center for Machine Learning (MCML). JK and FL are sup-

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

ported by the Deutsche Forschungsgemeinschaft (grant no. SFB 1502/1-2022 – Project No. 450058266). Xiao Xiang Zhu is supported by the Munich Center for Machine Learning and by the Excellence Strategy of the German federal government and the states through the TUM Innovation Network EarthCare. AI tools (Claude and Claude Code) were used to assist in editing portions of the text and programmatically generate and improve figures in this manuscript. All scientific content, experimental design, analysis, and interpretation are the responsibility of the authors. We want to thank Shan Zhao for sharing her knowledge on graph design and graph network implementations.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

References

Ahmed, M., Sultan, M., Elbayoumi, T., and Tissot, P.: Forecasting GRACE Data over the African Watersheds Using Artificial Neural Networks, Remote Sensing, 11, 1769, https://doi.org/10.3390/rs11151769, number: 15, 2019.

Cachay, S. R., Erickson, E., Bucker, A. F. C., Pokropek, E., Potosnak, W., Bire, S., Osei, S., and Lütjens, B.: The World as a Graph: Improving El Niño Forecasts with Graph Neural Networks, https://doi.org/10.48550/arXiv.2104.05089, arXiv:2104.05089 [cs], 2021.

Dufourg, C., Pelletier, C., May, S., and Lefèvre, S.: Forecasting water resources from satellite image time series using a graph-based learning strategy, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVIII-2-2024, 81–88, https://doi.org/10.5194/isprs-archives-XLVIII-2-2024-81-2024, conference Name: ISPRS TC II Mid-term Symposium "The Role of Photogrammetry for a Sustainable World" - 11–14 June 2024, Las Vegas, Nevada, USA, 2024.

Forootan, E., Kusche, J., Loth, I., Schuh, W.-D., Eicker, A., Awange, J., Longuevergne, L., Diekkrüger, B., Schmidt, M., and Shum, C. K.: Multivariate Prediction of Total Water Storage Changes Over West Africa from Multi-Satellite Data, Surveys in Geophysics, 35, 913–940, https://doi.org/10.1007/s10712-014-9292-0, 2014.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 monthly averaged data on single levels from 1940 to present, https://doi.org/10.24381/cds.f17050d7, 2023.

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

Humphrey, V., Rodell, M., and Eicker, A.: Using Satellite-Based Terrestrial Water Storage Data: A Review, Surveys in Geophysics, 44, 1489–1517, https://doi.org/10.1007/s10712-022-09754-9, 2023.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., and Reblitz-Richardson, O.: Captum: A unified and generic model interpretability library for PyTorch, https://doi.org/10.48550/arXiv.2009.07896, arXiv:2009.07896 [cs, stat], 2020.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S., and Battaglia, P.: Learning skillful medium-range global weather forecasting, Science, 382, 1416–1421, https://doi.org/10.1126/science.adi2336, 2023.

Lehner, B. and Grill, G.: Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems, Hydrological Processes, 27, 2171–2186, https://doi.org/10.1002/hyp.9740, data available at https://www.hydrosheds.org, 2013.

Li, F. and Kusche, J.: Observation-Driven Forecast of Global Terrestrial Water Storage and Evaluation for 2010–2024, Water Resources Research, 62, e2025WR041 710, https://doi.org/10.1029/2025WR041710, 2026.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

Li, F., Kusche, J., Chao, N., Wang, Z., and Löcher, A.: Long-Term (1979-Present) Total Water Storage Anomalies Over the Global Land Derived by Reconstructing GRACE Data, Geophysical Research Letters, 48, e2021GL093 492, https://doi.org/10.1029/2021GL093492, 2021.

Li, F., Kusche, J., Sneeuw, N., Siebert, S., Gerdener, H., Wang, Z., Chao, N., Chen, G., and Tian, K.: Forecasting Next Year's Global Land Water Storage Using GRACE Data, Geophysical Research Letters, 51, e2024GL109 101, https://doi.org/10.1029/2024GL109101, 2024.

Li, F., Springer, A., Kusche, J., Gutknecht, B. D., and Ewerdwalbesloh, Y.: Reanalysis and Forecasting of Total Water Storage and Hydrological States by Combining Machine Learning With CLM Model Simulations and GRACE Data Assimilation, Water Resources Research, 61, e2024WR037 926, https://doi.org/10.1029/2024WR037926, 2025.

Loshchilov, I. and Hutter, F.: Decoupled Weight Decay Regularization, https://doi.org/10.48550/arXiv.1711.05101, arXiv:1711.05101 [preprint], 2019.

Löcher, A. and Kusche, J.: A hybrid approach for recovering high-resolution temporal gravity fields from satellite laser ranging, Journal of Geodesy, 95, 6, https://doi.org/10.1007/s00190-020-01460-x, 2020.

Palazzi, A., Huanyu, and Pini, S.: ConvLSTM_pytorch: Implementation of Convolutional LSTM in PyTorch., https://github.com/ndrplz/ConvLSTM_pytorch, last accessed:2026-01-08, 2017.

Palazzoli, I., Ceola, S., and Gentine, P.: GRAiCE: reconstructing terrestrial water storage anomalies with recurrent neural networks, Scientific Data, 12, 146, https://doi.org/10.1038/s41597-025-04403-3, 2025.

Perez, E., Strub, F., Vries, H. d., Dumoulin, V., and Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer, https://doi.org/10.48550/arXiv.1709.07871, arXiv:1709.07871 [cs], 2017.

Save, H.: CSR GRACE RL06 Mascon Solutions, https://doi.org/10.18738/T8/UN91VR, 2019.

Save, H.: CSR GRACE and GRACE-FO RL06 Mascon Solutions v02, https://doi.org/10.15781/cgq9-nh24, 2025.

Save, H., Bettadpur, S., and Tapley, B. D.: High-resolution CSR GRACE RL05 mascons, Journal of Geophysical Research: Solid Earth, 121, 7547–7569, https://doi.org/10.1002/2016JB013007, 2016.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, http://arxiv.org/abs/1506.04214, arXiv:1506.04214 [cs], 2015.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M.: SmoothGrad: removing noise by adding noise, https://doi.org/10.48550/arXiv.1706.03825, arXiv:1706.03825 [cs], 2017.

Steidl, V. and Zhu, X. X.: Hierarchical Graph Networks for Forecasting Terrestrial Water Storage Anomalies, in: Machine Learning and the Physical Sciences Workshop @ NeurIPS 2025, https://ml4physicalsciences.github.io/2025/files/NeurIPS_ML4PS_2025_296.pdf, 2025.

Sun, A. Y., Scanlon, B. R., Save, H., and Rateb, A.: Reconstruction of GRACE Total Water Storage Through Automated Machine Learning, Water Resources Research, 57, e2020WR028 666, https://doi.org/10.1029/2020WR028666, 2021.

Tapley, B. D., Watkins, M. M., Flechtner, F., Reigber, C., Bettadpur, S., Rodell, M., Sasgen, I., Famiglietti, J. S., Landerer, F. W., Chambers, D. P., Reager, J. T., Gardner, A. S., Save, H., Ivins, E. R., Swenson, S. C., Boening, C., Dahle, C., Wiese, D. N., Dobslaw, H., Tamisiea, M. E., and Velicogna, I.: Contributions of GRACE to understanding climate change, Nature Climate Change, 9, 358–369, https://doi.org/10.1038/s41558-019-0456-2, 2019.

Wang, J. and Chen, Y.: The applicability of using NARX neural network to forecast GRACE terrestrial water storage anomalies, Natural Hazards, 110, 1997–2016, https://doi.org/10.1007/s11069-021-05022-y, 2022.

https://doi.org/10.5194/egusphere-2026-2312
Preprint. Discussion started: 23 June 2026

World Meteorological Organization (WMO), Intergovernmental Oceanographic Commission of UNESCO, International Science Council (ISC), United Nations Environment Programme (UNEP), and Copernicus Climate Change Service (C3S): The 2022 GCOS ECVs Requirements, Tech. Rep. GCOS-245, World Meteorological Organization, Geneva, https://library.wmo.int/idurl/4/58111, 2022.

Yin, J., Slater, L. J., Khouakhi, A., Yu, L., Liu, P., Li, F., Pokhrel, Y., and Gentine, P.: GTWS-MLrec: global terrestrial water storage reconstruction by machine learning from 1940 to present, Earth System Science Data, 15, 5597–5615, https://doi.org/10.5194/essd-15-5597-2023, 2023.

Yu, Q., Wang, S., He, H., Yang, K., Ma, L., and Li, J.: Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model, International Journal of Applied Earth Observation and Geoinformation, 102, 102 404, https://doi.org/10.1016/j.jag.2021.102404, 2021.