

---

# Hierarchical Graph Networks for Forecasting Terrestrial Water Storage Anomalies

---

**Viola Steidl**

\*Technical University of Munich,  
Munich Center for Machine Learning (MCML)  
Germany  
viola.steidl@tum.de

**Xiao Xiang Zhu**

Technical University of Munich,  
Munich Center for Machine Learning (MCML)  
Germany

## Abstract

The availability of fresh water is vital to the ecosystem and communities. In a changing climate, the increased risk of droughts makes it more important to have an accurate view of changes in terrestrial water storage (TWS). Predicting changes in TWS is inherently difficult since it integrates the changes of all water compartments, with underlying processes that operate on vastly different temporal and spatial scales. In our work, we explore a novel design of a hierarchical graph using domain knowledge of hydrological basins to encode these processes in a latent feature sequence using an encoder-processor-decoder style graph neural network. The subsequent recurrent neural network then forecasts changes in TWS from the latent feature sequence and historical TWS evolution for up to six months ahead. The gridded product of the seasonal forecast of global TWS evolution shows short-term improvement over a seasonal long-term mean.

## 1 Introduction

The total amount of water stored on land is called the terrestrial water storage (TWS) [6]. TWS is an essential climate variable for land hydrology, but can not directly be measured [14]. Since 2002, the Gravity Recovery and Climate Experiment (GRACE) and GRACE Follow-On (GRACE-FO) satellites have delivered measurements of changes in the Earth’s gravity field. From these, terrestrial water storage anomalies (TWSA) with respect to a long-term mean can be derived [6].

In recent years, notable effort has gone into extending the time series of GRACE-like TWSA to past decades when GRACE satellite data was not available (e.g., [5, 20, 18, 10, 19, 15]). These reconstructions can be used in hydrological modelling and benchmarking, for sea level budget studies or long-term assessment of changes in the frequency of droughts [5]. Conversely, forecasting GRACE-derived TWSA globally has received less attention, although Li et al. [12] has shown that assimilating forecasted GRACE-derived TWSA can increase the capacity of physics-based land surface models. In their work, Li et al. [11] determine lag correlations and teleconnections between TWSA and climate variables to hand-craft features. Then, they train separate models for each time step in the forecasting horizon using the matching subset of climate variables.

---

\*Chair of Data Science in Earth Observation

Graph-based learning methods are capable of learning these large-scale interactions for long timescales from data, if the graph is set up sensibly [1]. In this work, we build a hierarchical graph that represents the world on two spatial scales to incentivize our graph neural network (GNN) to resolve processes that act on different spatial and temporal scales. The GNN automatically generates a sequence of latent features and therefore takes on the task of hand-crafting input features for a subsequent recurrent neural network.

To train this model, we build a new dataset consisting of TWSA reconstruction and ERA5 climate variables. Analysis of the latent features shows that the model is able to pick up time-step-specific information.

## 2 Dataset

We use a reconstruction of GRACE-like TWSA by Li et al. [10] that extends the time series from 1979 until 2020 at a monthly  $0.5^\circ \times 0.5^\circ$  resolution. Li et al. [10] created their data product combining machine learning and statistical decomposition techniques. They extensively evaluated their product against TWSA data obtained from other satellites for the time period before the GRACE missions. The data product also comes in a detrended version, meaning long-term trends, which are difficult to model, have been removed. Additionally, we collect 11 variables, shown in Table 1, from ERA5 monthly averaged single levels [4]. Also, we include a trigonometric embedding of the month of the year as features to give the model a better perception of the seasonality.

All inputs are sampled to a  $1^\circ \times 1^\circ$  resolution, as this is the resolution of GRACE-derived TWSA products. The dataset is available at <https://doi.org/10.5281/zenodo.17340261>.

Table 1: Input features.

TWSA
Total precipitation
Surface pressure
2m temperature
10m wind speed
Soil moisture level 1 & 4
Evaporation
Potential evaporation
Runoff
Sea surface temperature
Land sea mask
Month of the year

## 3 Methodology

### 3.1 Graph construction

The model should be able to uncover long-range teleconnections or lag correlations between different parts of the world. Since these correlations operate on various spatial scales, we build a hierarchical graph of two levels: A high-resolution grid level and a coarser mesh level.

The first level of the graph consists of nodes that represent every grid cell on land. Their node features are the time series of the variables described in 2. Additionally, we generate 20 clusters from the grid cells of the ocean to reduce the size of the first graph level. Ocean grid cells are clustered by mean and variance of their sea surface temperature, and their geographic position. Their features are the averaged features of the ocean grid cells that belong to them. We call the nodes of the first graph level *grid nodes* in the following. At a resolution of  $1^\circ \times 1^\circ$ , we have 20953 grid nodes, representing the grid cells on land and the 20 ocean clusters.

The second level of the graph consists of nodes that represent the hydrological basins as they were defined in [9] and nodes that represent the 20 ocean clusters. We call these nodes the *mesh nodes*. The second-level node features are the sine and cosine embedded spatial coordinates of the nodes.

Three sets of edges connect the nodes of the first level and the nodes of the second level of the graph. The first set connects grid nodes to the nodes of the hydrological basin or ocean cluster that they belong to. Another set of edges connects the mesh nodes of the second graph level to their ten nearest neighbours, depending on spatial distance. The mesh nodes representing the ocean clusters connect to all basin nodes, as we assume that the ocean has a long-ranging influence on the TWSA. A final set of edges connects mesh nodes of the second graph level back to the grid nodes. Each grid node is connected to three nearest neighbours.

### 3.2 Grid–Mesh–Grid spatial encoding

We propose a Grid2Mesh  $\rightarrow$  Mesh2Mesh  $\rightarrow$  Mesh2Grid pipeline that bridges the gap between regular gridded Earth observation products and sparse graph-based representations. This design allows us to

leverage the flexibility of mesh structures for spatial dependency learning while remaining compatible with gridded TWSA data. The GNNs in the pipeline rely on simple Multi-Layer Perceptrons (MLPs), as proposed in GraphCast [8]. The GNN implementations are adapted from [2].

At first, an embedder maps the node and edge features into a common dimension ( $\text{dim} = 128$ ). From there, another MLP-based module (Grid2Mesh) transfers the information from the grid nodes to the mesh nodes along their edges by updating the destination node features and edge attributes.

In the processing step, the information is passed among the connected mesh nodes with  $n = 2$  layers of Mesh2Mesh GNNs in the same way as in the Grid2Mesh GNN. As already mentioned, we generate a separate encoding for every time step of the forecast. Therefore, the current timestep is encoded with Feature-wise Linear Modulation [16] in the Mesh2Mesh processing GNNs.

In the decoding step, the Mesh2Grid GNN passes the information back to the grid node features. An output layer fuses the spatio-temporal information into a 4-dimensional sequence of latent features as input for the temporal block, the LSTM (2 layers with 64 neurons). Residual connections for the update of the grid node, mesh node, and edge features stabilize training.

### 3.3 Spatio-temporal forecasting

The decoded latent feature sequences are integrated into an LSTM-based autoregressive forecaster together with the sequence of TWSA. For every timestep of the forecasting horizon, the GNNs create a new latent feature sequence. In this way, the GNNs are trained to create a time-step-specific latent feature sequence. The LSTM layers then output the change in TWSA for the next month. Figure 1 shows the complete model architecture.

## 4 Results

We split the dataset into training (July 1979 - February 2004), validation (March 2004 - April 2012), and test (May 2012 - June 2020) time series. The model is trained with Adam [7] (learning rate =  $1e^{-3}$ , batch size = 4) for 100 epochs. All experiments were conducted on a single GPU (NVIDIA A40).<sup>2</sup>

The model achieves a root mean squared error (RMSE) of 14.25 cm on the test set, averaged over all forecasting sequences and time steps in the forecast. Table 2 shows that the error rises abruptly when forecasting the second month. After the second month, the RMSE rises steadily but with smaller increases. We compare these results to the RMSE scores of a ConvLSTM [17]

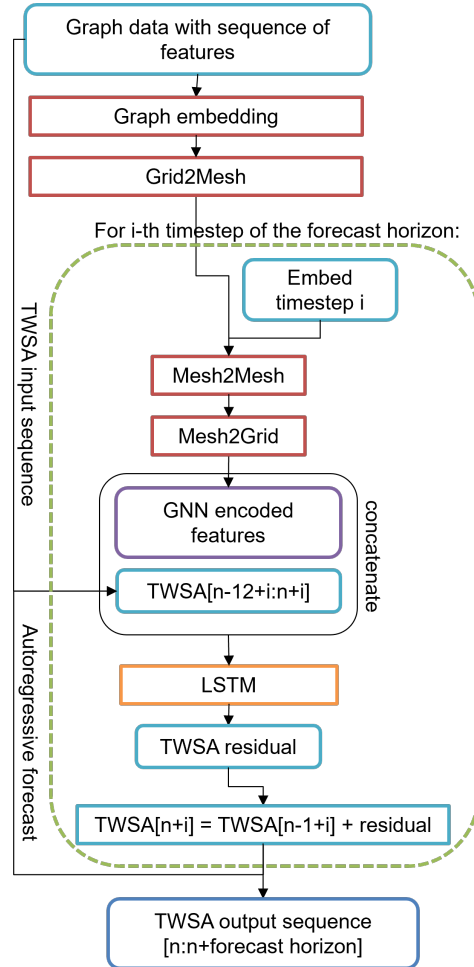


Figure 1: Scheme of the model architecture.

Table 2: RMSE [cm] mean and (std) over all forecasting periods in the test set.

Month	Ours	ConvLSTM	Climatology
1	7.95 (0.98)	8.24 (0.92)	15.72 (2.74)
2	12.03 (1.50)	12.67 (1.50)	15.69 (2.73)
3	14.33(1.95)	14.88 (1.81)	15.69 (2.73)
4	15.99 (2.09)	16.30 (2.03)	15.68 (2.73)
5	17.19 (2.31)	17.37 (2.15)	15.69 (2.73)
6	18.01 (2.49)	18.02 (2.23)	15.72 (2.73)
Mean	14.25	14.85	15.70

<sup>2</sup>Code available at [https://github.com/violal1593/TWSA\\_HiGNN](https://github.com/violal1593/TWSA_HiGNN).

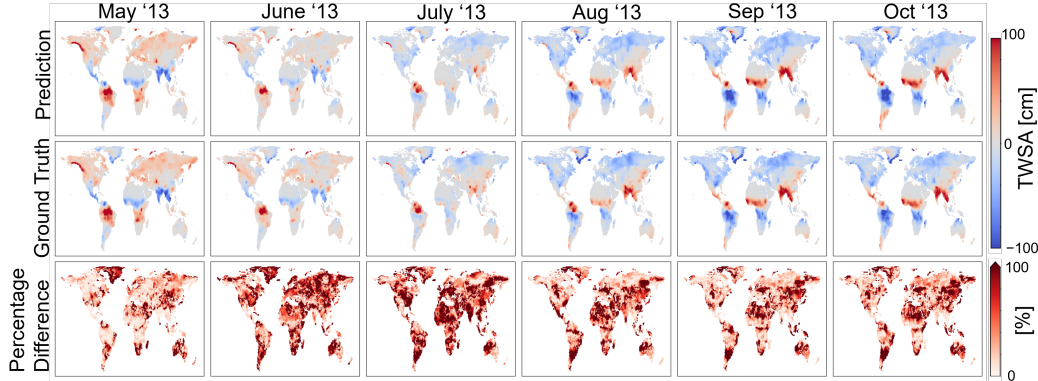


Figure 2: Forecast of the first sequence in the test dataset (May 2013 - October 2013).

network tested on the same time series and RMSE scores of a long-term seasonal mean (Climatology) derived from the training TWSA time series. Our model constantly achieves better scores than the ConvLSTM and outperforms climatology in the first three months. After that, climatology has lower RMSE scores, which shows that our model is not yet able to fully capture long-term dynamics. Figure 2 shows the six-month forecast for the first sequence in the test set. The spatial patterns are vastly matched, but the magnitude is not. Naturally, the percentage error amplifies the perception of errors in areas where only small changes in TWS occur.

#### 4.1 Basin-wide evaluation

We selected 15 basins to examine the spatial variability of the model performance. Table 3 shows that RMSE and Pearson’s correlation can vary greatly between basins. The correlation is measured along the temporal dimension, so it indicates how well the predictions match the dynamics of the target. However, it does not measure the accuracy of the predictions’ magnitude like the RMSE does. Therefore, pronounced (seasonal) signals that are easy to match for the model will have high correlations but might still have high RMSE scores, as for example in the Amazon basin. Conversely, small RMSE scores might be misleading for river basins where the TWSA signal is closer to 0. For example, in the Amur basin, the model prediction closely matches the ground truth if we are only considering RMSE. The correlation reveals that the model has difficulties capturing the dynamics of the basin.

Table 3: Mean (std) performance across test hydrological basins.

Basin	RMSE [cm] #	Correlation "
Amazon	13.40 (10.2)	0.986 (0.04)
Parana	10.17 (4.9)	0.931 (0.11)
Mississippi	5.48 (2.1)	0.964 (0.06)
Mackenzie	2.27 (1.1)	0.990 (0.01)
Danube	8.68 (4.4)	0.939 (0.09)
Nile	5.12 (2.9)	0.962 (0.08)
Congo	8.77 (3.9)	0.822 (0.34)
Orange	2.68 (1.8)	0.741 (0.36)
Zambezi	11.45 (7.2)	0.982 (0.03)
Euphrates	5.64 (3.6)	0.976 (0.03)
Volga	6.85 (3.8)	0.977 (0.03)
Yangtze	5.28 (2.4)	0.880 (0.16)
Amur	4.39 (2.4)	0.535 (0.47)
Ganges	8.11 (3.7)	0.977 (0.07)
Indus	5.39 (2.4)	0.941 (0.07)
Murray	6.52 (3.4)	0.744 (0.35)

## 5 Discussion and future work

**GNN encoding** The GNN block encodes all the features for every time step and grid node in the input sequence into four latent feature sequences. We analyze these latent feature sequences by letting the GNN predict on a dummy dataset where all input features are equal to 1. To calculate the kernel density estimate (KDE) over all grid nodes, we average over the length of the latent feature time series (twelve months) that the model creates. Figure 3 shows the KDE for the 4 latent features. The model seemingly is able to encode temporal progression in the latent features, where months 1 and 6 are well separated (e.g., Figure 3 Latent Feature 2). Latent feature sequences for months 2-5 still

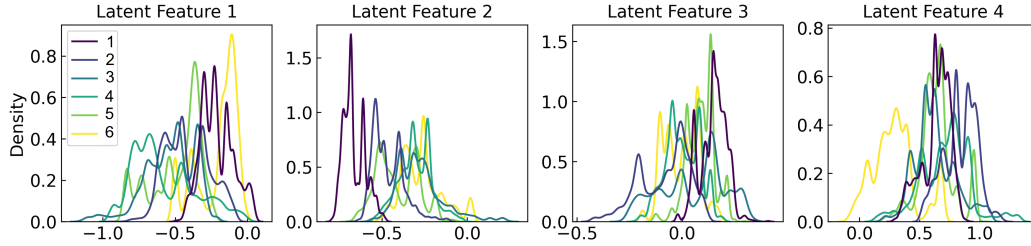


Figure 3: KDE plots of the GNN encoded latent features for months 1-6 in the forecasting horizon.

show some overlap, so future work should focus on enhancing the separability of mid-horizon latent features.

**Spatial variability** Spatially examining the model predictions reveals high variability in predictive performance (see Figure 2). Generating a graph with extended domain knowledge on the basin or ocean level could help to better account for the hydrological conditions in different areas. One option could be to generate sparser graph edges not only based on spatial distance but also based on spatio-temporal correlations between TWSA and features. Another could be to change the clustering of the ocean nodes such that it better represents known patterns that influence TWSA.

**Dataset** Our model relies on ERA5 variables, which are not direct hydrological observations but the output of a land-atmosphere model constrained by data assimilation. Key drivers of TWSA, such as precipitation, evapotranspiration, runoff, and soil moisture, may therefore contain systematic biases, particularly in dry regions where observations are sparse and model physics dominate [3]. This likely constrains the skill of forecasts at longer horizons. Additionally, the model trains on a reconstruction of GRACE-derived TWSA. The dataset itself has been tested extensively by its authors [10]. However, to properly assess the performance of our graph-based TWSA forecast, evaluating the predictions on the GRACE-derived TWSA is still needed.

## 6 Conclusion

Forecasting TWSA is inherently difficult because it reflects the combined dynamics of multiple water compartments with different temporal and spatial scales. In this work, we take a step toward tackling this challenge by representing the Earth as a hierarchical graph to fuse spatial and temporal information into a sequence of latent features. These time-step-specific latent features enable global, seasonal forecasting of TWSA. While the model can capture TWSA patterns and even outperform the long-term seasonal mean, its accuracy declines with longer horizons, and its performance remains spatially variable. Future progress will likely come from incorporating more domain knowledge into the construction of the graph and the mesh latent space to strengthen the informative power of the latent feature sequence.

## Acknowledgments and Disclosure of Funding

The project is funded by the German Federal Ministry for Economic Affairs and Energy under grant number 50EE2201C. The author is responsible for the content of this publication. The authors are 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). We thank Jürgen Kusche and Fupeng Li for supporting us with their hydrological expertise. We also want to thank Shan Zhao for sharing her knowledge on graph design and graph network implementations.

## References

- [1] S. R. Cachay, E. Erickson, A. F. C. Bucker, E. Pokropek, W. Potosnak, S. Bire, S. Osei, and B. Lütjens. The World as a Graph: Improving El Niño Forecasts with Graph Neural Networks, 2021. URL <http://arxiv.org/abs/2104.05089>. arXiv:2104.05089 [cs].

- [2] C. Dufourg, C. Pelletier, S. May, and S. Lefèvre. 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, 2024. ISSN 1682-1750. doi: 10.5194/isprs-archives-XLVIII-2-2024-81-2024. URL <https://iisprs-archives.copernicus.org/articless/XLVIII-2-2024/81-2024/isprs-archives-XLVIII-2-2024-81-2024.html>. Conference Name: ISPRS TC II Mid-term Symposium “The Role of Photogrammetry for a Sustainable World” - 11–14 June 2024, Las Vegas, Nevada, USA Publisher: Copernicus GmbH.
- [3] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730): 1999–2049, 2020. ISSN 1477-870X. doi: 10.1002/qj.3803. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803>.
- [4] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J.-N. Thépaut. ERA5 monthly averaged data on single levels from 1940 to present, 2023. URL <https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels-monthly-means?tab=overview>.
- [5] V. Humphrey and L. Gudmundsson. GRACE-REC: a reconstruction of climate-driven water storage changes over the last century. *Earth System Science Data*, 11(3):1153–1170, 2019. ISSN 1866-3508. doi: 10.5194/essd-11-1153-2019. URL <https://essd.copernicus.org/articless/11/1153/2019/>. Publisher: Copernicus GmbH.
- [6] V. Humphrey, M. Rodell, and A. Eicker. Using Satellite-Based Terrestrial Water Storage Data: A Review. *Surveys in Geophysics*, 44(5):1489–1517, 2023. ISSN 1573-0956. doi: 10.1007/s10712-022-09754-9. URL <https://doi.org/10.1007/s10712-022-09754-9>.
- [7] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, 2017. URL <http://arxiv.org/abs/1412.6980>. arXiv:1412.6980 [cs].
- [8] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, A. Merose, S. Hoyer, G. Holland, O. Vinyals, J. Stott, A. Pritzel, S. Mohamed, and P. Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, Dec. 2023. doi: 10.1126/science.adi2336. URL <https://www.science.org/doi/10.1126/science.adi2336>. Publisher: American Association for the Advancement of Science.
- [9] B. Lehner and G. Grill. 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, 2013. ISSN 1099-1085. doi: 10.1002/hyp.9740. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.9740>.
- [10] F. Li, J. Kusche, N. Chao, Z. Wang, and A. Löcher. Long-Term (1979-Present) Total Water Storage Anomalies Over the Global Land Derived by Reconstructing GRACE Data. *Geophysical Research Letters*, 48(8):e2021GL093492, 2021. ISSN 1944-8007. doi: 10.1029/2021GL093492. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2021GL093492>.
- [11] F. Li, J. Kusche, N. Sneeuw, S. Siebert, H. Gerdener, Z. Wang, N. Chao, G. Chen, and K. Tian. Forecasting Next Year’s Global Land Water Storage Using GRACE Data. *Geophysical Research Letters*, 51(17): e2024GL109101, 2024. ISSN 1944-8007. doi: 10.1029/2024GL109101. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2024GL109101>.
- [12] F. Li, A. Springer, J. Kusche, B. D. Gutknecht, and Y. Ewerdwalbesloh. 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(2):e2024WR037926, 2025. ISSN 1944-7973. doi: 10.1029/2024WR037926. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2024WR037926>.
- [13] J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. ISSN 0022-1694. doi: 10.1016/0022-1694(70)90255-6. URL <https://www.sciencedirect.com/science/article/pii/0022169470902556>.
- [14] W. M. Organization, I. O. C. of the United Nations Educational, I. S. C. (ISC), U. N. E. P. (UNEP), C. C. C. S. (C3S), and S. a. C. O. (IOC-UNESCO). The 2022 GCOS ECVs Requirements - Updated in 2025. Technical Report 245, Geneva, 2025. URL <https://library.wmo.int/records/item/58111-the-2022-gcos-ecvs-requirements>.

- [15] I. Palazzoli, S. Ceola, and P. Gentine. GRAiCE: reconstructing terrestrial water storage anomalies with recurrent neural networks. *Scientific Data*, 12(1):146, 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-04403-3. URL <https://www.nature.com/articles/s41597-025-04403-3>. Publisher: Nature Publishing Group.
- [16] E. Perez, F. Strub, H. d. Vries, V. Dumoulin, and A. Courville. FiLM: Visual Reasoning with a General Conditioning Layer, 2017. URL <http://arxiv.org/abs/1709.07871>. arXiv:1709.07871 [cs].
- [17] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Sept. 2015. URL <http://arxiv.org/abs/1506.04214>. arXiv:1506.04214 [cs].
- [18] A. Y. Sun, B. R. Scanlon, H. Save, and A. Rateb. Reconstruction of GRACE Total Water Storage Through Automated Machine Learning. *Water Resources Research*, 57(2):e2020WR028666, 2021. ISSN 1944-7973. doi: 10.1029/2020WR028666. URL <https://onlinelibrary.wiley.com/doi/abs/10.1029/2020WR028666>.
- [19] J. Yin, L. J. Slater, A. Khouakhi, L. Yu, P. Liu, F. Li, Y. Pokhrel, and P. Gentine. GTWS-MLrec: global terrestrial water storage reconstruction by machine learning from 1940 to present. *Earth System Science Data*, 15(12):5597–5615, 2023. ISSN 1866-3508. doi: 10.5194/essd-15-5597-2023. URL <https://essd.copernicus.org/articles/15/5597/2023/>. Publisher: Copernicus GmbH.
- [20] Q. Yu, S. Wang, H. He, K. Yang, L. Ma, and J. Li. Reconstructing GRACE-like TWS anomalies for the Canadian landmass using deep learning and land surface model. *International Journal of Applied Earth Observation and Geoinformation*, 102:102404, 2021. ISSN 15698432. doi: 10.1016/j.jag.2021.102404. URL <https://linkinghub.elsevier.com/retrieve/pii/S0303243421001112>.

Table 4: Graph construction summary.

Component	Count
grid nodes	20 953
mesh nodes	313
grid to mesh edges	20 953
mesh to mesh edges	9390
mesh to grid edges	62 859

## A Supplementary Material

### A.1 Graph Construction

Table 4 lists the exact number of nodes and edges in the hierarchical graph.

### A.2 Spatial evaluation

To assess the spatial performance over time, we calculate the Nash-Sutcliffe Efficiency (NSE) per pixel for the first test sequence. NSE measures the model’s predictive skill relative to the mean of observations and is defined as

$$\text{NSE}(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{N-1} (y_i - \hat{y}_i)^2}{\sum_{i=0}^{N-1} (y_i - \bar{y})^2} \quad (1)$$

with  $\hat{y}$  being the predicted quantity and  $y$  being the observed mean [13]. The optimal value is 1.0, and negative values indicate that the model predictions are worse than the observed mean.

Figure A.2 shows that the model performance has a high spatial variability. For example, for almost all of Greenland, the performance is worse than predicting the mean, while the forecast in vast parts of North America reaches high NSE scores.

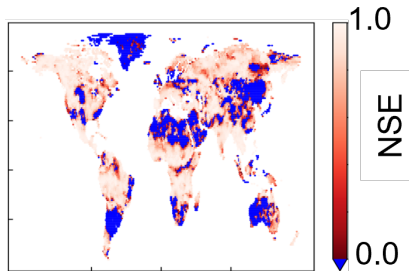


Figure 4: NSE for the sequence from May 2013 to October 2013.

### A.3 GNN latent features

We analyze latent feature sequences by letting the GNN predict on a dummy dataset where all features are equal to 1. Figure A.3 shows latent feature 1 for every time step in the forecasting sequence. There are differences in the magnitudes and patterns of the feature depending on the time step in the forecasting horizon. Therefore, we assume that the model not only learns that for different time steps it needs to create different latent features, but also that for different time steps it should put more or less weight on input from a specific region.

