Dear Reviewer #1,

Thank you very much for your thorough review and questioning of our manuscript. We are grateful for your constructive discussion, which helped us significantly improve the manuscript.

The results changed substantially during the revision, partly because we found and fixed several errors in the data preprocessing pipeline. Partly because, motivated by the reviewers' requests to focus on TWSA anomalies, the model now predicts TWSA anomalies relative to the training climatology, and climatology is added back afterwards to retrieve full TWSA. 'Anomaly' in this response letter therefore refers to non-seasonal TWSA, where climatology has been subtracted.

Some of the conclusions in the original manuscript have changed:

1. The model generates useful predictors from ERA5 data: withdrawn. Removing all ERA5 inputs has no significant impact at any lead, and has an ambiguous impact in the per-basin evaluation. Skill comes mainly from TWSA inputs.

2. The model encodes spatial teleconnections: withdrawn. Removing ocean nodes changes the skill significantly only at lead six and only for 11 out of 226 basins, which is what chance alone would give.

3. Graph connections are better than local convolutions: restricted. HiGNN-LSTM outperforms the ConvLSTM on the reconstruction set but not on GRACE/-FO.

4. Strong performance where seasonal signal dominates: altered. Predicting anomalies and evaluating anomaly correlations show the highest scores where anomalies persist the longest.

5. HiGNN-LSTM outperforms climatology baseline: strengthened. The model is now evaluated against persistence and climatology baselines and beats both in area-weighted global RMSE at every lead, on the reconstruction test set and the GRACE/-FO.

At the end of this document, we have attached a supplement that lists the code changes, the revised results and the new results from the experiments requested by the reviewers. In the answers to the comments, we will refer to the Supplement. The original comments are in *blue italics*, our answers are in black.

We hope to address all comments to your satisfaction.

Best regards

Viola Steidl

On behalf of all co-authors

I appreciate the substantial technical effort involved in developing and implementing this global hierarchical graph model. However, I have more fundamental concerns about the study's hydrological framing. The manuscript is primarily organized around a machine-learning architecture and then applies that architecture to TWSA, but it does not yet demonstrate that the resulting framework addresses a clearly formulated hydrological question. In particular, the central interpretation that the graph learns useful hydrological teleconnections is not supported by the current experiments. At present, the manuscript demonstrates that the proposed network can reproduce a reconstructed TWSA product more clearly than it demonstrates a new hydrological understanding or a useful advance in seasonal hydrological prediction. I therefore consider major revision necessary, and I am uncertain whether the required changes can be achieved without a substantial redesign of the study.

We agree that our study was organised around the machine learning architecture. In our work, we investigate whether a graph-based architecture is able to extract and propagate hydrological information from local water storage, climate, and hydrological variables, and distant ocean regions.

We now test more directly where the model's skill comes from. We included an ablation study that removes key parts of the model's architecture (message passing between graph nodes and ocean nodes) and all ERA5 inputs, and retrains the model. Additionally, we replaced the latent attribution analysis with a latent-removal experiment that replaces feature information with the training-set mean and measures the performance loss.

All experiments are now run with 8 seeds to separate signal from seed noise, the model predicts anomalies relative to the training climatology, and we compare it to persistence and climatology on the reconstruction and on GRACE/-FO observations, using anomaly correlation and anomaly NSE as additional metrics.

The results lead to different conclusions than our original manuscript claimed, as summarised above: 1) Storage memory is the main source of skill and the model beats anomaly persistence and climatology in global RMSE at every lead; 2) Message passing reduces global RMSE significantly from lead 2 onward, although the effect is small 34 out of 64 basins with significant effect improve their RMSE without message passing; 3) Removing ERA5 does not change global RMSE significantly, but changes are significant in 99 basins, with 52 basins getting better RMSE scores; 4) Ocean nodes only have significant effect at lead 6, and do not significantly change results per basin.

We think these results, including the negative ones, are relevant for the hydrological community. They show which information our graph-based model is actually able to extract and that teleconnections are not automatically learned during the training. We revised the introduction, the hypotheses and the conclusions accordingly.

1) The current interpretability analysis does not yet establish which hydrological information the model has learned. The attribution analysis only shows that the output is sensitive to the GNN latent features, but it does not identify where the predictive information comes from. TWSA is primarily controlled by local water-balance processes and storage memory, and remote climate connections may provide additional information in specific regions and seasons. I suggest adding process-oriented attribution experiments and determining whether the learned remote connections correspond to known hydroclimatic relationships.

We agree that the original interpretability analysis (SmoothGrad on the GNN latent features) can not show where the predictive information comes from. Therefore, we replaced it with a process-oriented ablation study where we retrain the model while systematically removing information and measure the change in forecast skill. We use 8 seeds and evaluate on the test period, since it was held out from training and hyperparameter tuning, while the validation dataset was used to develop the model. All p-values come from a Welch t-test over the seeds. They show whether a difference exceeds seed noise on the test dataset, not how much it depends on the test dataset. The results are shown in Table 1 and Figure 1 of the Supplement.

1. local water-balance processes only: We remove the hierarchical graph structure and just keep the grid-level embedding to create the four latent feature sequences per grid cell. No messages are passed between grid nodes or basin/ocean nodes, and only local information is forwarded to the LSTM module. Globally averaged, this increases RMSE scores significantly after lead 1, indicating that,

for longer leads, the graph adds skill. Evaluating per basin and averaging across all leads shows opposite effects in individual basins, yielding a small overall gain in message passing.

2. water storage memory only: We remove all ERA5 inputs. On the global average, RMSE scores increase but not significantly for any lead. Averaging over leads and evaluating per basin also changes RMSE in both directions, so the effects cancel out in the global average. This shows that the model's performance mostly comes from the TWSA input and gains only minimal benefit from ERA5 variables.

3. No oceanic influence: We remove all ocean nodes from the graph. The results show that while the RMSE scores increase, the increase is significant only for lead 6. It remains unclear if that is because SST correlation to TWSA is more prominent at later leads (Li et al. 2024), but we remove the claim that the model has learned teleconnections. Averaged across all leads and evaluated per basin, the results show that removing ocean nodes significantly impacts only 11/226 basins, as would be expected by chance.

2) The practical hydrological value of forecasting total TWSA over a six-month horizon needs to be clarified. A technically accurate prediction of this integrated signal does not necessarily translate into useful forecasts of drought, groundwater deficit, reservoir conditions, or other hydrological impacts. I suggest that the authors define the intended hydrological application and evaluate metrics relevant to that application.

Thank you for pointing this out. We agree that we did not clearly define an application of the forecast. We see the main application in bridging the latency period (2-3 months) before standard GRACE TWSA products become available after processing the raw GRACE measurements.

Additionally, GRACE TWSA can be applied in Early Warning Systems, such as the NASA Hydrological Forecast System for food security or groundwater forecasting (Arsenault et al., 2020; Getirana et al., 2020). We added this in the introduction.

For both applications, RMSE is a relevant metric because it measures the difference between the target and predicted TWSA. We extended our evaluation to include a persistence baseline, as this would be the best alternative to bridge a latency of 2-3 months. Climatology becomes the more relevant baseline at longer leads. The model beats both baselines on the test dataset and the GRACE/-FO dataset (see Supplement Table 3 and Table 4).

3) More importantly, because the seasonal cycle is retained in the target and the month of the year is explicitly provided as an input, the reported RMSE and correlation can be strongly influenced by the model's ability to reproduce seasonality. For seasonal hydrological prediction, skill in the anomaly component is usually more informative than skill in the full signal. I recommend evaluating deseasonalized TWSA anomalies directly. The results should be reported separately for each lead time and basin.

Thank you for this remark. We agree that the correlation will be dominated by the seasonal cycle, which the model predicted well. As we switched to predicting anomalies relative to the training climatology, we added NSE and anomaly correlation as metrics.

Additionally, we included the per-basin, per-lead scores in the Appendix and discuss results in the main text:

Across the 226 basins, the median anomaly correlation falls from 0.85 at lead 1 to 0.10 at lead 6 (persistence: 0.83 to 0.08). The model beats climatology in 222 basins at lead 1 but only 106 at lead 6, and it beats persistence in 193 and 219 basins, respectively (see Supplement Table 2). In the Amazon, anomaly correlation falls from 0.90 to 0.34 and skill against climatology (1 − RMSE/RMSE_clim) from 0.60 to 0.08 (see Supplement Table 6).

4) The relationship between the predictor data and the reconstructed TWSA target requires a more careful discussion. The Li et al. (2021) reconstruction was itself produced using climatic and hydrological predictors and machine-learning methods, but the present model uses ERA5 climatic and hydrological variables to predict that reconstruction. Good performance on the reconstructed product may partly reflect HiGNN-LSTM's ability to reproduce relationships already built into that product. Ideally, the author may wish to repeat the analysis with more than one reconstruction product or provide a stronger observation-based evaluation.

We agree that overlap among predictors could make skill on the reconstruction partly circular. Li et al. (2021) built the reconstruction using precipitation, temperature, SST, soil moisture, evaporation, runoff, and climate indices, and our ERA5 inputs cover most of these variables. However, two things limit the risk of circularity here:

1) The model never receives ERA5 inputs for the months it forecasts: its input ends the month before the first forecast month, so it cannot learn a direct mapping from target-month climate to target-month TWSA, as it would be done in the reconstruction.

2) The climate and hydrological predictors come from different sources.

Lastly, the added ablation study shows that removing all ERA5 inputs globally increases the RMSE on the test period by only 0.001-0.010 cm, not significant at any lead.

So at least the globally averaged performance on the reconstructed product does not reflect the model's ability to reproduce the relationships. Nevertheless, we added this concern to the limitations in Section 5.3, as it could still affect individual basins. Since other reconstruction datasets are also based on the variables we chose to include in our model, we don't think repeating the analysis on a different reconstruction would resolve the concern. However, we agree that evaluation on GRACE/-FO is strictly necessary: HiGNN-LSTM scores are lower on the GRACE/-FO dataset, evaluated per lead and per basin for the 16 test basins. However, on the global average, HiGNN-LSTM still outperforms a persistence baseline and the two climatology baselines (derived from the training dataset and GRACE-derived TWSA) for every lead.

5) Right now, it seems every ocean cluster is connected to every basin. I cannot immediately understand this strong structural assumption of universal ocean-to-basin connectivity. I suggest testing graph variants without ocean nodes, with only local basin connections, and with river-network connections. Otherwise, I don't think improved performance can be interpreted as evidence that the model has identified physically meaningful teleconnections.

At the moment, the graph is set up using Level 3 basins, so all basins drain only into the ocean or an inland sink, and their river-network connections are confined within the basins.

The ocean-to-basin connectivity was motivated by the correlation between sea surface temperature (SST) and TWSA. Li et al. (2024) found SST to be one of the variables with high correlation to TWSA. Because SST measurements are located far from TWSA grid cells, the graph design should link ocean

cells to the hydrological basins. Ocean nodes contain not only SST but also other input variables available for ocean grid cells (surface pressure, 2 m temperature, 10 m wind speed, total precipitation, evaporation).

As mentioned above, we have now included an ablation experiment testing a graph variant without any ocean nodes. The results show that ocean nodes do not add significant skill to the model's performance except for lead 6. An open question remains if this is due to SST only playing a more significant role at later lead times or because the information was not sufficiently propagated through the graph network. We added this in the Results section for the process-based ablation study.

Specific comments:

L24: However, these effects are not explicitly represented in the model inputs or graph structure, and their treatment in the reconstructed TWSA target remains unclear. The authors should clarify the extent to which the reconstruction preserves signals of human interventions. They should also discuss how detrending may remove persistent anthropogenic storage changes and how this affects transferability to GRACE observations and performance in strongly managed basins.

Thank you for this comment. The reconstruction dataset we work with comes from a model trained to fit GRACE-derived TWSA. It does not explicitly include human interventions, but because the reconstruction pipeline is trained to fit GRACE TWSA closely, it should implicitly preserve the signals of human interventions. We add this to the description of the dataset in Section 3.1 as 'As the set of predictors for the TWSA reconstruction consists of climatic and hydrological variables only, human intervention is not explicitly accounted for. However, since the models are trained to fit GRACE observations, the reconstruction should implicitly include effects from anthropogenic water use.'

Additionally, in Section 3.1 we explain what using the detrended version of the reconstruction dataset by Li et al. means: 'Without the linear trend, long-term storage changes from ice sheet and glacier mass loss or anthropogenic activities, as well as part of the multi-year variability between wet and dry periods, are not represented in the data (Rodell et al., 2018). Our target, therefore, contains only the variability around these trends, and a model trained on this data can not forecast persistent storage loss.'

L60: Please specify which parts are hydrologically informed. The basin aggregation might be hydrologically motivated, but the nearest-neighbor and universal ocean-to-basin connections appear primarily geometric. Not sure of the value of basin aggregation here.

Thank you for pointing that out. 'Hydrological domain knowledge' here refers purely to knowledge about hydrological basins. We extended the sentence to make that more obvious.

It now reads: 'This approach ensures that the computation remains tractable while incorporating knowledge of hydrological belonging of grid cells.'

The ablation experiments show that the nearest neighbour connections between basins, together with the grid-mesh-grid message passing, have a small but significant positive effect on the model's scores. The ocean-to-basin connections, however, do not add significant skill for leads 1-5.

L69: For a hydrology journal, the manuscript should formulate one or more hydrological hypotheses in addition to this architectural objective. Otherwise, I have to say this manuscript is more suitable for a general Earth data science journal.

We agree that we did not clearly state the (hydrological) hypotheses. The underlying hypothesis was that a graph-based network could extract information from local and non-local water storage, climate, and hydrological variables, as well as distant ocean regions, to forecast TWSA on a seasonal timescale. Thanks to the very extensive comments of both reviewers, we now have a much better experimental setup to test these hypotheses, and we clearly state it in the introduction.

L82: Please explain why remote relationships are expected to contribute materially to TWSA prediction relative to local storage memory and local water-balance inputs with literature.

We added citations to studies that examine the influence of remote relationships, such as teleconnections and lag correlations: (Li et al., 2024; Phillips et al., 2012; Humphrey et al., 2016)

L92: Please specify the basin-selection procedure used to obtain the basin nodes. How are grid cells that intersect multiple basins treated?

We assume that, in this comment, "basin-selection procedure" refers to the process by which grid cells are assigned to hydrological basins. Each grid cell is purely defined by its centre point, and this point will be mapped to exactly one basin by finding the "nearest" basin. Since basins are defined as geometries, each grid point within a basin has a distance of 0 to that basin, making it the nearest basin. If the centre of a grid cell would lie exactly on the boundary between two basins, the grid cell would be arbitrarily assigned to one of them, which is a rare edge case. We have rephrased the section to better explain this: 'The graph's connectivity is defined by three sets of edges connecting the nodes of the two levels. Each grid-level node lies within a hydrological basin or ocean cluster. The grid-to-mesh edges link nodes at the grid level to the mesh-level node that represents the basin they lie within.'

L97: Connecting each basin to its geographically nearest neighbors does not necessarily represent hydrological interaction. Please explain how information transfer between hydrologically unrelated neighboring basins should be interpreted.

We agree that this does not necessarily reflect hydrological interaction. The mesh-level edges are purely geometric, and we do not interpret the information passed between neighbouring basins as a hydrological exchange. As this study is a proof of concept, they provide a simple way to give the GNN regional context when it generates input features for the LSTM. We address this in the 'Limitations and future work' section, noting that connecting nodes based on known spatio-temporal correlations, rather than purely on spatial proximity, could improve the model's ability to distinguish regional signals.

L133: Because the Li et al. (2021) reconstruction uses climatic and hydrological predictors, please compare its predictor set with the variables used in this study.

We now list the climatic and hydrological predictors: 'The set of predictors includes 2 m temperature, precipitation fields, sea surface temperature and 17 climate indices as meteorological inputs.

Additionally, they use soil moisture, surface and subsurface runoff, and evaporation as hydrological inputs.' Of these predictors, we use precipitation, temperature, SST, evaporation, and runoff, all from ERA5.

L145: Removing linear trends changes the scientific target. Please explain which hydrological signals are removed by detrending; some regions may be affected by long-term groundwater depletion or glacier mass loss.

Thanks for the remark. We added the explanation; see the answer to the comment on L24.

L161: The CSR mascon product is just distributed on a 0.25-degree grid, but this grid spacing does not represent its independent spatial resolution. Please make sure this sentence is precise.

Thank you for the remark; we updated that by exchanging 'resolution' for 'grid spacing'.

Section 4.1: Aggregating attribution values over all nodes removes the spatial and temporal information that is most relevant to the graph hypothesis. Please provide basin-specific examples and show how attribution changes across input months and lead times.

We agree that aggregating attribution hides information relevant to the graph hypothesis. However, the underlying limitation is, as rightfully stated in the following comment, that nonzero attribution does not demonstrate distinct or useful information. Therefore, we removed this experiment entirely and now rely on the process-based ablation experiments to test the model's design and inputs rather than expanding the previous attribution experiment. Additionally, we included a revised 'latent-removal ' experiment (see Supplement Figures 2 and 3) in the Appendix. Both experiments are evaluated across lead times and basins.

L191: All four latent dimensions having nonzero attribution by themselves does not demonstrate that they contain distinct or useful information. A latent-dimension removal experiment would provide more direct evidence.

We agree, and we removed the statement and replaced it with the attribution analysis for the process-based ablation studies. We ran the suggested removal experiment:

Each of the four dimensions is replaced by its training-period mean in the trained model.

Lead-wise evaluation shows that, as in the original analysis, the latent feature sequences are mostly important at short leads. The most influential channel per seed raises RMSE by 9–53% at lead 1 and by 0–6 % at lead 6. This is the opposite of the lead embedding pattern and is far below the TWSA history (124–144% at lead 1).

Evaluating the experiment per basin (seed 45) shows that the model's reliance on each latent feature varies between basins, and latent features have two distinct patterns with varying influence between all four features, with one feature sequence having overall low influence (see Supplement Figure 3).

This suggests that the latent feature sequences are spatially distinct between channels, but over lead time they share the same pattern of having the largest effect relative to RMSE at short leads and decreasing towards longer leads. We included the experiment in the Appendix.

L261: I cannot understand here. Predicting the full signal may be harder in absolute RMSE terms, but the strong seasonal cycle can make correlation and normalized skill easier to obtain. The comparison should use common anomaly-based metrics.

We agree that the explanation was inadequate. In the original HiGNN-LSTM formulation, the model predicted TWSA, including the seasonal cycle, from the 12 input months, whereas FCast (Li et al. 2026) predicts nonseasonal TWSA and extrapolates seasonal cycles and trends from the previous year to obtain the full TWSA prediction. We meant to explain that for a model that is trained on seasonal TWSA, the loss will be dominated by the seasonal cycle; therefore, the model learns less about anomalies relative to the seasonal cycle.

The revised model is trained on deseasonalised TWSA and forecasts anomalies relative to the training-period climatology. The comparison shows that FCast remains more accurate than HiGNN-LSTM, averaged over the comparison period and lead times. HiGNN-LSTM has lower RMSE in two of the 16 test basins (Euphrates and Zambezi), one more than before. The Figure now shows anomaly correlation instead of correlation (see Supplement Figure 7).

L290: The poor performance in anomaly-dominated basins may also reflect human water management, groundwater abstraction, snow processes, or reconstruction errors. The discussion currently focuses mainly on seasonality and graph connectivity.

We agree that seasonality and graph connectivity were not the only possible causes of the spatial variability in the model's performance. These processes will lower the skill, especially when transferring any model trained on a reconstruction to predict GRACE-derived TWSA. We added this to the discussion of the spatial variability in the forecast.

L306: The experiments show that the full model outperforms ConvLSTM and an embedder-only variant, but they do not identify the learned connections as teleconnections.

We agree, that the margin over the ConvLSTM model does not show that the model has learned teleconnections and additionally ConvLSTM outperforms HiGNN-LSTM when evaluating on GRACE/-FO. We removed that statement and rather refer to the ablation results for any statements about learned connections.

References:

Li, F., Kusche, J., Sneeuw, N., Siebert, S., Gerdener, H., Wang, Z., et al. (2024). Forecasting next year's global land water storage using GRACE data. Geophysical Research Letters, 51, e2024GL109101. https://doi.org/10.1029/2024GL109101

Arsenault, K. R., and Coauthors, 2020: The NASA Hydrological Forecast System for Food and Water Security Applications. Bull. Amer. Meteor. Soc., 101, E1007–E1025, https://doi.org/10.1175/BAMS-D-18-0264.1.

Getirana, A., M. Rodell, S. Kumar, H. K. Beaudoing, K. Arsenault, B. Zaitchik, H. Save, and S. Bettadpur, 2020: GRACE Improves Seasonal Groundwater Forecast Initialization over the United States. J. Hydrometeor., 21, 59–71, https://doi.org/10.1175/JHM-D-19-0096.1.

Phillips, T., Nerem, R. S., Fox-Kemper, B., Famiglietti, J. S., & Rajagopalan, B. (2012). The influence of ENSO on global terrestrial water storage using GRACE. Geophysical Research Letters, 39(16). https://doi.org/10.1029/2012gl052495.

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(2), 357–395. https://doi.org/10.1007/s10712-016-9367-1

Rodell, M., Famiglietti, J.S., Wiese, D.N. et al. Emerging trends in global freshwater availability. Nature 557, 651–659 (2018). https://doi.org/10.1038/s41586-018-0123-1

Supplement

During the revision, we reviewed the code for the preprocessing, graph construction, training and evaluation pipeline and found several bugs that we fixed. Additionally, following the reviewer comments we changed how the forecast target is formulated: the model now predicts deseasonalised TWSA and climatology is added to get the full prediction. We retrained and re-evaluated every model, baselines included. Results of new experiments as requested by the reviewers can be found in Section 2. The revised figures and tables can be found in Section 3. The code changes are listed in detail in Section 4.

1. Revised conclusions

Thanks to the thorough review of both reviewers we conducted an extended analysis of where the HiGNN-LSTM's skill comes from. The results do not confirm some of our earlier conclusions.

1. **The model generates useful predictors from ERA5 data**: Withdrawn. The ablation study shows that when removing ERA5 data from the inputs, the model only performs slightly worse, and the changes are not significant for any lead. Meaningful latent features rather come from the TWSA inputs.
2. **The model encodes spatial teleconnections**: Withdrawn. When ocean nodes are removed from the graph in the ablation study, the model's performance also drops but not significantly. Only at lead 6, the 95% confidence interval does not span 0, hinting at a weak positive effect of the ocean nodes. We hypothesise that this might be due to SST being more correlated to longer lead times. We don't claim the model finds spatial teleconnections.
3. **Graph connections are better than local convolutions**: Restricted. This is only true for our reconstruction test set. Evaluating HiGNN-LSTM and ConvLSTM on the GRACE/-FO-derived TWSA shows that ConvLSTM performs better, even though both have substantially increased RMSE scores (see Table 4).
4. **Strong performance where seasonal signal dominates**: Altered. As the revised model gets the seasonal cycle from the training period climatology and only predicts anomalies to TWSA, the previous correlation measurement is not suitable anymore. We instead score on anomaly correlation, which shows that performance is not necessarily best where seasonal cycle dominates but where anomalies persist and only slowly decay.
5. **HiGNN-LSTM outperforms climatology baseline**: Strengthened. The model is now tested against climatology and persistence and beats both in area-weighted global RMSE at all six leads on the test set and GRACE/-FO.

2. New experiments and figures

2.1 Process-based ablation studies

We test the design choices by running an ablation study. Each time we retrain the model with 8 different seeds and remove 1) the message passing between the grid nodes and the basin nodes (No MP); 2) the edges to the ocean nodes (No Ocean nodes); 3) ERA5 inputs to the model (no ERA5). Note that removing ERA5 inputs from the model input also removes the influence of ocean nodes, as ocean nodes carry no TWSA information. The models are evaluated on the test set as it is not involved in any model development or model checkpoint selection. The results are averaged over the 8 training seeds, 95% confidence intervals over the 8 models are given. P-values come from a Welch t-test over the seeds.

Table 1 shows globally averaged results per lead. Only message passing has a significant influence, and only after lead 1. Ocean nodes do not show significant influence, except for the last lead time. ERA5 input variables seem to have no significant influence at all.

Table 1 Ablations per lead on test dataset d RMSE of HiGNN-LSTM (mean ± standard deviation over 8 seeds) and the difference of each ablation to it in % of its RMSE, with the 95 % confidence interval and p-value of a Welch t-test over the seeds.

	HiGNN-LSTM	No MP		No ocean nodes		No ERA5	
Lead	RMSE [cm]	ΔRMSE [%] (95 % CI)	p	ΔRMSE [%] (95 % CI)	p	ΔRMSE [%] (95 % CI)	p
L1	1.404 ± 0.009	+0.04 (−0.53, +0.60)	0.890	−0.00 (−0.80, +0.79)	0.996	+0.72 (−0.04, +1.47)	0.061
L2	2.036 ± 0.008	+0.45 (+0.05, +0.85)	0.031	+0.13 (−0.39, +0.65)	0.592	+0.21 (−0.26, +0.68)	0.351
L3	2.439 ± 0.005	+0.44 (+0.13, +0.74)	0.009	+0.23 (−0.28, +0.73)	0.332	+0.06 (−0.33, +0.44)	0.753
L4	2.694 ± 0.007	+0.65 (+0.38, +0.93)	< 0.001	+0.49 (−0.07, +1.06)	0.079	+0.26 (−0.17, +0.68)	0.211
L5	2.899 ± 0.013	+0.48 (+0.00, +0.95)	0.049	+0.29 (−0.30, +0.88)	0.311	+0.10 (−0.34, +0.54)	0.643
L6	3.034 ± 0.008	+0.56 (+0.34, +0.78)	< 0.001	+0.38 (+0.05, +0.70)	0.025	+0.31 (−0.06, +0.68)	0.094

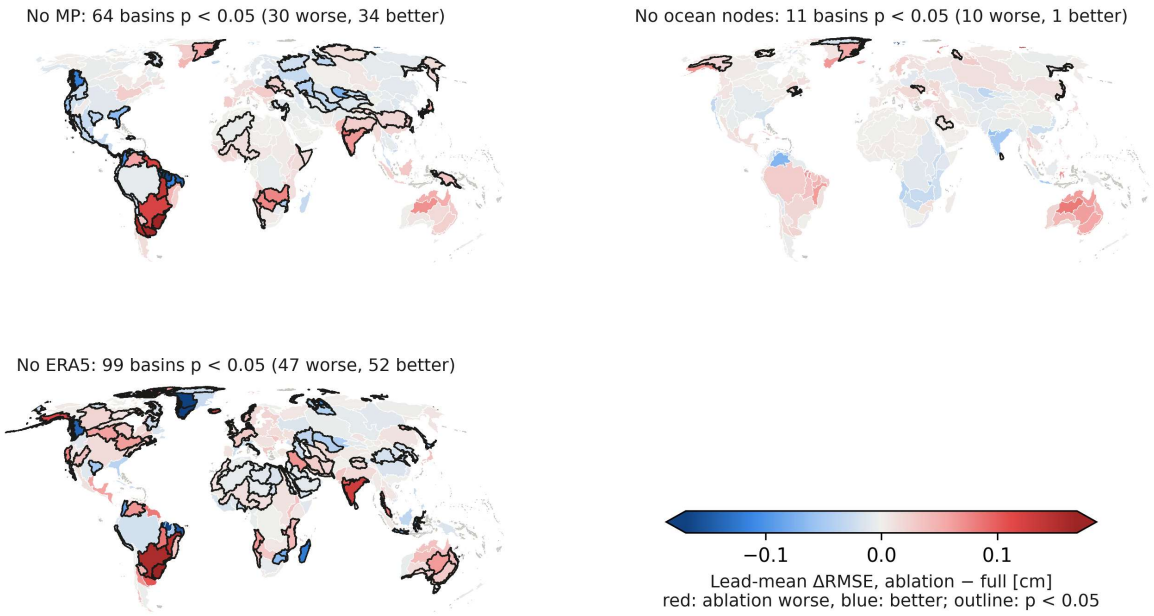


Figure 1 Basin-wise evaluation of ablation study. RMSE differences are averaged over all leads and seeds. Basins outlined in black show significant changes. The colour indicates if the ablation decreased RMSE (blue) or if it increased RMSE (red).

Figure 1 shows the results averaged over all leads per basin. The colours show absolute values of RMSE differences, not relative to how much the RMSE for each of these basins is. Therefore, the relative impact in each basin can not be derived from the figure. Significant basins have black outlines. Significance here is calculated from a Welch test per basin over the seeds. For 226 basins, 11 should be expected to be significant by chance at $p < 0.05$.

No MP has significant influence on 64 basins, but they split almost evenly: 30 get worse RMSE without MP and 34 get better. The basins that suffer from removing the MP cluster in South America, India, southern Africa, and Greenland. Basins that get better are on the North American Pacific coast, in northeastern Brazil and parts of Central Asia and Siberia.

Removing ERA5 from the input affects 99 basins significantly but again they split almost evenly in basins that get worse and better RMSE scores. The effects would cancel out overall, as seen in the globally averaged evaluation. The RMSE rises again in South America, India, the Euphrates region and western and Arctic North America. Basins that get better RMSE are in Greenland and the North American Pacific coast, northeastern Brazil and Madagascar, for example.

Removing ocean nodes has overall little effect and only 11 basins show significant effects, which is what would be expected by chance alone.

2.2 GNN-generated latent feature sequences

To measure how much influence the GNN-generated latent feature sequences have on the final prediction, we set up an experiment where each input to the LSTM is replaced one by one. The inputs are replaced by their training-period mean over all cells, pseudo-steps and leads. For example, the lead embedding becomes the same vector for every lead. We compare the results of the model with the removed information from the latent feature sequences $RMSE_{removal}$ and compare it to the full model as $\Delta RMSE = RMSE_{removal} - RMSE_{full}$.

This experiment can not measure what information the GNN-generated latent features store, only how much influence it has on the subsequent LSTM prediction. The inputs are replaced at the LSTM only, so the encoder still sees the TWSA history and FiLM still sees the lead. The experiment is repeated for all 8 seeds.

Figure 2 shows the results for all 8 seeds, the results of seed 45 is drawn with thick lines. Because each training run learns its own four latent features, their order is arbitrary, and averaging over them would maybe average entirely unrelated features. The experiment shows that removing the TWSA history increases the RMSE by 1.76 to 2.01 cm at lead 1. At lead 6, this goes down to 0.17 to 0.29 cm. Removing the lead embedding has little effect at short leads, but the effect increases at longer leads, where removal increases RMSE by 0.11 to 0.35 cm at lead 6. The most useful latent feature sequences increase RMSE by 0.12 to 0.74 cm at lead 1, decreasing to 0.02 to 0.17 cm at lead 6. From lead 5 onwards, the lead timestep embedding becomes more important than the latent feature sequences. The experiment shows that the LSTM uses the GNN latent features as input mostly at short leads. Their importance decays more slowly than that of the TWSA input, but overall, the TWSA input is the most influential input to the LSTM.

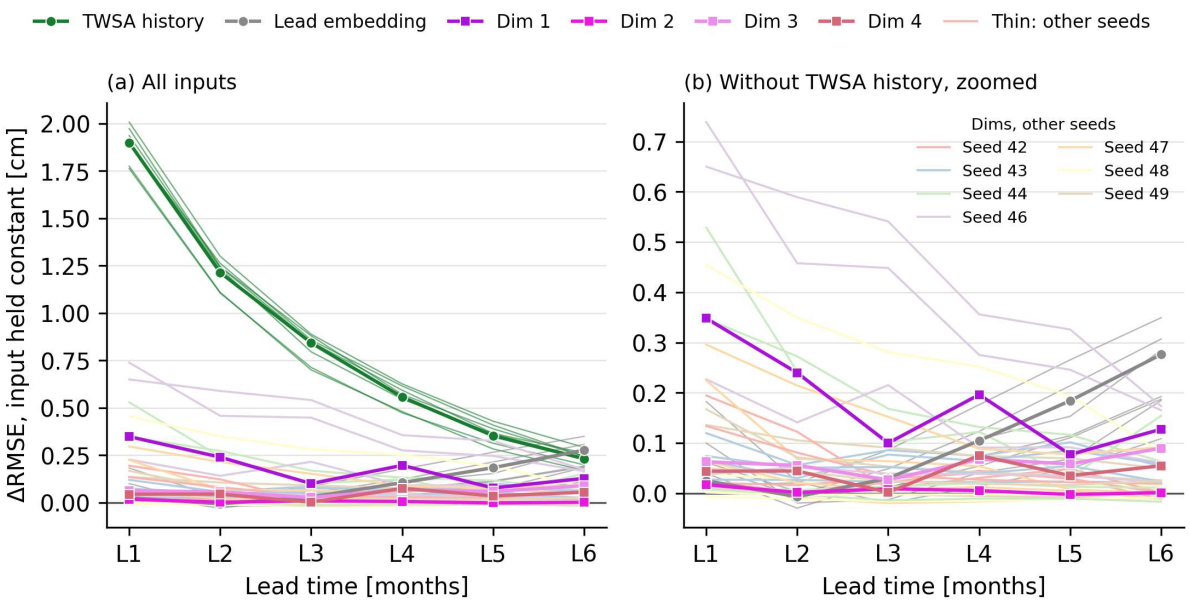


Figure 2 LSTM inputs substituted by mean values evaluated on the test period: change in area-weighted RMSE per lead against the full model with nothing removed. The thick lines show the results for seed 45. The thin lines show the results for all the other seeds.

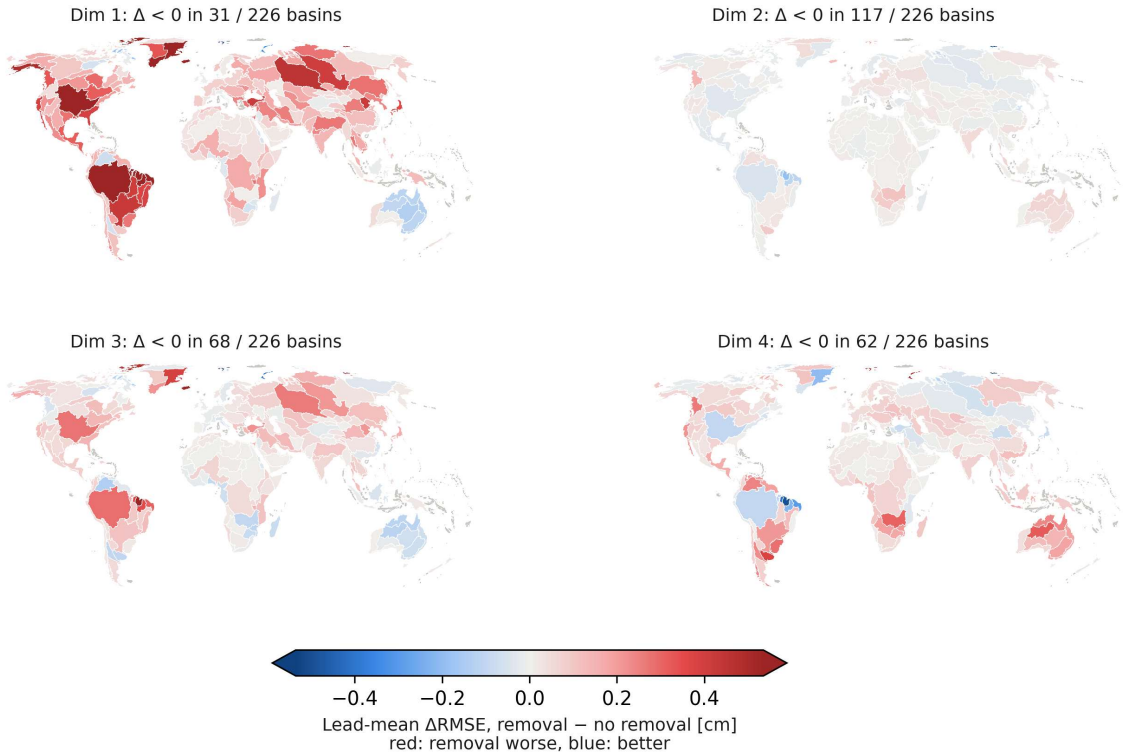


Figure 3 Basin-wise evaluation of latent-feature removal experiment evaluated for one seed (seed 45). RMSE differences are averaged over all leads. The colour indicates if the removal of a feature decreased RMSE (blue) or if it increased RMSE (red).

Figure 3 plots the per basin RMSE differences for each latent feature dimension of seed 45 to qualitatively investigate the spatial importance of each feature sequence. It also shows how many basins decreased their RMSE score with the replaced latent features.

Latent feature dim 1 and dim 3 are highly correlated (Spearman correlation of 0.89) and show very similar pattern although the magnitude is different. Dim 2 seems to have close to no influence in any of the basins and the number of basins that improve vs the number of basins that get worse when the features are replaced almost balanced. Dim 4 is roughly the opposite pattern of Dim 3 (Spearman −0.56).

2.3 Skill per basin per lead on the test set

Table 2 quantitatively evaluates HiGNN-LSTM's RMSE per basin as skill versus the persistence and the climatology baseline. For each lead, the number of basins with RMSE on the test set is lower than that of anomaly persistence or the training climatology is given. The model beats persistence in most basins at every lead (193 at lead 1, 219 at lead 6), but it beats climatology in fewer basins as lead time grows (222 at lead 1, 106 at lead 6), while the median anomaly correlation falls from 0.85 to 0.10. Median skill scores over climatology decreases until it becomes negative at L6. On global average HiGNN-LSTM beats climatology at every lead.

Table 2 Per-basin per lead scores: for HiGNN-LSTM on the test set for all 226 basins. Mean over 8 seeds. Counts give the number of basins where model's skill score SS_{ref} = 1-RMSE/$RMSE_{ref}$ is positive.

	L1	L2	L3	L4	L5	L6
No. of Basins > persistence	193	215	216	218	220	219
No. of Basins > climatology	222	196	172	143	123	106
Median r_{anom}	0.85	0.66	0.48	0.37	0.23	0.10
Median SS_{pers}	0.06	0.08	0.11	0.14	0.17	0.19
Median SS_{clim}	0.51	0.27	0.14	0.07	0.01	−0.01

3. Revised figures and tables

The new results include updates requested by the reviewers, such as training the model on multiple seeds, reporting standard deviations of the results, anomaly scores that measure skill independent of climatology, and a persistence baseline.

3.1 Evaluation on the test set

Table 3 shows the new results on the reconstruction test set. RMSE scores at lead 1 went from 1.83 cm to 1.40 cm, and at lead 6 from 3.77 to 3.03. The model now beats climatology also at lead 6. Remark that due to the new preprocessing also the climatology RMSE was decreased at every lead. HiGNN-LSTM also beats ConvLSTM at every lead.

NSE_{anom} and r_{anom} are calculated against climatology, therefore NSE_{anom} of climatology is 0, and r_{anom} is undefined.

Table 3 Globally averaged scores on the test set (June 2017 - June 202) per lead time (mean ± standard deviation over 8 seed). Persistence: last input month's TWSA anomaly added to the target month. Climatology derived from the reconstruction's training period.

Model	Metric	L1	L2	L3	L4	L5	L6
HiGNN-LSTM	RMSE [cm]	1.404 ± 0.009	2.036 ± 0.008	2.439 ± 0.005	2.694 ± 0.007	2.899 ± 0.013	3.034 ± 0.008
	NSE_{anom}	0.821 ± 0.002	0.617 ± 0.003	0.450 ± 0.002	0.331 ± 0.003	0.229 ± 0.008	0.161 ± 0.005
	r_{anom}	0.811 ± 0.001	0.607 ± 0.001	0.457 ± 0.002	0.340 ± 0.003	0.237 ± 0.003	0.146 ± 0.004
ConvLSTM	RMSE [cm]	1.429 ± 0.008	2.050 ± 0.005	2.459 ± 0.007	2.725 ± 0.009	2.918 ± 0.011	3.054 ± 0.012
	NSE_{anom}	0.814 ± 0.002	0.611 ± 0.002	0.441 ± 0.003	0.314 ± 0.005	0.219 ± 0.006	0.149 ± 0.007
	r_{anom}	0.805 ± 0.002	0.604 ± 0.001	0.452 ± 0.001	0.336 ± 0.002	0.235 ± 0.004	0.145 ± 0.005
Persistence	RMSE [cm]	1.474	2.207	2.718	3.097	3.409	3.662
	NSE_{anom}	0.802	0.550	0.317	0.116	−0.063	−0.219
	r_{anom}	0.800	0.592	0.441	0.321	0.215	0.123
Climatology	RMSE [cm]	3.323	3.300	3.303	3.309	3.317	3.325

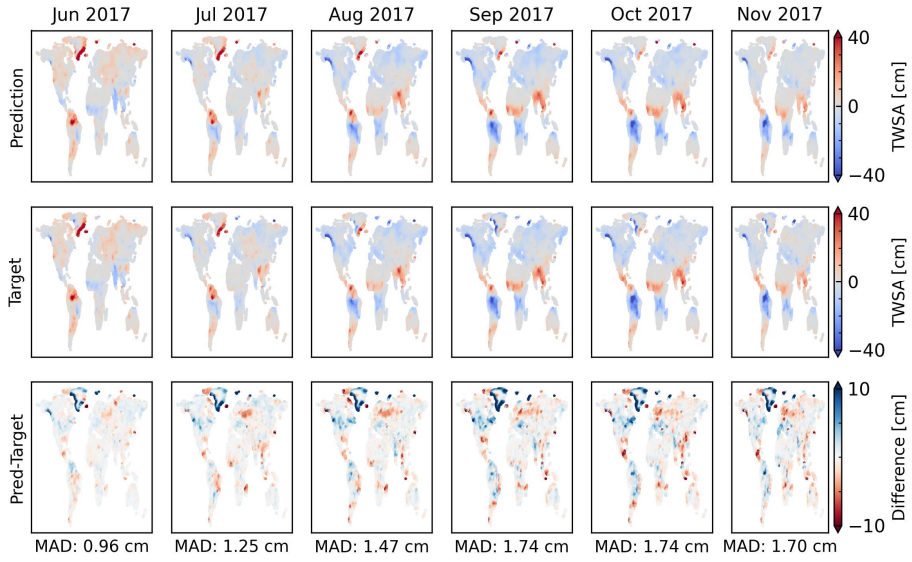


Figure 4 Six-month forecast (June 2017 - November 2017) from input time series June 2016 to May 2017. Rows show (top) model prediction from seed 45, (middle) target TWSA from the reconstruction, and (bottom) their difference, with the Mean Absolute Difference (MAD) indicated below each difference map.

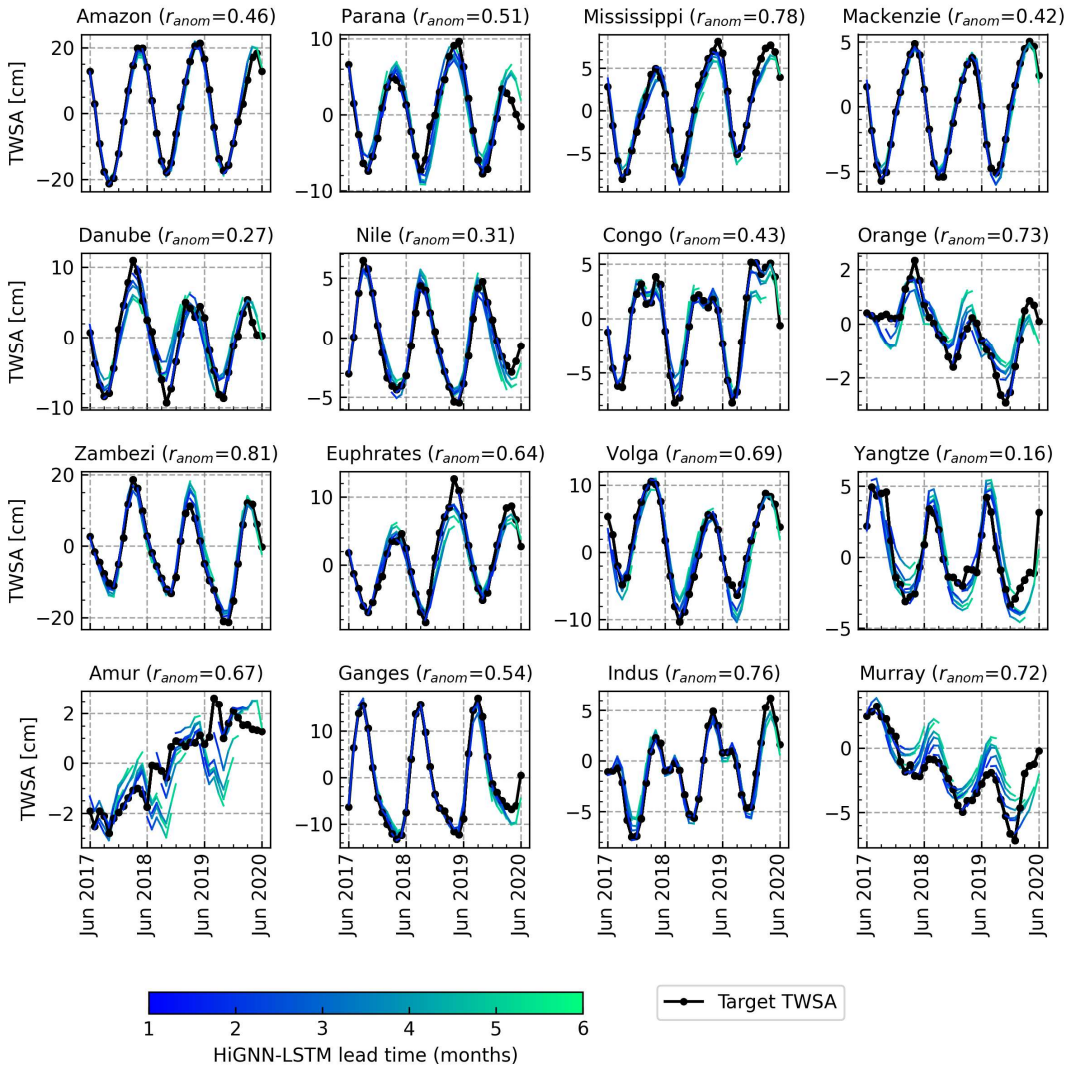


Figure 5 Predicted TWSA from seed 45 (blue) vs target TWSA (black) in 16 test basins together with anomaly correlation coefficient (r_{anom}) averaged over all forecast sequences for the given basin from June 2017 until June 2020.

3.2 Evaluation on GRACE-derived TWSA

We evaluate how HiGNN-LSTM transfers to observations using GRACE/-FO-derived TWSA of uninterrupted sequences between March 2008 and November 2023. The scores on this dataset get substantially worse, but the model can still beat the persistence and the climatology baseline. The persistence baseline takes the anomaly value from the last input and adds it to the climatology derived from the reconstruction training dataset. We also added a climatology$_{GRACE}$ baseline that is derived from the GRACE-derived TWSA before March 2008. As this is only a short sequence it has even worse scores than the climatology derived from the reconstructed TWSA dataset.

Table 4 Globally averaged scores on the GRACE-derived TWSA for valid sequences between March 2008 and November 2023 per lead time (mean ± standard deviation over 8 seed). Persistence: last input month's TWSA anomaly added to the target month. Climatology is derived from the training set and Climatology$_{GRACE}$ is derived from the available GRACE-derived TWSA

available during the training period.

Model	Metric	L1	L2	L3	L4	L5	L6
HiGNN-LSTM	RMSE [cm]	4.593 ± 0.023	5.322 ± 0.022	5.676 ± 0.026	5.843 ± 0.020	5.948 ± 0.025	6.077 ± 0.016
	NSE_{anom}	0.470 ± 0.005	0.300 ± 0.006	0.209 ± 0.007	0.158 ± 0.005	0.123 ± 0.007	0.085 ± 0.005
	r_{anom}	0.526 ± 0.002	0.377 ± 0.002	0.301 ± 0.003	0.253 ± 0.002	0.226 ± 0.002	0.169 ± 0.002
ConvLSTM	RMSE [cm]	4.453 ± 0.007	5.213 ± 0.022	5.572 ± 0.015	5.763 ± 0.012	5.885 ± 0.013	6.032 ± 0.010
	NSE_{anom}	0.502 ± 0.002	0.329 ± 0.006	0.238 ± 0.004	0.181 ± 0.003	0.142 ± 0.004	0.099 ± 0.003
	r_{anom}	0.539 ± 0.001	0.387 ± 0.002	0.310 ± 0.001	0.262 ± 0.001	0.231 ± 0.002	0.174 ± 0.002
Persistence	RMSE [cm]	4.907	5.959	6.532	6.858	7.102	7.417
	NSE_{anom}	0.392	0.119	−0.047	−0.159	−0.248	−0.358
	r_{anom}	0.520	0.366	0.292	0.246	0.222	0.165
Climatology	RMSE [cm]	6.394	6.408	6.413	6.395	6.376	6.372
$Climatology_{GRACE}$	RMSE [cm]	6.832	6.819	6.806	6.775	6.743	6.724

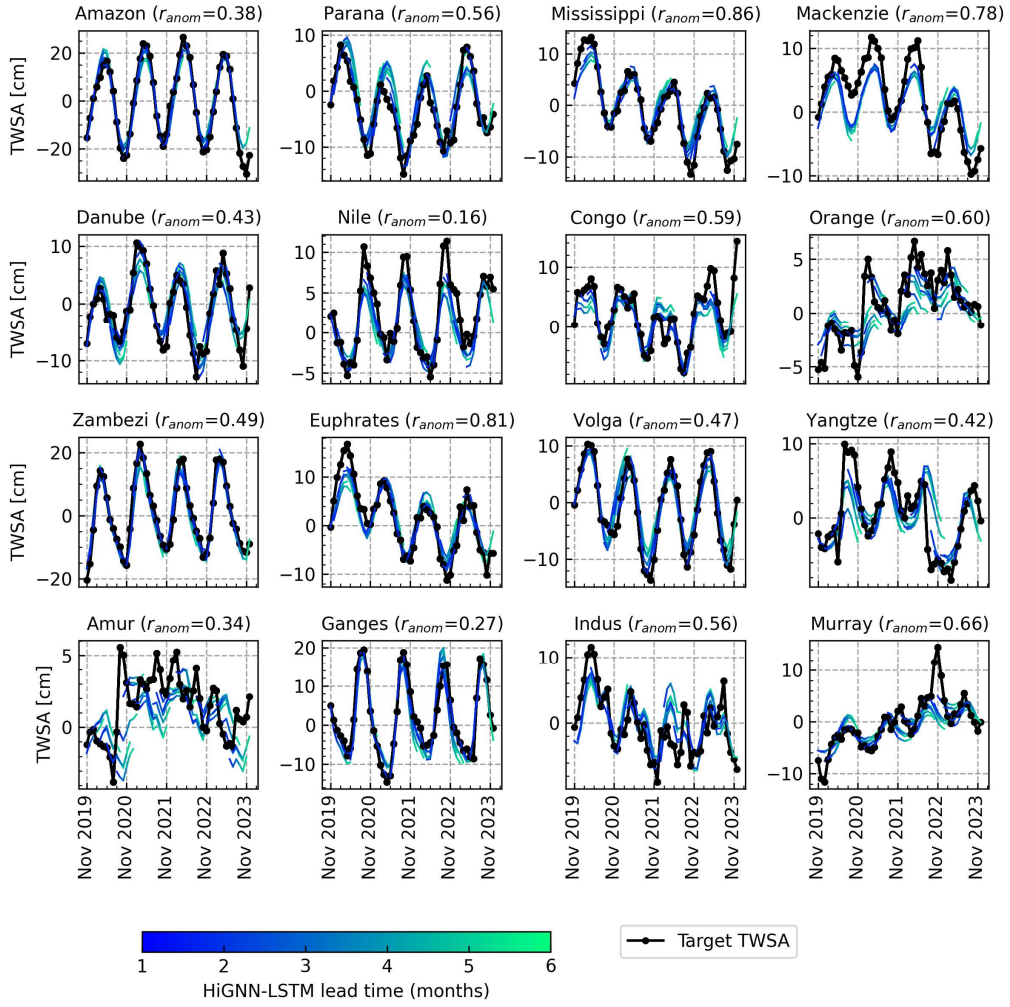


Figure 6 Predicted TWSA from seed 45 (blue) vs target GRACE-derived TWSA (black) in 16 test basins. r_anom: lead-averaged correlation of deseasonalised basin averages.

3.3 Comparison to other forecasts

Also with the new target formulation (predicting only anomalies and adding climatology) the HiGNN-LSTM model can only beat Li et al. (2026) RMSE scores in two of our 16 test basins (Euphrates and Zambezi).

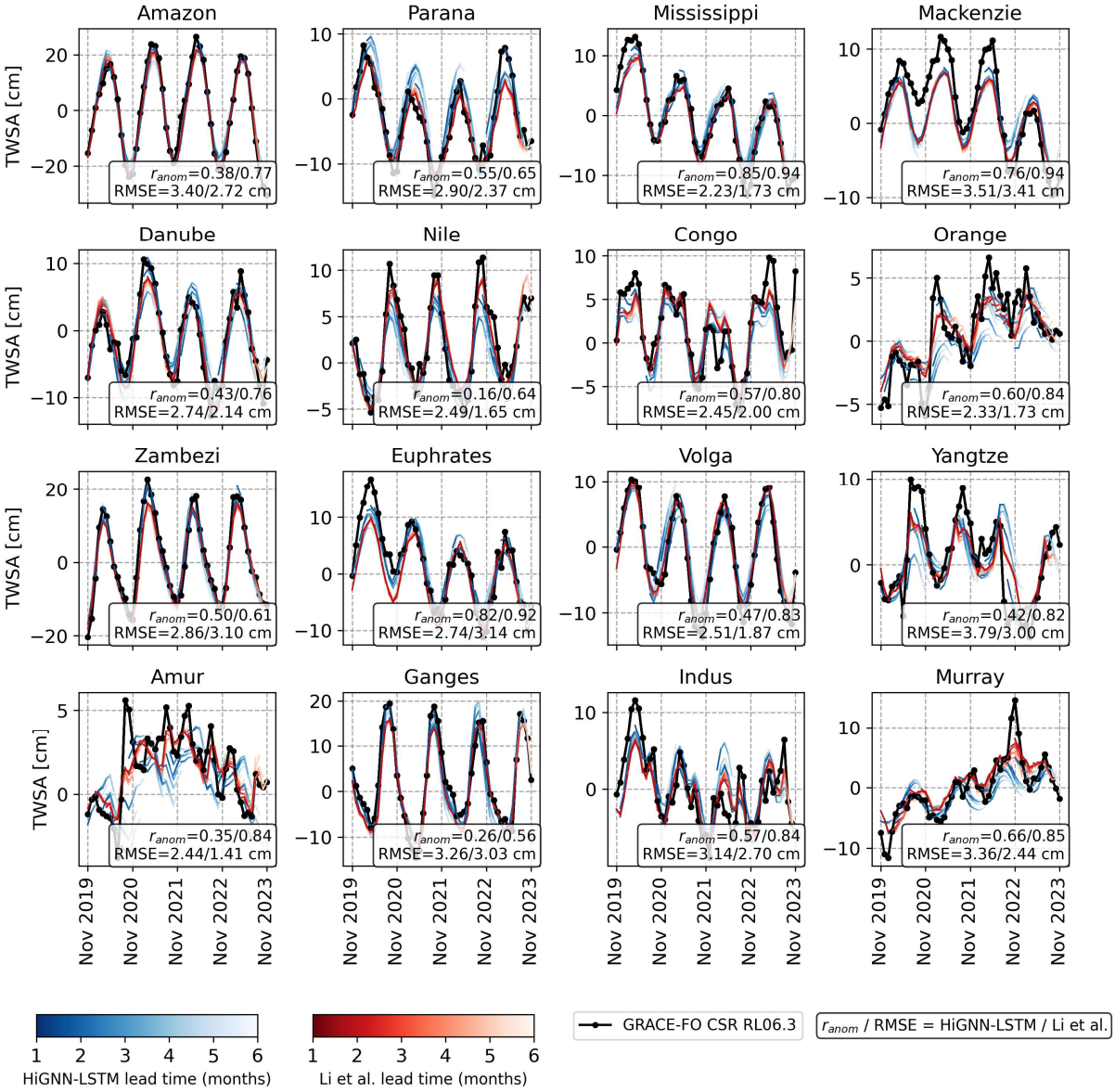


Figure 7 HiGNN-LSTM forecast of seed 45 (blue) and Li et al. (red) vs. GRACE-derived TWSA (black) in the 16 test basins. The lighter the line color of the forecasted TWSA, the longer the lead time. Anomaly correlation coefficient ($r_{_anom}$) and Root Mean Squared Error (RMSE) for the basins are given for the HiGNN-LSTM architecture/Li et al. (2026).

3.4 Graph construction: new ocean nodes

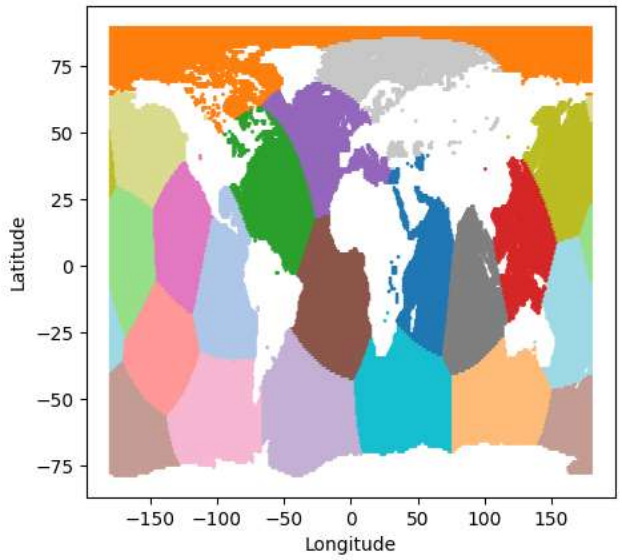


Figure 8 Ocean clusters as they are represented in the hierarchical graph. Each colour indicates a cluster formed by K-means clustering of grid-cell positions.

3.5 Graph parameter ablation

We reran the study to assess the extent to which our choice of parameters affects HiGNN-LSTM performance. To save computational time, we ran these experiments on only four seeds. We tested two values for the number of ocean clusters (10 and 50), two values for the number of mesh-to-mesh edges (5 and 20) and two values for the number of mesh-to-grid edges (1 and 5). The base configuration of HiGNN-LSTM is 20 ocean clusters, 10 mesh-to-mesh edges, and 3 mesh-to-grid edges. None of the tested parameters produces significantly different RMSE scores compared to the base configuration. The confidence intervals all include 0.

Table 5 Graph parameter study: area-weighted test RMSE of HiGNN-LSTM in cm (mean ± standard deviation over 8 seeds), and for each variant (4 seeds) ΔRMSE, the difference to HiGNN-LSTM in cm with 95 % confidence intervals over the seeds below. Each variant changes only one parameter and retrains the model. Ocean clusters (OC), M2M: mesh-to-mesh edges, M2G: mesh-to-grid edges.

Model	L1	L2	L3	L4	L5	L6
HiGNN-LSTM	1.404 ± 0.009	2.036 ± 0.008	2.439 ± 0.005	2.694 ± 0.007	2.899 ± 0.013	3.034 ± 0.008
OC = 10	+0.006 0.004, +0.016	+0.008 0.005, +0.021	+0.004 0.000, +0.009	+0.005 0.003, +0.012	+0.016 0.020, +0.053	+0.007 0.002, +0.016
OC = 50	+0.003 0.005, +0.010	+0.006 0.011, +0.023	+0.001 0.009, +0.011	+0.002 0.008, +0.011	+0.014 0.027, +0.054	+0.007 0.001, +0.014
M2M = 5	+0.002 0.005, +0.010	−0.000 0.014, +0.013	−0.000 0.011, +0.010	−0.002 0.010, +0.007	+0.007 0.037, +0.051	+0.002 0.010, +0.014
M2M = 20	+0.003 0.006, +0.012	+0.006 0.014, +0.026	+0.001 0.013, +0.014	+0.001 0.013, +0.016	+0.014 0.030, +0.057	+0.006 0.001, +0.013
M2G = 1	+0.001 0.018, +0.019	+0.002 0.015, +0.019	+0.006 0.008, +0.020	+0.004 0.019, +0.028	+0.009 0.012, +0.030	−0.001 0.013, +0.011
M2G = 5	+0.001 0.008, +0.011	+0.006 0.019, +0.030	−0.002 0.020, +0.016	+0.002 0.010, +0.014	+0.016 0.034, +0.066	+0.005 0.006, +0.016

Input: grid inputs $\mathbf{X}_{n-11:n}$, ocean inputs $\mathbf{O}_{n-11:n}$, TWSA anomaly $a_{n-11:n}$

Output: $\widehat{\mathrm{TWSA}}_{n+1}, \ldots, \widehat{\mathrm{TWSA}}_{n+6}$

$\mathbf{x} \leftarrow \textsc{Embed}(\mathbf{X}, \mathbf{O})$

$\mathbf{g} \leftarrow \textsc{Grid2Mesh}(\mathbf{x})$ ▷ shared by all leads

for $i = 1, \ldots, 6$ **do**

  $\mathbf{e}_i \leftarrow \textsc{LeadEmbed}(i)$

  $\mathbf{g}^{(i)} \leftarrow \textsc{Mesh2Mesh}(\mathbf{g}; \mathbf{e}_i)$

  $\mathbf{z}^{(i)}_{n-11:n} \leftarrow \textsc{Mesh2Grid}(\mathbf{g}^{(i)}; \mathbf{e}_i)$

  $\mathbf{h}_0, \mathbf{c}_0 \leftarrow \mathbf{0}$ ▷ state not carried over

  $\mathbf{h} \leftarrow \mathrm{LSTM}\big([\mathbf{e}_i, a_k, \mathbf{z}^{(i)}_k]_{k=n-11}^{n}; \mathbf{h}_0, \mathbf{c}_0\big)$ ▷ same a every lead

  $\hat{a}_{n+i} \leftarrow \textsc{Readout}(\mathbf{h})$

  $\widehat{\mathrm{TWSA}}_{n+i} \leftarrow \sigma\, \hat{a}_{n+i} + \mu + C\big(m(n+i)\big)$ ▷ unnormalize and add climatology

end for

Figure 9 Algorithm of HiGNN-LSTM forecast for 6 months.

Table 6 Anomaly skill of HiGNN-LSTM for the 16 test basins per lead time on the test set. r_{anom}: anomaly correlation; SS_{pers} and SS_{clim}: skill score 1 − RMSE/$RMSE_{ref}$ against anomaly persistence and the training climatology, positive where the model is better.

Basin	Metric	L1	L2	L3	L4	L5	L6
Amazon	r_{anom}	0.90	0.67	0.48	0.40	0.36	0.34
	SS_{pers}	−0.02	0.03	0.06	0.09	0.11	0.15
	SS_{clim}	0.60	0.33	0.16	0.12	0.09	0.08
Parana	r_{anom}	0.89	0.73	0.54	0.43	0.36	0.32
	SS_{pers}	0.14	0.19	0.22	0.23	0.23	0.24
	SS_{clim}	0.54	0.30	0.13	0.08	0.06	0.05
Mississippi	r_{anom}	0.93	0.81	0.76	0.74	0.73	0.70
	SS_{pers}	0.07	−0.01	−0.04	−0.07	−0.07	−0.06
	SS_{clim}	0.64	0.41	0.33	0.29	0.26	0.22
Mackenzie	r_{anom}	0.87	0.65	0.43	0.27	0.20	0.14
	SS_{pers}	0.12	0.12	0.13	0.15	0.17	0.20
	SS_{clim}	0.61	0.38	0.22	0.13	0.08	0.05
Danube	r_{anom}	0.82	0.46	0.25	0.13	0.06	−0.03
	SS_{pers}	0.16	0.12	0.13	0.16	0.18	0.21
	SS_{clim}	0.41	0.05	−0.08	−0.11	−0.12	−0.13
Nile	r_{anom}	0.80	0.53	0.31	0.14	0.06	0.05
	SS_{pers}	0.07	0.09	0.10	0.14	0.14	0.15
	SS_{clim}	0.50	0.25	0.10	0.00	−0.03	−0.02
Congo	r_{anom}	0.77	0.59	0.46	0.31	0.27	0.11
	SS_{pers}	−0.02	0.05	0.06	0.07	0.11	0.14
	SS_{clim}	0.47	0.30	0.19	0.09	0.07	0.00
Orange	r_{anom}	0.95	0.88	0.79	0.72	0.64	0.56
	SS_{pers}	0.02	0.07	0.08	0.09	0.13	0.15
	SS_{clim}	0.69	0.52	0.37	0.29	0.22	0.16
Zambezi	r_{anom}	0.97	0.91	0.86	0.79	0.71	0.62
	SS_{pers}	0.05	0.02	0.03	0.06	0.09	0.10
	SS_{clim}	0.75	0.58	0.47	0.37	0.29	0.22
Euphrates	r_{anom}	0.91	0.78	0.69	0.59	0.49	0.37
	SS_{pers}	0.11	0.08	0.09	0.10	0.12	0.15
	SS_{clim}	0.60	0.39	0.29	0.21	0.15	0.09
Volga	r_{anom}	0.94	0.83	0.75	0.66	0.56	0.41
	SS_{pers}	0.11	0.10	0.14	0.18	0.21	0.23
	SS_{clim}	0.69	0.47	0.36	0.27	0.18	0.09
Yangtze	r_{anom}	0.79	0.44	0.23	0.05	−0.19	−0.40
	SS_{pers}	0.10	0.09	0.13	0.18	0.20	0.26
	SS_{clim}	0.37	0.02	−0.10	−0.16	−0.26	−0.33
Amur	r_{anom}	0.94	0.83	0.72	0.61	0.48	0.41
	SS_{pers}	0.06	0.05	0.07	0.10	0.12	0.15
	SS_{clim}	0.65	0.42	0.27	0.16	0.06	0.02
Ganges	r_{anom}	0.78	0.62	0.53	0.44	0.48	0.52
	SS_{pers}	0.02	0.04	0.06	0.05	0.03	0.02
	SS_{clim}	0.47	0.31	0.22	0.14	0.12	0.10
Indus	r_{anom}	0.92	0.82	0.73	0.70	0.69	0.69
	SS_{pers}	−0.01	0.01	−0.00	−0.00	0.02	0.07
	SS_{clim}	0.70	0.55	0.43	0.36	0.32	0.28
Murray	r_{anom}	0.97	0.91	0.79	0.67	0.57	0.44
	SS_{pers}	0.00	−0.01	0.01	0.03	0.04	0.06
	SS_{clim}	0.79	0.63	0.48	0.38	0.31	0.25

4. Code changes in detail

1. Data preprocessing

- Coarsening of the TWSA reconstruction: The area-weighted averaging from 0.5° to 1° counted the area of ocean (NaN) sub-cells in the denominator, which pulled coastal land cells towards zero. Only valid land area is now used in the denominator.
- Latitude weights in the ERA5 coarsening: The cell areas were shifted by half a cell, and the southernmost latitude row was dropped: Areas are now computed from proper cell edges.
- Regridding: TWSA used to be reindexed onto the ERA5 grid, which added an all-NaN row at the −90° pole. ERA5 is now regridded onto the TWSA grid, so TWSA defines the grid and the land mask.
- GRACE/GRACE-FO evaluation data:
    - Two GRACE solutions were matched to the same ERA5 month.
    - ERA5 was detrended over a shorter period than in the training data, so the model's inputs at GRACE evaluation differed from those it saw in training.

2. Graph construction

- Node embedding: The old "trigonometric" embedding read longitude as latitude. It encoded $(\cos \lambda, \sin \phi, \cos \phi)$, so cells at $+\lambda$ and $-\lambda$ were indistinguishable. This affected the ocean clustering, the mesh-node features and every k-nearest-neighbour edge set, which could also be seen in Figure A1 showing the ocean clusters. We substituted for cartesian unit-sphere embedding $(\cos \phi \cos \lambda, \cos \phi \sin \lambda, \sin \phi)$, where nearest neighbours are nearest in great-circle distance.
- Ocean-cluster and basin positions: Cluster positions were the planar mean of lon/lat. For clusters crossing the dateline, this was off by up to 54°. Ocean clusters and basins are now placed at the spherical mean of their member cells.

3. Normalization

- Train vs val/test: For validation and test data, NaNs were filled before normalisation, while for training data they were filled after. A missing value was therefore encoded differently between the datasets.
- Standard-deviation floor: Cells with small temporal variability had their standard deviation replaced by 1 (below 10 % of the median std), so many low-variability cells were effectively left unscaled. The std is now clipped from below at 1 % of the median.
- Runoff clipping: Runoff was clipped at zero before normalisation to remove unphysical values. However, ERA5 inputs are linearly detrended, so this clip set about 59 % of land runoff values to zero, which had valid values. It has been removed.

4. Forecast target and inputs

- The model now predicts deseasonalised TWSA relative to the training-period monthly climatology. The climatology is added back to give TWSA in cm. Both reviewers noted that keeping the reproduced seasonal cycle can dominate correlation and asked for skill on deseasonalised anomalies. This was also the step we proposed as future work in the original manuscript.

- No residual prediction: The model used to output an increment added to the last observed TWSA, which assumes the next month starts from the current state. When predicting the deseasonalised TWSA, we remove this residual framing, such that predicting a zero output corresponds to predicting climatology, not persistence of the last input month's value.
- ERA5 inputs are deseasonalised as well. We subtract each variable's training-period monthly climatology.

5. Architecture and training

From the corrections and adaptations described above, we re-tuned the model on the validation set:

hidden size 64 → 32; mesh processing layers 2 → 1; LSTM layers 2 → 1; learning rate $5 \cdot 10^{-4} \rightarrow 2 \cdot 10^{-3}$ with cosine annealing; 50 instead of 100 epochs; effective batch size unchanged at 16.

Every configuration is trained with 8 random seeds. We report mean ± std.

6. Evaluation

- Metrics: All metrics are area-weighted (latitude-weighted RMSE, anomaly NSE, anomaly r) and reported per lead time.
    - Anomaly NSE (NSE_anom): Nash–Sutcliffe efficiency with the training climatology as the benchmark: $\text{NSE_anom} = 1 - \Sigma\, w\, (\hat{y} - y)^2 / \Sigma\, w\, (C - y)^2$. At each lead, it is pooled over all land cells, with area weights $w = \cos(\text{lat})$. Climatology gets NSE=0, values above 0 mean the forecast beats it. There is no lower bound.
    - Anomaly correlation (r_anom): Pearson correlation between forecast and observed anomalies ($\hat{y} - C$ and $y - C$). In the per-lead tables it is computed per cell and then averaged with area weights; as per-basin score it is computed on area-weighted basin-mean series.
    - Skill score (SS_ref): $\text{SS_ref} = 1 - \text{RMSE_model} / \text{RMSE_ref}$, where the reference is either the training climatology (SS_clim) or anomaly persistence (SS_pers). Anomaly persistence means the climatology plus the anomaly of the last input month, held constant over all six leads. A positive value means the model is closer to the observations than the reference, 0 means it does as well, and 1 is a perfect forecast.

- Baselines and ablations:
    - Added anomaly persistence baseline: last input anomaly is carried forward unchanged and added to climatology for each lead time.
    - the ConvLSTM baseline gets the same data and target changes;
    - new ablations: removing ocean nodes, removing ERA5 data from the inputs, removing message passing between graph nodes.