Dear Reviewer #2,

We thank you for your review. Your comments helped us improve the manuscript a lot, for which we are very grateful.

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 The model generates useful predictors from ERA5 data: withdrawn.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 The model encodes spatial teleconnections: withdrawn.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 we can answer all comments to your satisfaction.

Best regards

Viola Steidl

On behalf of all co-authors

1. The model isn't tested by isolating it from ERA5. Is there any evidence that adding ERA5 substantially improves the model? Retrain the model with only TWSA and month of the year and see if the performance drops. Then check if the ERA5 variables have contribution to the model or just adding complexity to the model. If the paper's claim is to improve performance by ERA5, there should be a clear comparison with and without ERA5.

Thank you for the comment. We added an ablation study that addresses this question (see Supplement Sec. 2.1). We retrained HiGNN-LSTM using only TWSA and the month of the year as inputs, without ERA5 data. Removing ERA5 from the input features increases globally averaged RMSE on the test period at every lead, but not significantly. Basin-wise evaluation shows that, averaged

over all leads, ERA5 has significant influence on the forecast score in 99 of 226 basins. Out of these, 52 basins actually achieve better RMSE scores if ERA5 is removed from the input, so for these basins, ERA5 just adds complexity instead of improving the forecast.

Most of the predictive skill comes from the memory of the TWSA anomaly itself. Therefore, we removed the claim that the GNN "automatically learns useful predictors from ERA5" from the abstract, introduction and conclusions, and apologise for not providing evidence in the original manuscript.

2. The evaluation against GRACE observations (Table 3, Sect. 4.3) lacks sufficient baselines and diagnostics.

2a) Inconsistency between reconstruction and GRACE evaluation

Table 2 and Table 3 need meaningful and strong interpretation based on the comparison made against observation. The model loses to climatology at lead 6 on the reconstruction (3.77 vs. 3.69 cm) but beats climatology at every lead against GRACE, including lead 6 (6.96 vs. 7.96 cm). It is not clear why or how the model performs worse than the baseline on the reconstruction but better on the GRACE observation. It should be clearly stated how Table 3 climatology was computed.

From the released code in the repository evaluation_HiGNN.py, it seems that the climatology is computed from the reconstruction training period get_climatology() and reused unchanged for the GRACE evaluation. If so, then Table 3 compares the model against climatology derived from different smoother product which would explain the values in the table.

This is correct. The table compares the climatology computed from the reconstruction TWSA for the training period. This was intended to show that the model adds skill over the simplest baseline that can be derived from the training data. We agree that this was not clearly stated and will clarify in the table's caption.

In the original version, the model had to predict the seasonal cycle together with the anomaly from the 12-month input. At lead 6, it got slightly worse than climatology from the training data. However, on the GRACE-derived data, the model could outperform the reconstruction climatology because it predicted TWSA as a change to the last TWSA input, and therefore it already had the correct offset that the climatology did not.

In the revised manuscript, the model predicts the anomaly relative to the training climatology, which is added back to the output. With this formulation, the HiGNN-LSTM beats climatology at every lead on both the reconstruction test set and the GRACE/-FO dataset.

In Table 4 (Supplement), we added a climatology baseline calculated from GRACE-derived TWSA prior to the evaluation period (April 2002 - February 2008). The GRACE-derived climatology is a poorer baseline than the reconstruction climatology because it is based on only a few years of observations.

2b) The error in Table 3 stops growing as the lead time increases, and this is never discussed

Table 2 almost doubles across the horizon (1.83 → 3.77 cm), as expected. Table 3 goes 5.59 → 6.63 → 6.98 → 6.98 → 6.89 → 6.96 cm: it rises for two months, then flattens and even dips at lead 5. This suggests a fixed error floor near 7 cm caused by the mismatch between training data and observations rather than by forecast difficulty. Please discuss this, and clarify whether leads 3 to 6 in Table 3 measure forecast skill or a lead-independent error floor.

Thank you for the observation. The preprocessing pipeline had errors that we addressed during the review (see Supplement). With the corrected data and the anomaly formulation of the target, the error relative to GRACE/-FO-derived TWSA increases steadily with lead time.

However, the revised model predicts anomalies and the reconstruction climatology is added to them, so part of the error at every lead comes from the systematic difference between the reconstruction and GRACE/-FO observations. RMSE does not separate forecast skill from this error floor. We added the anomaly NSE and anomaly correlation metrics, as they are less affected by this offset. NSE compares model predictions against climatology, which, of course, has the same systematic difference. Anomaly correlation is computed per grid cell over time, so a constant offset between the two datasets does not affect it. Both decrease steadily with lead time (see Supplement Table 4). On GRACE/-FO, the anomaly NSE falls from 0.47 at lead 1 to 0.09 at lead 6, and the anomaly correlation falls from 0.53 to 0.17.

2c) The ConvLSTM baseline is missing from the GRACE evaluation

Table 2 compares HiGNN-LSTM with a ConvLSTM and climatology but in Table 3, ConvLSTM is dropped without any explanation. Since the ConvLSTM supports the claim at lines 214–215 that GNN features beat local convolutions, its absence removes that comparison from the one test of whether the advantage holds on real observations.

Thanks for the remark, originally we did not report ConvLSTM scores on GRACE/-FO as the section was intended to show the limited transferability to observations. We have now added the ConvLSTM results (see Supplement Table 4).

On GRACE/-FO, the ConvLSTM performs better than HiGNN-LSTM at every lead. The advantage of the GNN features does not hold when transferring directly to real observations. The differences between the models are small compared with the loss of skill when moving from the reconstruction to GRACE/-FO, which affects both models alike (lead-1 anomaly NSE drops from 0.82 to about 0.5). Both models beat persistence and climatology on GRACE/-FO at all leads (see Supplement Table 4).

Accordingly, we restrict our claim about graph-based features outperforming local convolutions to the reconstructed test dataset. We add this to our conclusion that already states transferability to GRACE/-FO derived TWSA remains an open challenge.

3. No uncertainty quantification is reported.

It appears from table 2 and table 3 that results are reported based on a single run without any seed variability, confidence interval, and significance test. Appendix B (l. 334–338) states 0.05 to 0.12 cm difference "confirming" while treating differences of at most 0.05 cm as negligible, which is self-contradictory with no variance estimate.

Thank you for this remark. We reran the experiments with 8 seeds and included standard deviations (see Supplement Tables 3 and 4). For the ablation studies, we additionally report 95% confidence intervals and p-values from a Welch t-test over the seeds. They show whether a difference exceeds seed noise on the test period, not how much it depends on that period. We use the test period because it was held out from training and hyperparameter tuning.

The ablation shows that the benefit of message passing is significant after lead 1, while the ocean nodes' benefit is not significant except for lead 6 (see Supplement Table 1).

As in the original manuscript, the new ablations on graph design choices (number of ocean nodes, number of mesh-to-mesh or mesh-to-graph edges) show no significant change to the model's performance at any lead on the test set. The results are now reported with 95% confidence intervals. The ablations on the graph design were run with 4 seeds instead of 8 to reduce computational effort (see Supplement, Table 5).

4. The overall workflow what is described in Figure 1 is ambiguous

It is not clear where the autoregressive forecast and the TWSA input sequence are implemented in the workflow, and what part of the model contributes to the autoregressive component. If it gets fixed for 12 months then it is not autoregressive. It should be formulated clearly if it is fixed offset prediction or recursive prediction from previous month. The figure itself with a minimal formulation of the method is ambiguous. There is only one equation in the manuscript which has no role in the model description. Although Figure 1 is useful for the overall workflow and spatial pipeline but does not have enough information about the temporal formulation. I would ask the authors to correct or explain the arrow (autoregressive label) and to add a short algorithm block, or a small set of equations, making the forecast procedure explicit. In particular, what the LSTM receives at each lead, whether the 12-month history is held fixed, what the residual is added to, and whether the recurrent state is carried across leads.

Indeed, the model is not autoregressive, and the figure was tailored to an earlier version. We apologise for this mistake and are very thankful for the reviewer noticing the ambiguity. We updated the figure and described the model workflow in more detail:

For every lead, the model uses the same 12-month input of TWSA. In detail, for each lead i = 1, …, 6, the Grid–Mesh–Grid encoder takes the same 12-month input window and produces four latent features per grid cell for each of the 12 input months. The lead time enters through a learned lead-time embedding, which modulates the Mesh2Mesh and Mesh2Grid layers via FiLM. Therefore, the GNN generates different latent features for every lead time. The LSTM then processes a 12-step sequence. At each step, it receives the TWSA anomaly of that input month, the four latent features of that month for lead i, and the lead-time embedding. We also included an algorithm block in the Appendix (see Supplement Figure 9).

To answer your questions directly:

- What the LSTM receives at each lead: the 12-step sequence described above. Only the latent features and the lead-time embedding change with the lead.
- Whether the 12-month history is fixed: yes, it is identical for all six leads.
- What the residual is added to: before it was added to the last input TWSA. In the revised model, there is no residual, as we directly predict anomalies. The LSTM predicts the TWSA anomaly of the target month, and the training climatology of that month is added back (see our answer to comment 2a).
- Whether the recurrent state is carried across leads: no. The LSTM state starts from zero for every lead.

5. The role of seasonality in model performance

The seasonal cycle is kept in target, and the month of the year is given as an input to the model. This model will then reproduce seasonality very well and because of this the correlation will be very high in

the basins which are dominated by the seasonal cycles like Amazon (r=0.99, Fig. 4). In these cases, seasonal cycle accounts for most of the variance. This is mentioned in manuscript section 4.2 but the model's skill on anomaly components hasn't been reported. Please report anomaly correlation with the seasonal cycle removed from both forecast and target, using a training-period climatology, per basin and per lead, together with skill scores against climatology and persistence.

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 also adapted the metrics to only measure these anomaly components.

This is updated in Figures 4 and 5, showing the predictions for the 16 test basins. Also, we added NSE and anomaly correlation to the results tables and compared them against climatology and a persistence baseline.

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).

Closing assessment

I recommend major revision. The main issues are the need to provide evidence for the model's usefulness and performance in comparison to the baseline, uncertainty quantification of the model, clear description of the model and workflow in different mentioned parts. If these issues are addressed, the manuscript could make a valuable contribution to literature.

Thank you for your assessment! We addressed the issues with adding the persistence baseline and anomaly correlation and NSE as metrics; all models are now trained and evaluated over 8 seeds to differentiate between signal and seed noise; the model description and workflow are updated and clarified. We hope the revised manuscript meets your requests and can make a valuable contribution to the scientific community.

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. The results of the new experiments 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 detailed 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

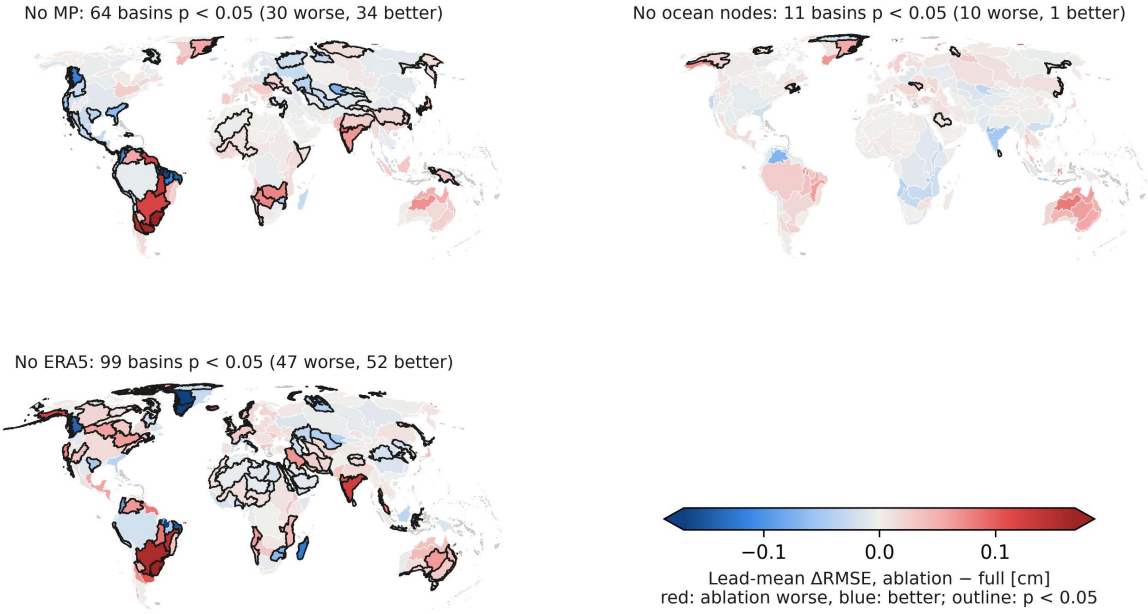


Figure 1 Basin-wise evaluation of ablation study. RMSE differences are averaged over all leads and seeds. Basins outlined in black show significant changes. The colour indicates if the ablation decreased RMSE (blue) or if it increased RMSE (red).

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.

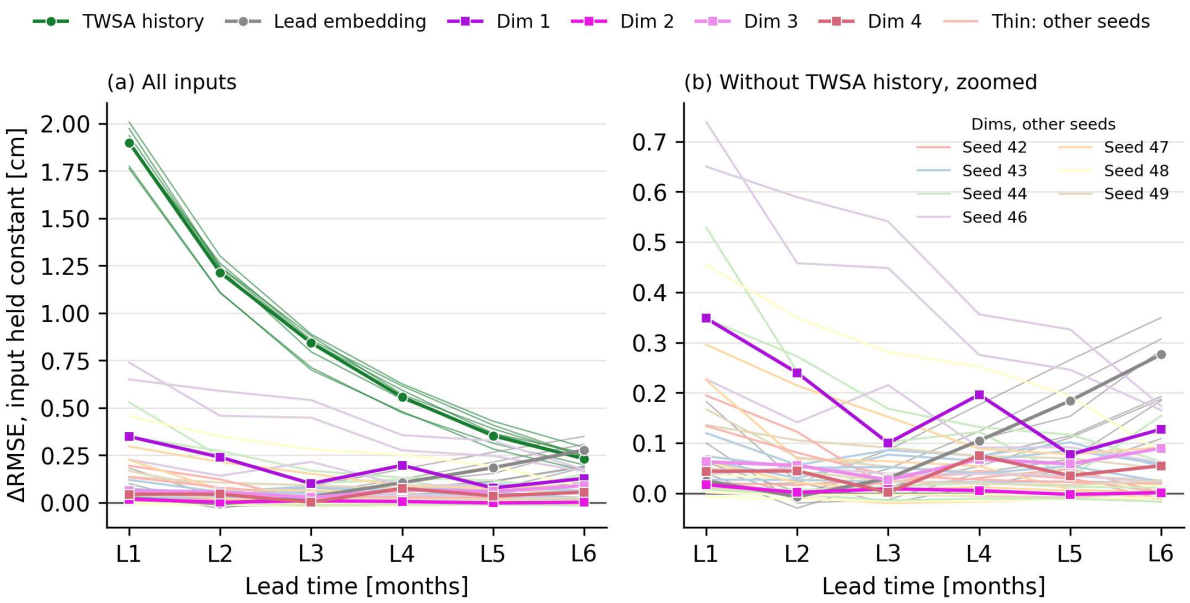


Figure 2 LSTM inputs substituted by mean values evaluated on the test period: change in area-weighted RMSE per lead against the full model with nothing removed. The thick lines show the results for seed 45. The thin lines show the results for all the other seeds.

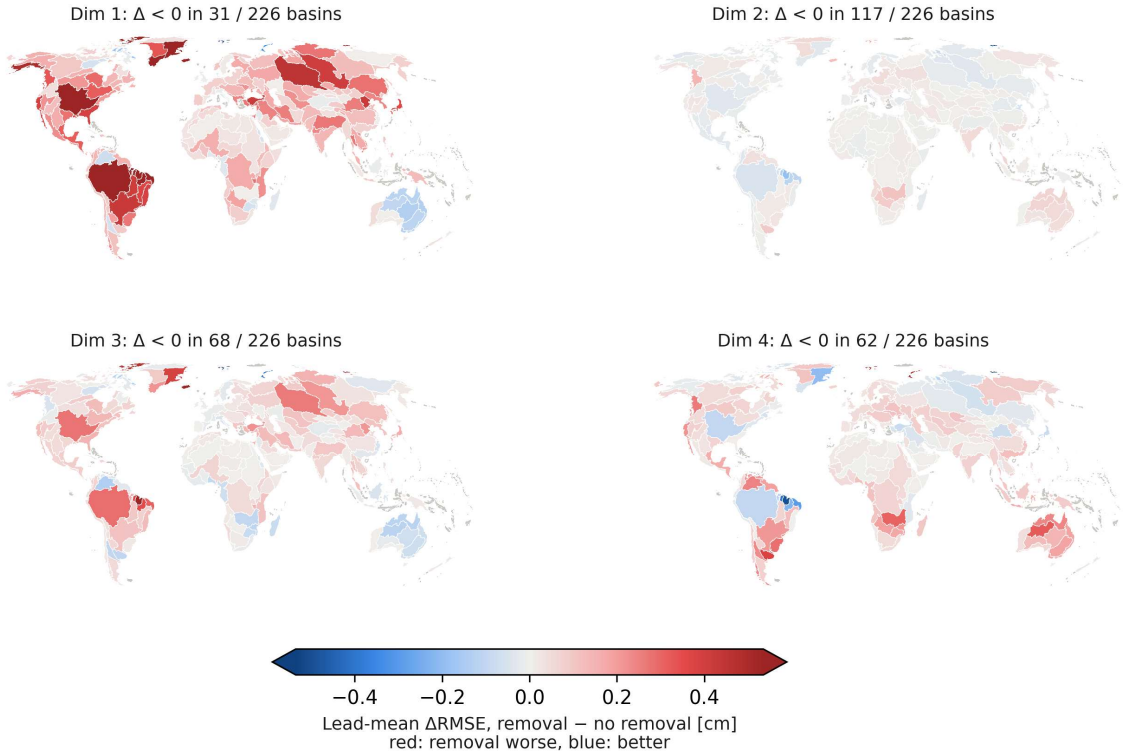


Figure 3 Basin-wise evaluation of latent-feature removal experiment evaluated for one seed (seed 45). RMSE differences are averaged over all leads. The colour indicates if the removal of a feature decreased RMSE (blue) or if it increased RMSE (red).

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

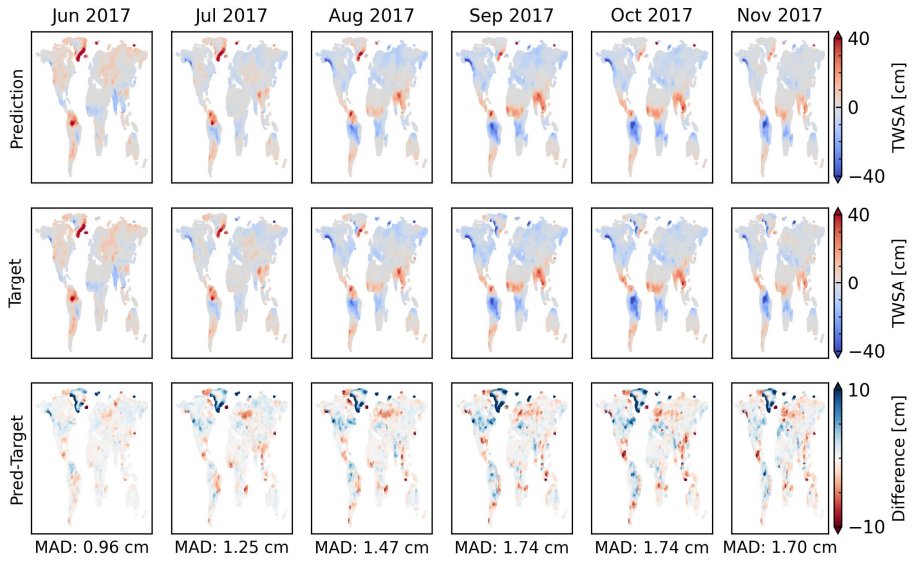


Figure 4 Six-month forecast (June 2017 - November 2017) from input time series June 2016 to May 2017. Rows show (top) model prediction from seed 45, (middle) target TWSA from the reconstruction, and (bottom) their difference, with the Mean Absolute Difference (MAD) indicated below each difference map.

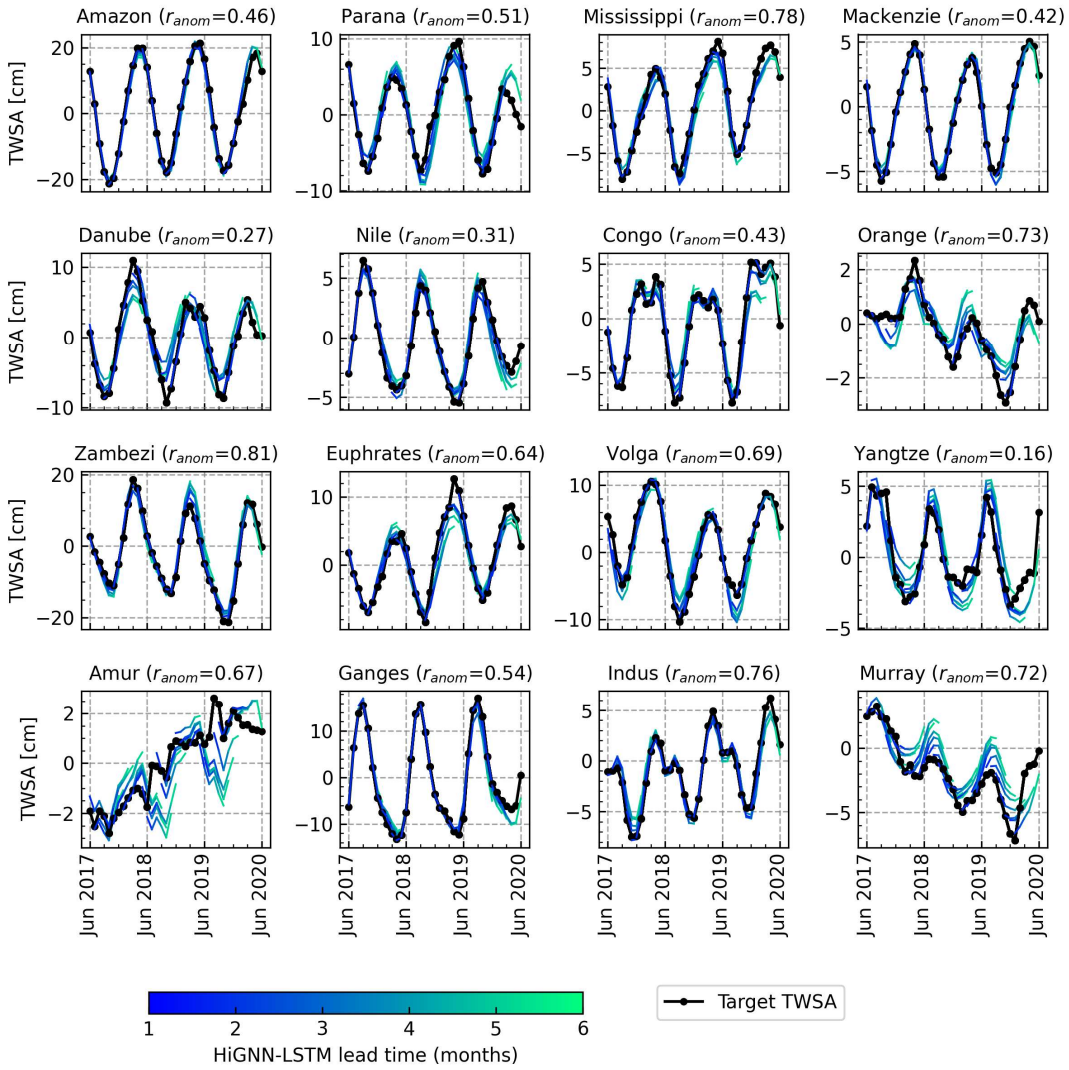


Figure 5 Predicted TWSA from seed 45 (blue) vs target TWSA (black) in 16 test basins together with anomaly correlation coefficient (r_{anom}) averaged over all forecast sequences for the given basin from June 2017 until June 2020.

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 $\text{Climatology}_{\text{GRACE}}$ 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
$\text{Climatology}_{\text{GRACE}}$	RMSE [cm]	6.832	6.819	6.806	6.775	6.743	6.724

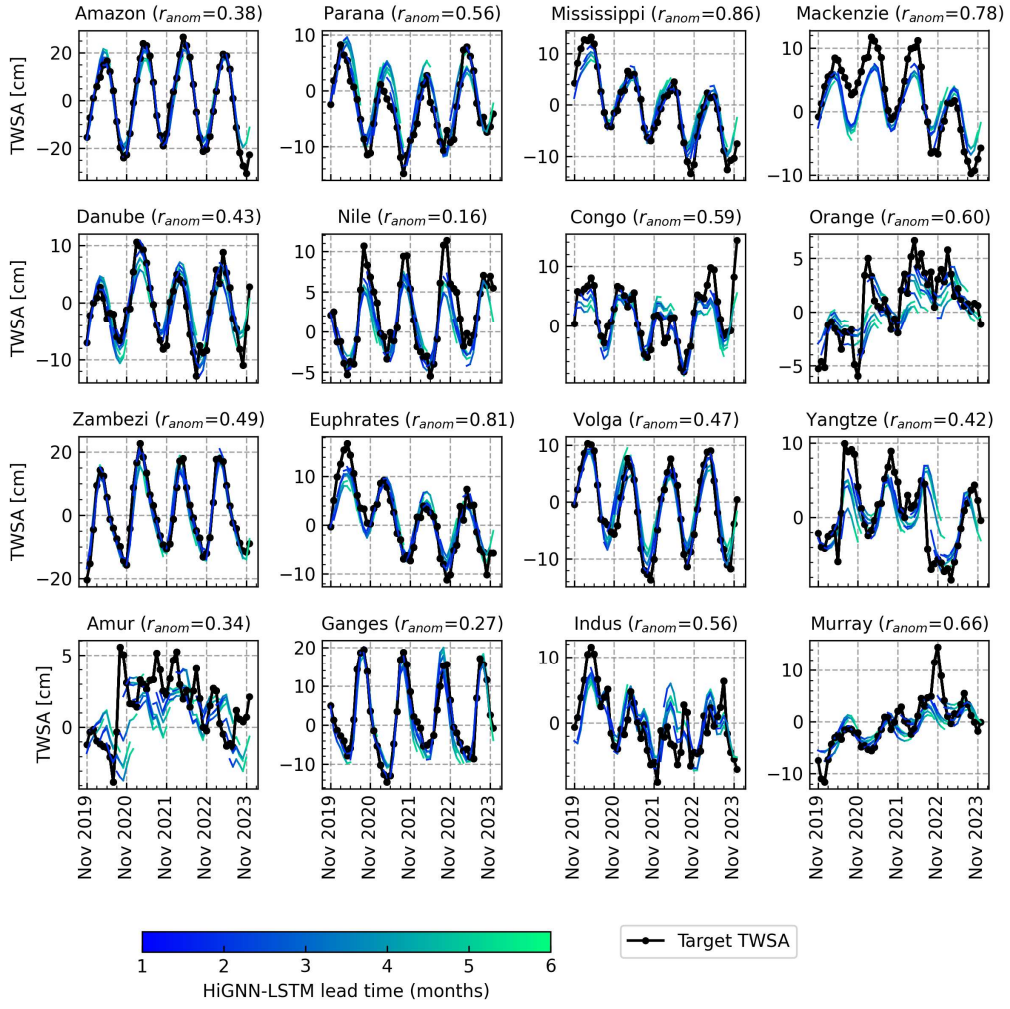


Figure 6 Predicted TWSA from seed 45 (blue) vs target GRACE-derived TWSA (black) in 16 test basins. r_anom: lead-averaged correlation of deseasonalised basin averages.

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).

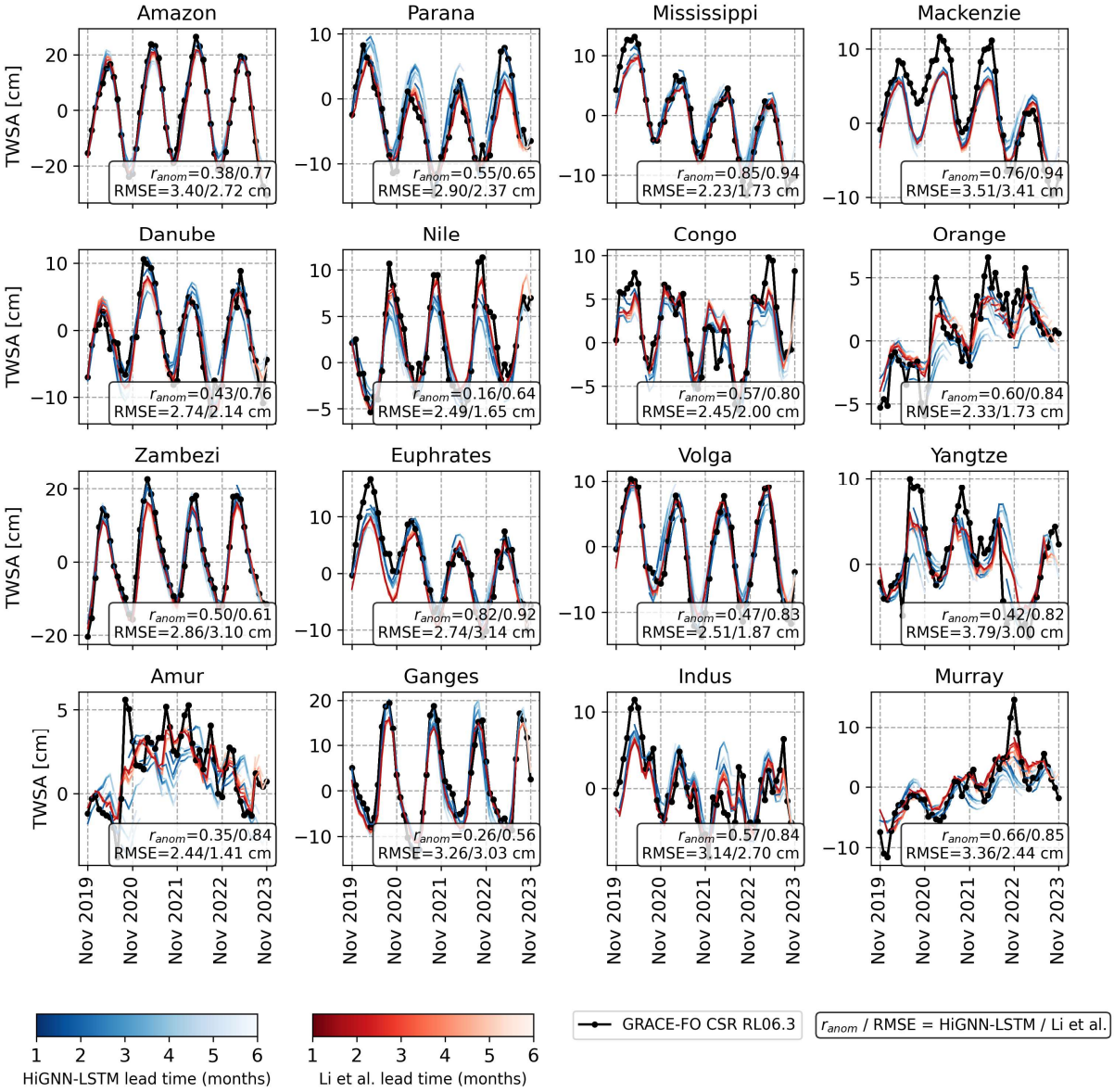


Figure 7 HiGNN-LSTM forecast of seed 45 (blue) and Li et al. (red) vs. GRACE-derived TWSA (black) in the 16 test basins. The lighter the line color of the forecasted TWSA, the longer the lead time. Anomaly correlation coefficient ($r_{_anom}$) and Root Mean Squared Error (RMSE) for the basins are given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

3.4 Graph construction: new ocean nodes

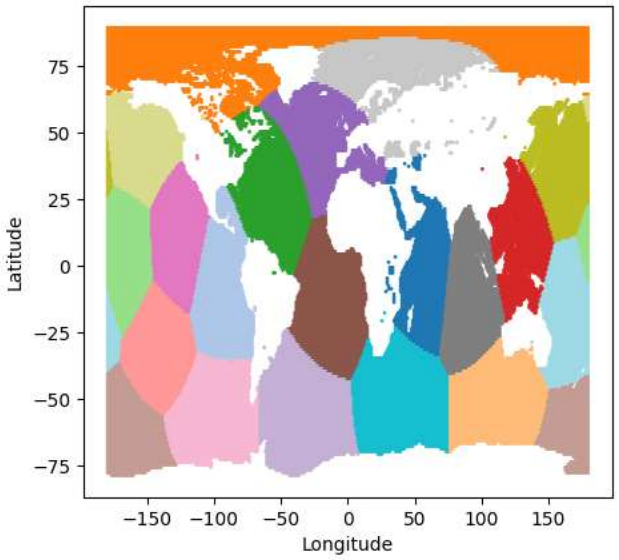


Figure 8 Ocean clusters as they are represented in the hierarchical graph. Each colour indicates a cluster formed by K-means clustering of grid-cell positions.

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.