Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Expert Complementarity and Residual Correction for Coastal Sea-Level Forecasting

Enze Cai
Wenzhou-Kean University
Yujing Wang
yujingwang2028@u.northwestern.edu
Northwestern University
Yiling Yang
Wenzhou-Kean University
Guanchao Tong
Wenzhou-Kean University

Research Article

Keywords:

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11531941/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Expert Complementarity and Residual Correction for Coastal Sea-Level Forecasting

Enze Cai[1,†], Yujing Wang[1,2,†,*], Yiling Yang[1,†], Guanchao Tong[1,†]

[1]Wenzhou-Kean University, Wenzhou, China
[2]Northwestern University, Evanston, USA
Emails: caie@kean.edu,
yujingwang2028@u.northwestern.edu,
yangyili@kean.edu,
tguancha@kean.edu
[†]These authors contributed equally to this work and share first authorship.
[*]Corresponding author: yujingwang2028@u.northwestern.edu

Abstract

Coastal water levels vary with tides, weather, and ocean conditions. We study forecasts of the non-tidal residual over the next 24 h. Two Graph WaveNet experts use the same historical inputs but learn different targets. One predicts the residual alone; the other also predicts current and wave states. HS-DT combines their forecasts with a fixed rule. C4 then adds a bounded correction and a predictive uncertainty estimate while keeping the experts frozen. We compare these methods with deep and linear baselines at seven northeastern U.S. stations. Further tests examine later periods and a separate ten-station network. HS-DT improves sequence forecasts in several temporal evaluations. C4 improves historical sequence accuracy and probability scores, with most of the point-forecast gain coming from residual correction. Its component effects change in the later backtest. VARX remains a strong baseline, and the tested physical priors give small, mixed gains. These results support expert combination and bounded correction, while showing that their benefits depend on the evaluation setting.

Keywords: coastal sea level; non-tidal residual; Graph WaveNet; forecast combination; residual correction

1 Introduction

Astronomical tides explain a large part of coastal water-level variation. Wind, air pressure, currents, and waves also affect the observed level. The difference between observed water level and predicted tide is the non-tidal residual. Forecasting this residual helps assess changes that a tide prediction alone cannot describe.

Nearby stations often respond to the same weather systems. A graph model can use these links together with each station's recent history. Models trained to predict different environmental states may learn different aspects of the same process. Their forecasts may therefore complement each other, even when they use the same inputs.

Combining forecasts does not remove every error. Two models may share a bias, and a fixed average cannot correct it. A learned correction may help, but extra flexibility can also make a forecast worse. The effects of expert combination, correction, and physical constraints need to be tested separately.

This study asks whether two graph-based experts can improve coastal residual forecasts through a fixed combination and a bounded correction. We also examine whether the gains persist across periods and stations. Deep models and a linear baseline provide reference forecasts. Controls with similar capacity help assess the contribution of physical priors.

We first compare the two experts and their fixed HS-DT combination. We then study C4, which adjusts the combination and learns an additional correction. Component tests identify which changes help. Temporal and regional tests examine the limits of those results.

2 Related work

Coastal water levels can be predicted with hydrodynamic models or statistical models. Hydrodynamic models describe tides, circulation, and storm surges, but require forcing data, boundary conditions, and bathymetry [7]. Machine learning has also been used for storm-surge forecasts and coastal extremes [8–12, 14]. Other studies address sea-level changes over months to years [13]. These tasks differ in timescale and target, so their accuracy scores are not directly comparable.

Graph neural networks use connections between locations to share information [1–3, 15]. They have improved forecasts in several spatial time-series tasks [16–18]. Graph WaveNet combines graph operations with dilated temporal convolutions [3]. Coastal station networks are much smaller than many traffic networks, and their inputs can change with weather and season. Performance therefore needs to be checked on coastal data.

Direct models predict several future steps together and avoid feeding their own predictions back into the input [19]. Multitask learning uses related targets to train shared features [20, 21]. For coastal prediction, current and wave targets may help a model learn information that differs from residual-only training. Whether those differences improve the final residual forecast remains an empirical question.

Forecast combination can reduce error when models make different mistakes [22, 23]. Mixtures of experts learn how to weight their outputs [24, 25]. Learned weights add flexibility, but may also overfit. A fixed combination provides a useful reference. Strong linear forecasts are another important comparison, as deep models do not always outperform them [6].

Forecast evaluation must account for the order and dependence of observations [26, 27]. A model that performs well on one period may not retain that advantage later. Different error measures also emphasize different failures [28, 29]. We therefore report paired differences, errors in physical units, and results across forecast leads.

Physical information can enter a model through inputs, equations, or penalties on its predictions [4, 5, 30–33]. Sparse coastal stations do not fully resolve the spatial derivatives and boundary conditions of a hydrodynamic model. The constraints studied here are simplified priors. We compare them with controls of similar capacity to separate their contribution from that of added model flexibility.

3 Data and study design

3.1 Stations, inputs, and target

The main network contains seven northeastern U.S. stations: New London, Montauk, Kings Point, The Battery, Sandy Hook, Atlantic City, and Cape May. The historical data cover 2023-2025. Water levels and tide predictions come from NOAA CO-OPS. ERA5 provides atmospheric inputs, Copernicus Marine provides currents and waves, and GEBCO provides bathymetry [34, 35, 39–44].

Let z denote observed water level and T the astronomical tide prediction at station i and time t. The target is the non-tidal residual:

$$\eta_{i,t} = z_{i,t} - T_{i,t}, \tag{1}$$

Each sample contains the preceding 24 h and predicts the next 24 hourly residuals in one step. The 34 input classes describe residual history, atmospheric forcing, currents, waves, local weather, depth, and station

location. Adding the predicted astronomical tide to a residual forecast gives total water level. All accuracy scores in this paper concern the residual.

Input windows end before the forecast target window. A sample is excluded when its residual history or future target window fails the validity rules.

All GWN variants use a fixed distance graph. Edge weights depend on station locations and are normalized by row. The graph is unchanged across seeds and variants within each network. Validation and test observations are not used to estimate its weights.

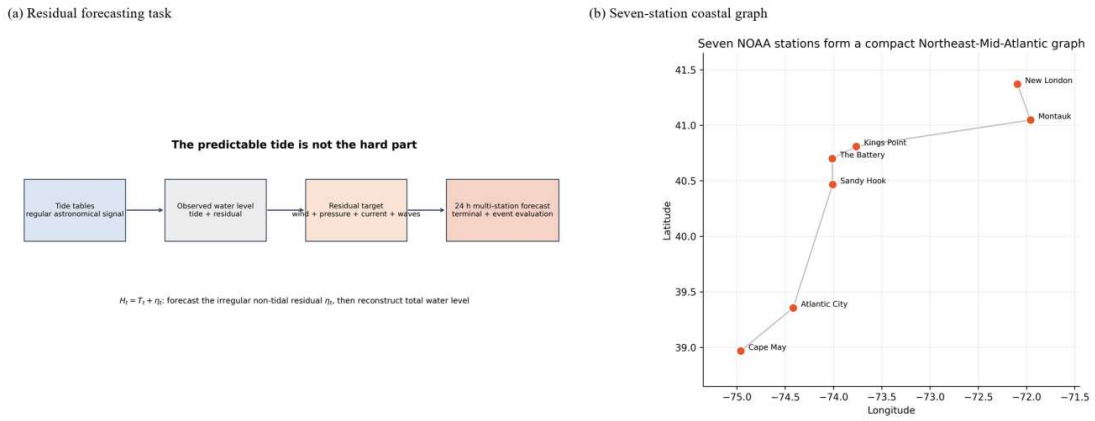


Figure 1: Study setting. The target is the non-tidal residual rather than tide-dominated total water level. Stations span the New Jersey, New York, and Connecticut coast.

3.2 Temporal and regional evaluations

We use four evaluation settings. They address different questions and are reported separately.

Tier A is the historical benchmark. The 2023-2025 series is divided chronologically into approximately 70% training, 15% validation, and 15% test data.

Tier B trains on 2023-2024, selects models on January-June 2025, and tests on July-December 2025.

Tier C applies the frozen GWN experts and HS-DT rule to later 2026 data. VARX was added after some outcomes from those periods had been seen.

Tier D tests the methods in the Delaware Bay-River network. The stations do not overlap the original network. Station selection, preprocessing, model definitions, metrics, and decision rules were fixed before external scoring. The split uses 2023-2024 for training, January-June 2025 for validation, and July-December 2025 for testing.

Reported scores are means of the five separate runs. Scores for an average of their prediction arrays are identified as ensemble scores.

3.3 External stations and data quality

The external network contains ten Delaware Bay-River stations: Lewes, Brandywine Shoal Light, Ship John Shoal, Reedy Point, Delaware City, Marcus Hook, Philadelphia, Bridesburg, Burlington, and Newbold.

A station could be excluded if water-level or tide coverage fell below 97% in any split, provided that at least eight stations remained. All ten stations had complete water-level, tide, and residual targets across the three splits.

The external network uses the same 34 input classes. Local weather records are less complete: wind-related variables have about 30% coverage and pressure-related variables about 71%. The ERA5, ocean, wave, and depth inputs are checked separately. Origins with incomplete residual history or target windows are excluded.

Three checks were specified before external scoring. Both experts should lead at least one forecast hour and one station. HS-DT should exceed both experts in sequence R^2. The ordering of VARX and HS-DT should change between Leads 1-8 and Lead 24, or reverse the original sequence ordering.

3.4 The overlapping 84-day backtest

A further backtest covers 2,016 hourly origins over 84 consecutive days. It compares frozen VARX, eta-only, multistate, HS-DT, C4, and physical-prior models.

Model parameters, checkpoints, hyperparameters, graph, scalers, feature order, calibration, and fusion rules stayed fixed. The input builder was corrected only to restore the preprocessing used by the checkpoints. A common target mask retains 323,832 valid station-lead cells and excludes 14,856 invalid cells. There are 1,374 origins with complete targets at every station and lead.

The primary comparison uses all common valid target cells. Sensitivity checks use complete origins or exclude the station with substantial missingness. The restored inputs reproduce the original 2025 H2 scores.

4 Models and controls

4.1 Baselines and GWN experts

The baseline experiments include GNN-BiGRU, DCRNN, and Graph WaveNet [2, 3, 36, 37]. Earlier GNN-BiGRU experiments varied the graph, inputs, supervision, and physical loss. Their scores compare complete model configurations; they do not isolate the effect of physical loss.

Graph WaveNet combines dilated temporal convolutions with graph operations [3]. We disable its adaptive adjacency and use the fixed distance graph. This gives the variants the same spatial support. FS-GWN in the external table denotes this fixed-graph implementation.

The two GWN experts share inputs, time splits, backbone settings, and training schedules. The eta-only expert predicts the non-tidal residual. The multistate expert also predicts current components and a wave-related state. Neither expert uses a physical loss. Both the output heads and training objectives differ, so this comparison does not isolate supervision as a single causal factor.

4.2 HS-DT-GWN expert combination

Horizon-Specialized Dual-Task Graph WaveNet (HS-DT-GWN) combines the two trained experts. It averages their forecasts at Leads 1–23 and uses the multistate forecast at Lead 24:

$$\widehat{\eta}_h^{HS-DT} = \begin{cases} \frac{1}{2}(\widehat{\eta}_h^{eta} + \widehat{\eta}_h^{multi}), & h = 1, \ldots, 23, \\ \widehat{\eta}_{24}^{multi}, & h = 24. \end{cases} \tag{2}$$

HS-DT adds no trainable parameters and leaves both experts unchanged. The rule was developed after the initial benchmark and kept fixed for subsequent evaluations. Its Tier-A score is therefore treated as a development result. At Lead 24, HS-DT uses the multistate prediction directly, so the two forecasts are identical.

4.3 C4 residual correction and uncertainty

C4 keeps both experts frozen. It adds a gate, a bounded correction, and a head that estimates predictive uncertainty. The gate and correction use the final hidden states of the two experts:

$$\begin{aligned} g_{i,h} &= \mathrm{clip}\left(g_h^{(0)} + 0.25 \tanh f_g([c_i^\eta, c_i^m])_h, 0, 1\right), \\ r_{i,h} &= s_\eta \, 0.25 \tanh f_r([c_i^\eta, c_i^m])_h, \end{aligned} \tag{3–4}$$

The gate sets the expert weights. Its adjustment is bounded around the HS-DT weight. The correction is limited by the training-set residual scale. The point forecast is

$$\widehat{\eta}_{i,h} = g_{i,h}\widehat{\eta}^{\eta}_{i,h} + (1 - g_{i,h})\widehat{\eta}^{m}_{i,h} + r_{i,h}. \tag{5}$$

The new point-prediction branches are initialized to match HS-DT. Training then tests whether adapting the weights or correcting the residual improves that starting forecast. Gate-only learns the weights. Correction-only keeps the fixed weights and learns the residual term. Full C4 learns both. The main C4 model has no physical loss; physical priors are evaluated separately.

A separate head predicts $\log \sigma_{i,h}$, where $\sigma_{i,h}$ is the Gaussian standard deviation. Its inputs are the point forecast, both expert outputs, the gate, and the forecast lead. Training uses Gaussian negative log-likelihood:

$$\mathcal{L}_{NLL} = \frac{1}{2N} \sum_{i,h} \left[\frac{(\eta_{i,h} - \widehat{\eta}_{i,h})^2}{\sigma_{i,h}^2} + 2 \log \sigma_{i,h} + \log(2\pi) \right]. \tag{6}$$

4.4 VARX-Ridge

VARX-Ridge predicts all 24 leads with a linear model. Its inputs include 24 h of residual history from all stations, the latest complete exogenous feature vector, and the 24-h feature mean. It uses the same input classes and time splits as the neural models. Validation MSE selects the ridge strength from $\{0.1, 1, 10, 100\}$. Both the historical and chronological protocols select $\alpha = 100$. The 2026 diagnostic retains the model fitted on 2023–2024 and does not refit it on later targets.

4.5 Physical priors and matched controls

Physical-prior experiments test a graph residual loss, ODE gating, ODE injection into both experts, and an ODE residual adapter. We assess the physical contribution against controls with matched trainable structure wherever available. Comparisons with different structures are reported separately because added capacity may explain part of the gain.

For an R^2-based comparison, the effect is

$$\Delta_{\mathrm{phys}} = M(\mathrm{backbone} + \mathrm{physics}) - M(\mathrm{matched\ control}), \tag{7}$$

A positive difference favors the physical-prior model, but it identifies a physical effect only when the control is appropriately matched. The ODE modules describe persistence, relaxation, and trend. The graph residuals relate sea level to current and wave states. These simplified constraints do not form a complete hydrodynamic model.

5 Evaluation

We describe expert differences at each lead using the R^2 difference, error correlation, and eta-only win probability. The R^2 difference is

$$S_h = R^2_{multi,h} - R^2_{eta,h}, \tag{8}$$

A positive value favors multistate. We also report error correlation, $\rho(h) = \mathrm{Corr}(e_\eta(h), e_m(h))$, and eta-only win probability, $W(h) = P(|e_\eta(h)| < |e_m(h)|)$. These summaries are calculated by lead, station, seed, and evaluation period.

We report sequence and Lead-24 residual R^2, MSE, RMSE, and MAE. Descriptive absolute-q95 scores assess large absolute residuals; they do not define an operational storm-surge event. For observations y_j and predictions $\widehat{y}_j$,

$$R^2 = 1 - \frac{\sum_j (y_j - \widehat{y}_j)^2}{\sum_j (y_j - \bar{y})^2}. \tag{9}$$

The scale-aware errors are

$$MSE = \frac{1}{N}\sum_j e_j^2, \quad RMSE = \sqrt{MSE}, \quad MAE = \frac{1}{N}\sum_j |e_j|, \tag{10}$$

with $e_j = \widehat{y}_j - y_j$. For probabilistic output we additionally use the Gaussian continuous ranked probability score

$$CRPS(\mu, \sigma; y) = \sigma\left[u(2\Phi(u) - 1) + 2\phi(u) - \pi^{-1/2}\right], \quad u = \frac{y - \mu}{\sigma}, \tag{11}$$

and empirical coverage $C_{1-\alpha} = N^{-1}\sum_j \mathbf{1}\{y_j \in [\mu_j - z_{1-\alpha/2}\sigma_j, \mu_j + z_{1-\alpha/2}\sigma_j]\}$. Here, $z_{1-\alpha/2}$ is the standard-normal quantile for a nominal $1 - \alpha$ interval. For serial dependence, origins are resampled in circular blocks of $B = 168$ h. If $\mathcal{I}^{(b)}$ is a bootstrap index sequence, the paired effect is

$$\Delta_m^{(b)} = m\left(y_{\mathcal{I}^{(b)}}, \widehat{y}^{(A)}_{\mathcal{I}^{(b)}}\right) - m\left(y_{\mathcal{I}^{(b)}}, \widehat{y}^{(B)}_{\mathcal{I}^{(b)}}\right), \tag{12}$$

We use 5,000 paired circular moving-block resamples to estimate uncertainty intervals [38]. The models are not retrained during resampling. The main block length is 168 h, with 84- and 336-h sensitivity checks. Each resample uses the same consecutive origin blocks for both models and retains all stations and leads together. The C4 analysis also shares these blocks across seeds. We report effect sizes, 95% intervals, and the number of positive seed differences. Exact one-sided Wilcoxon tests are descriptive.

A separate physics diagnostic considers high forcing. Its index combines wind, pressure tendency, current, wave flux, and a wave-setup proxy at the last input hour. The q95 threshold is estimated from training data. This subset differs from the absolute-residual q95 subset used in the forecast tables.

Predictive scales are fitted on validation data and then frozen. Eta-only, multistate, and HS-DT each use a constant Gaussian scale estimated from validation errors. C4 applies a validation-fitted multiplier to its predicted scale. These comparisons evaluate the complete predictive distributions. We added two controls using saved predictions. For ordinary ensembles, we evaluated all distinct-seed Eta+Eta and Multi+Multi pairs, and all distinct-seed Eta+Multi pairs. Each ordinary ensemble used equal weights at every lead. We averaged the scores across pairs, rather than combining all predictions into one larger ensemble.

For the scale control, both distributions used the same C4 mean prediction. The constant scale was the validation residual RMSE for each seed. The dynamic scale used the existing validation calibration. These post-hoc controls used 5,000 paired circular moving-block resamples with 168-h blocks and 84- and 336-h sensitivity checks. The same time blocks were shared across all pairs and seeds. The intervals describe time uncertainty conditional on the five saved runs; overlapping pairs were not treated as independent runs.

6 Results

6.1 Historical baseline comparison

Table 1 summarizes the historical benchmark. GWN scores higher than the tested physical-loss GNN-BiGRU configuration. This compares complete models and does not isolate the effect of physical loss. HS-DT has the highest sequence score among the deep forecasts in this table. VARX-Ridge has the highest sequence, Lead-24, and descriptive q95 scores.

Table 1: Historical Tier-A benchmark. Neural scores are means over five seeds. The q95 scores describe large absolute residuals.

Model	**Sequence R^2**	**Lead-24 R^2**	**q95 R^2**
Physical-loss GNN-BiGRU	0.6732	0.5553	0.5174
Eta-only GWN	0.7261	0.5825	0.6271
Multistate GWN, no physics	0.7250	0.6153	0.6179
HS-DT-GWN	0.7357	0.6153	0.6292
VARX-Ridge	0.7897	0.6442	0.7186

6.2 Differences between the two experts

In the 2025 H2 refit, eta-only has the higher point estimate at 13 of 24 leads, while multistate leads at the other 11. The five-seed mean error correlation is 0.932, with a range of 0.876–0.963. Relative performance changes with lead, without a clear monotonic pattern (Figure 2). Most uncertainty bands cross zero, and element-wise win probabilities stay near 0.5. This suggests modest error differences rather than a fixed division of work.

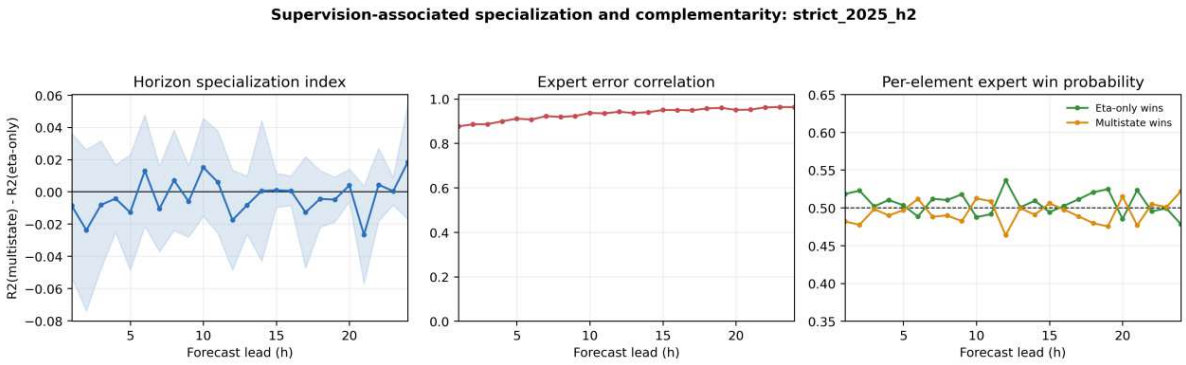


Figure 2: Expert differences in the 2025 H2 refit. Panels show the lead-wise R^2 difference, error correlation, and eta-only win probability. Shading shows variation across seeds.

The station-level comparison also varies with lead. Multistate has a higher mean Lead-24 score at all seven stations. Eta-only leads at several earlier hours. Neither expert is uniformly better across all station-lead combinations.

6.3 HS-DT across temporal evaluations

HS-DT reaches sequence $R^2 = 0.6563$ in the 2025 H2 refit. It improves on eta-only by 0.00915 and on multistate by 0.01243. All five seeds have positive differences in both comparisons. The respective 168-h intervals are [0.00141, 0.01955] and [0.00371, 0.02124]. The gain over multistate is 0.01525 in 2026 H1 (95% CI [0.00828, 0.02165]) and 0.01773 in July ([0.00907, 0.02839]). Both later comparisons also have five positive seeds.

These gains concern the full sequence. HS-DT and multistate give identical forecasts at Lead 24. The sequence improvements also do not establish better tail prediction: all deep models have negative q95 scores in July.

6.4 C4 component effects

Table 2 compares C4 with its matched HS-DT reference on the historical test set. The reference has sequence $R^2 = 0.731914$. It uses a different eta-only run set from Table 1, so 0.7357 is not the baseline for this ablation. Full C4 reaches 0.739055, an improvement of 0.007141 (95% CI [0.002121, 0.013309]). All five registered seeds improve in sequence R^2. MSE decreases by 2.66%; RMSE falls from 0.095161 to 0.093884 m and MAE from 0.065714 to 0.064805 m. The Lead-24 difference is 0.003780, but its interval includes zero ([-0.009380, 0.016587]).

Table 2: C4 component comparison on the Tier-A test set. Values are five-run means from the C4 run set. MSE is in m^2; RMSE and MAE are in m. Paired effects use the HS-DT reference saved with each component.

Model	Sequence R^2	Lead-24 R^2	Seq MSE	Seq RMSE	Seq MAE
HS-DT fixed	0.731914	0.615267	0.009056	0.095161	0.065714
Gate-only	0.730351	0.615965	0.009109	0.095438	0.065923
Correction-only	0.738353	0.617439	0.008838	0.094011	0.065009
Full C4	0.739055	0.619047	0.008815	0.093884	0.064805

Correction-only provides most of the sequence gain: +0.006441 (95% CI [0.001190, 0.012310]), with five positive seeds. Gate-only reduces sequence R^2 by 0.001561 (difference CI [-0.002573, -0.000615]). Full C4 exceeds Correction-only by 0.000702 in mean sequence R^2. Table 2 and Figure 3 therefore identify residual correction as the main source of historical improvement. Adapting the weights alone does not help in this test.

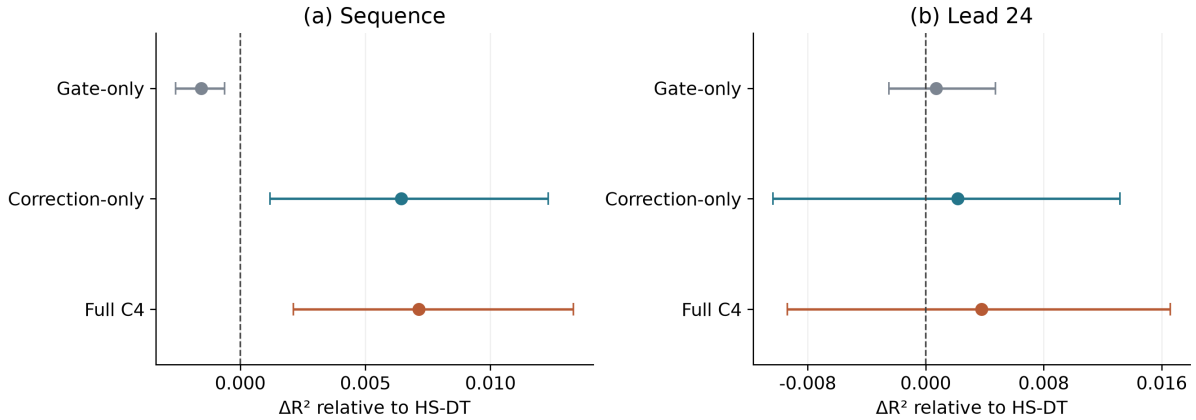


Figure 3: C4 component effects relative to their saved HS-DT references. Error bars are paired 95% intervals from 5,000 circular moving-block replicates with 168-h blocks.

6.5 Predictive uncertainty and block-length checks

The Full C4 sequence interval stays above zero with 84-, 168-, and 336-h blocks (Table 3). These are sensitivity checks on the same window, not independent tests. The Lead-24 intervals include zero. C4 has a CRPS of 0.04704 m, compared with 0.04950 m for HS-DT. Its 95% coverage is 94.00%, with a mean width of 0.31695 m. HS-DT gives 93.42% coverage and a width of 0.34859 m. C4 coverage at the 50% level is 53.81% (Figure 4).

C4 produces a narrower 95% interval with slightly higher coverage. Its 50% coverage is also closer to the nominal level, although calibration remains imperfect. Both the point forecast and predicted scale change, so the improvement cannot be attributed to the uncertainty head alone. A separate control held the C4 mean prediction fixed. On the seven-station test, the dynamic scale reduced CRPS from 0.048927 to 0.047039 m. The reduction was 0.001888 m (95% CI [0.001515, 0.002261]). On the ten-station test, CRPS fell from 0.064824 to 0.061739 m, a reduction of 0.003085 m (95% CI [0.002526, 0.003627]). All five seeds improved in both regions. Gaussian NLL also decreased. The mean 95% interval width fell by 11.05% and 14.56%, respectively, while the coverage differences were not statistically clear. Thus, the dynamic scale improves probability scores beyond the change in the C4 mean prediction, although calibration remains imperfect.

Table 3: C4 block-length sensitivity and calibration. R^2 differences use paired time blocks. Positive seeds count separate training runs. Calibration scores use validation-fitted scales.

Check	Setting	Mean ΔR^2	95% CI low	95% CI high	Positive seeds
C4 - HS-DT	Sequence, 84 h	0.00714	0.00239	0.01261	5/5
C4 - HS-DT	Sequence, 168 h	0.00714	0.00212	0.01331	5/5
C4 - HS-DT	Sequence, 336 h	0.00714	0.00194	0.01421	5/5
C4 - HS-DT	Lead 24, 168 h	0.00378	-0.00938	0.01659	4/5
Calibration	C4 NLL / CRPS	-1.098 / 0.0470	—	—	—
Calibration	Coverage 50% / 95%	0.538 / 0.940	—	—	—

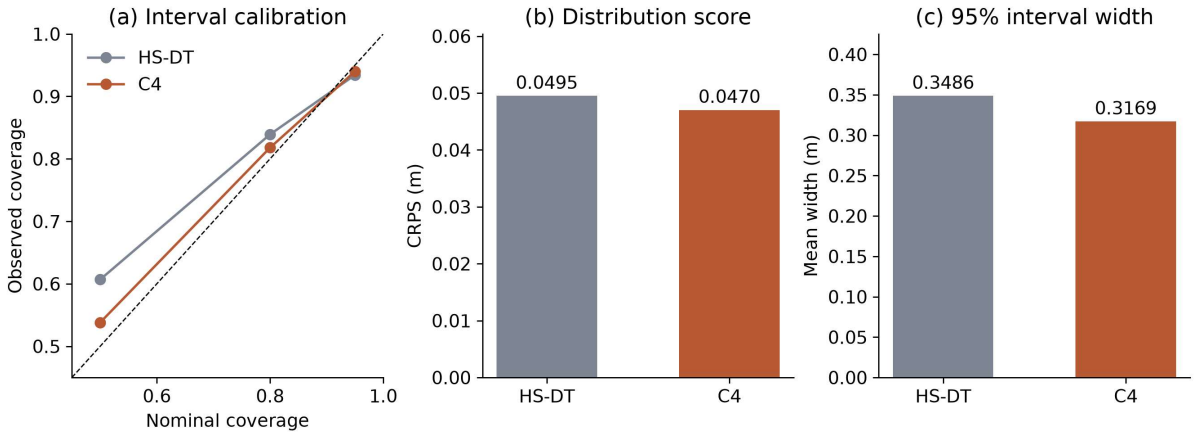


Figure 4: Predictive uncertainty after validation calibration. Panel (a) compares nominal and observed coverage. Panels (b) and (c) show CRPS and mean 95% interval width. HS-DT uses a constant Gaussian scale.

6.6 Expert combination in the Delaware network

All ten external stations pass the data-quality rule. Table 4 reports their 2025 H2 results. VARX has the highest sequence R^2, at 0.7657. HS-DT reaches 0.6803 ± 0.0058. Its gain over eta-only is 0.01634, with five positive seeds and a 168-h interval of [0.00949, 0.02420]. The gain over multistate is 0.00453, with four positive seeds, but its interval includes zero ([-0.00144, 0.01012]).

The specific station-and-lead pattern from the original network is not reproduced externally. Eta-only leads at Leads 1 and 5. Multistate leads at the other 22 hours and at all ten stations in sequence R^2.

Figure 5(b) scores an ensemble formed by averaging the five neural prediction arrays. HS-DT trails VARX by 0.11845 in mean R^2 over Leads 1–8, but exceeds it by 0.01334 at Lead 24. Table 4 instead averages separate-run scores, where HS-DT is slightly below VARX at Lead 24. Only two individual runs

Table 4: Delaware Bay-River results. Neural values are the mean and standard deviation of five separate-run scores. VARX is deterministic. These are not prediction-ensemble scores.

Model	Sequence R^2	Lead-24 R^2	Lead-24 RMSE (m)
VARX–Ridge	0.7657	0.5763	0.1405
Eta-only FS-GWN	0.6639 ± 0.0115	0.5217 ± 0.0207	0.1493 ± 0.0032
Multistate FS-GWN	0.6757 ± 0.0061	0.5743 ± 0.0105	0.1409 ± 0.0017
HS-DT-GWN	0.6803 ± 0.0058	0.5743 ± 0.0105	0.1409 ± 0.0017

exceed VARX there. Two of the three point-estimate checks specified before external scoring are satisfied. This does not confirm the original specialization pattern or establish a statistically clear gain over multistate.

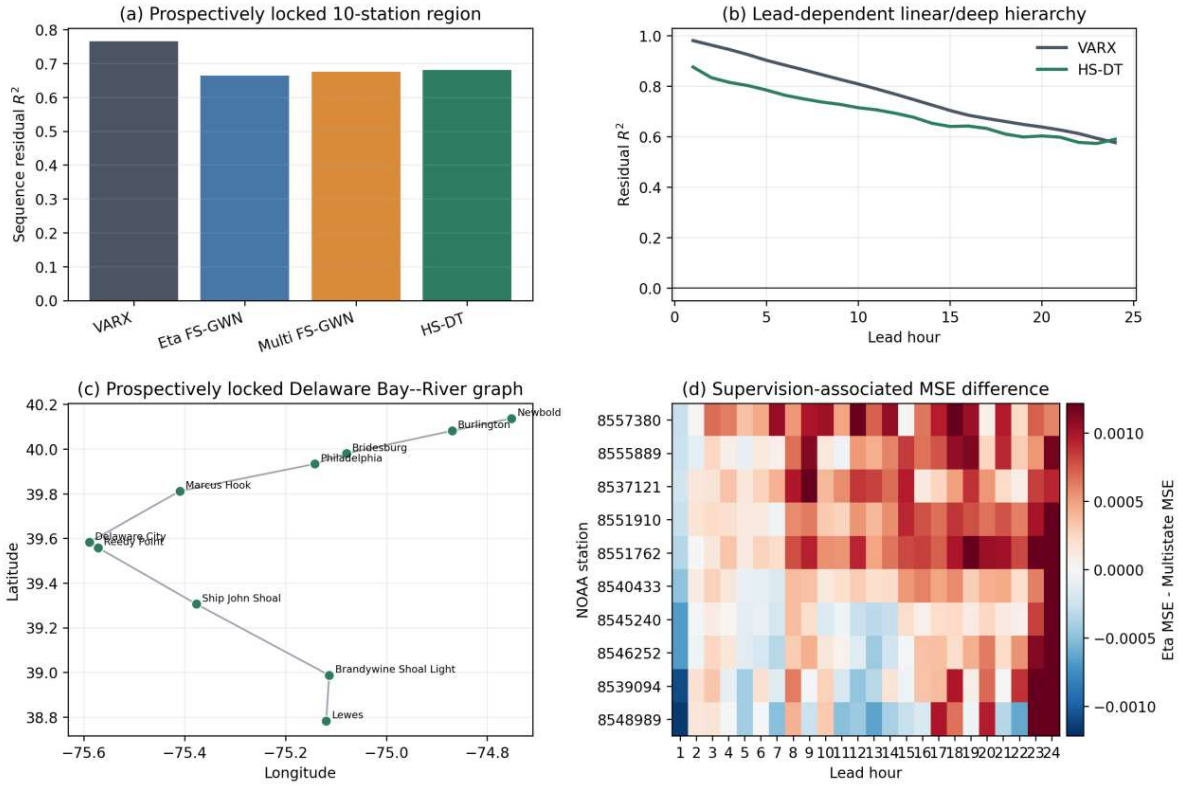


Figure 5: External network evaluation. Panel (a) shows separate-run mean sequence scores. Panel (b) scores the five-run prediction ensemble at each lead. Panel (c) shows the ten stations. Panel (d) shows eta-only MSE minus multistate MSE; positive values favor multistate.

We also evaluated C4 in the ten-station Delaware network (Table 5). C4 increased mean sequence R^2 from 0.6803 to 0.6880, with positive differences in all five seeds. The paired difference was 0.00778, with a 95% interval of [0.00229, 0.01462] from 5,000 moving-block bootstrap replicates using 168-h blocks. The sequence intervals remained above zero with 84-h and 336-h blocks.

Lead-24 performance was similar for the two models. The mean difference was -0.00021, with a 95% interval of [-0.00971, 0.00846]. Thus, the additional correction improved sequence accuracy in this network, but did not show a clear benefit at Lead 24. This analysis was added after the original external results had been examined and is reported separately from the predeclared replication checks.

Table 5: Additional C4 evaluation in the ten-station Delaware network. Values are the mean and standard deviation of five separate-run scores. This analysis was added after the original external evaluation.

Model	Sequence R^2	Lead-24 R^2
HS-DT-GWN	0.6803 ± 0.0058	0.5743 ± 0.0105
C4	0.6880 ± 0.0052	0.5741 ± 0.0102

6.7 Linear and deep forecasts across periods

Table 6 and Figure 6 compare the temporal results. VARX exceeds HS-DT by 0.0582 in sequence R^2 in 2025 H2. In the later diagnostic periods, HS-DT exceeds VARX by 0.1070 in 2026 H1 and 0.0441 in July. VARX remains stronger at early leads. The curves cross around Leads 9–13, and VARX has negative Lead-24 R^2 in both later periods. As described in Section 3.2, these VARX comparisons were designed after the periods had been examined and are treated as diagnostic evidence.

Ordinary ensemble controls. On the historical seven-station test, mean sequence R^2 was 0.730785 for Eta+Eta, 0.733883 for Multi+Multi, and 0.733394 for Eta+Multi. The Eta+Multi minus Multi+Multi difference was -0.000489 (95% CI [-0.004353, 0.004188]). On the ten-station test, the corresponding scores were 0.676089, 0.685877, and 0.681830. The cross-supervision difference from Multi+Multi was -0.004047 (95% CI [-0.006405, -0.001219]). Using the HS-DT rule with distinct-seed experts also did not outperform Multi+Multi. These controls do not establish an advantage of cross-supervision combination over ordinary same-supervision ensembles.

Table 6: Sequence R^2 across periods. Neural entries are means of five separate-run scores. The 2026 experts were frozen before evaluation; VARX was added later as a diagnostic comparison.

Period	Evidence status	Eta-only GWN	Multistate GWN	HS-DT-GWN	VARX-Ridge
2025 H2	chronological refit	0.6472	0.6439	0.6563	0.7146
2026 H1	frozen deep; VARX diagnostic	0.5185	0.5232	0.5385	0.4315
2026 July	frozen deep; VARX diagnostic	0.3915	0.3916	0.4093	0.3653

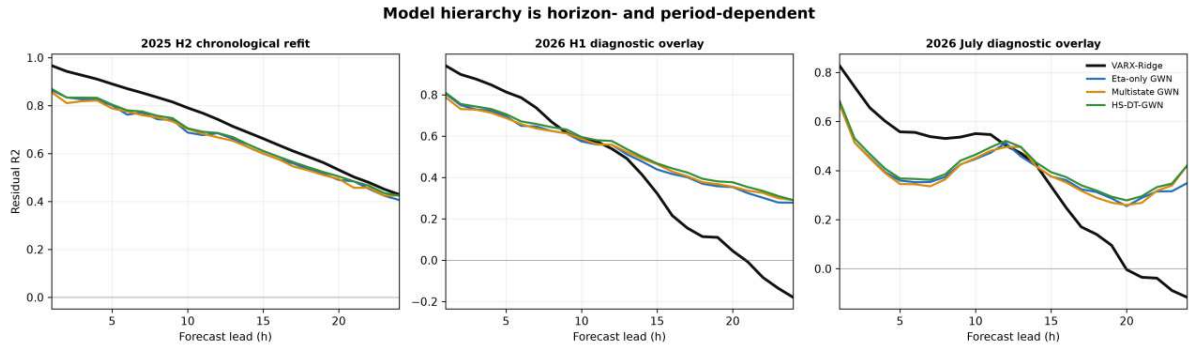


Figure 6: R^2 by forecast lead and period. VARX leads in 2025 H2 and at early leads in 2026. Its performance falls at longer leads in the later periods. These curves describe temporal changes; the 2026 VARX comparisons were not prospectively specified.

In the 84-day backtest, VARX has higher sequence R^2 than HS-DT (0.3439 versus 0.2884). HS-DT minus VARX changes from about -0.4049 over the early leads to +0.6363 at Lead 24. The terminal scores are 0.4107 and -0.2256. Average sequence skill and terminal skill therefore give different rankings in this window.

6.8 The 84-day backtest

This backtest assesses the stability of frozen models; it is not a new model-selection stage. With the common observed-target mask, HS-DT exceeds multistate by 0.015298 in sequence R^2 (95% CI [0.008513, 0.024079]). Sequence RMSE decreases by 0.000954 m ([0.000539, 0.001417]), and both differences are positive in all five seeds. The gain remains positive when using complete origins, excluding the station with substantial missingness, or changing the block length. These checks use the same 84-day window and are not independent confirmations.

Table 7: Paired effects in the overlapping 84-day backtest. Intervals use 5,000 moving-block replicates with 168-h blocks. Positive R^2 differences and positive RMSE reductions favor the first model.

Comparison	Metric	Mean Δ	95% CI	Positive seeds
HS-DT–Multistate	Sequence R^2	+0.015298	[0.008513, 0.024079]	5/5
HS-DT–Multistate	RMSE improvement	+0.000954	[0.000539, 0.001417]	5/5
Gate–HS-DT	Sequence R^2	-0.002303	[-0.004073, -0.000646]	1/5
Correction–HS-DT	Sequence R^2	+0.001895	[-0.006681, 0.009851]	2/5
Full–Correction	Sequence R^2	+0.006883	[0.003395, 0.009811]	5/5

C4 component effects differ from those in the historical test (Table 7). Gate-only again reduces sequence skill. Correction-only has a small positive difference from HS-DT, but its interval includes zero. Full C4 exceeds Correction-only. This comparison alone does not establish a significant Full C4 gain over HS-DT in the 84-day window. Fixed expert combination is more consistent across these tests than either adaptive component used alone.

6.9 Physical priors and matched controls

Figure 7 compares the physical-prior variants. In the chronological GNN-BiGRU test, direct physical loss reduces both sequence and Lead-24 R^2. Joint ODE injection improves GWN sequence R^2 by 0.00444 over its matched control, with four positive seeds, but reduces Lead-24 skill. The dual-expert ODE comparison shows the same trade-off: a sequence gain of 0.00221 and a negative terminal effect. Adding direct physical loss after ODE injection makes little further difference.

In the retrospective high-forcing q95 subset, the ODE correction lowers MSE by 1.06% relative to plain HS-DT, with five positive seeds. The reduction falls to 0.30% against a zero-signal adapter and 0.07% against a persistence adapter. Both matched comparisons have four positive seeds. Much of the gain over plain HS-DT is therefore shared by the control adapters.

The 84-day matched physics effects are close to zero. The sequence difference is +0.000056 (95% CI [-0.000062, 0.000194]; four positive seeds). The Lead-24 difference is -0.000128 ([-0.000514, 0.000235]; two positive seeds). These tests give no consistent positive effect across architectures and settings.

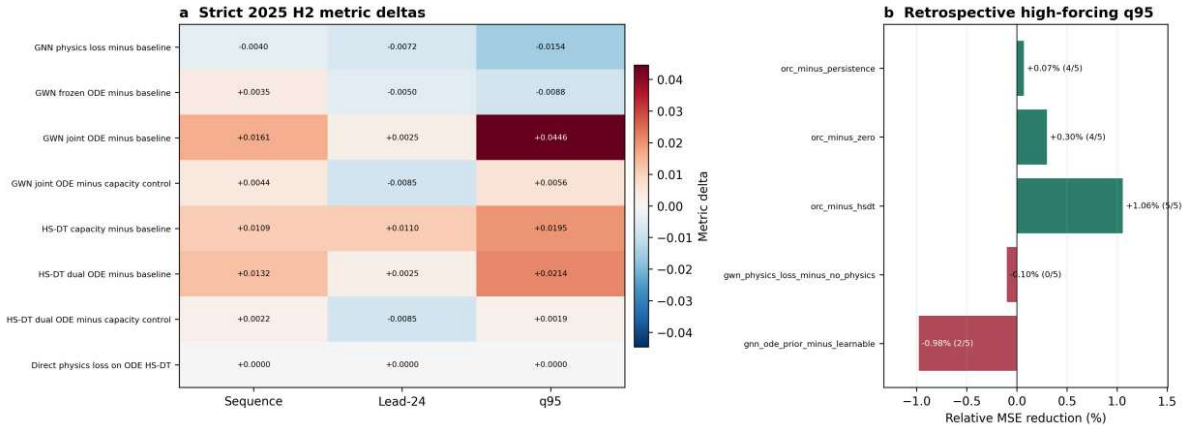


Figure 7: Physical-prior comparisons. Panel (a) shows R^2 differences in the chronological refit. Panel (b) shows MSE reductions in the retrospective high-forcing subset. Rows with unmatched capacity are displayed separately from matched controls.

7 Discussion

7.1 Expert combination and residual correction

The experts make similar but non-identical errors. Their fixed combination improves sequence forecasts in several temporal evaluations. The external gain is clear against eta-only, but smaller and uncertain against multistate. This supports useful complementarity without implying a universal division of work between the experts.

C4 addresses the errors left after combination. Its historical component tests point to the bounded correction as the main source of improvement. Gate-only adaptation does not help in that setting. The component effects change in the 84-day backtest, so the historical result does not establish that the same component will help in every period.The Delaware extension also shows a sequence gain for C4, while Lead-24 performance remains similar to HS-DT. The correction is therefore useful in some settings, but its benefit is not uniform across forecast leads.

The physical comparisons also show why matched controls matter. An added module can help even without the proposed physical signal. The small or mixed matched effects apply to the simplified priors tested here. They do not show that physical information is generally unhelpful. The ordinary ensemble controls show that the observed gains over individual experts do not establish a unique benefit from different supervision. In contrast, the fixed-mean scale control supports the value of C4's dynamic scale for probabilistic prediction in both evaluated regions.

7.2 Implications for model selection

Model choice depends on the forecast lead and period. VARX is strong in the historical and external sequence comparisons. The later diagnostics show cases where its long-lead forecasts deteriorate. A single overall score would hide these differences.

C4 improves historical errors and probability scores, but the absolute point-forecast gain is modest. It is best viewed as an extension of the deep experts, rather than a replacement for every baseline. Forecast lead, error magnitude, calibration, and computational cost all matter when choosing a model for use.

7.3 Limitations

The two station networks are located in the same broad Atlantic coastal region and use the same external test calendar. The additional C4 experiment shows a sequence gain in the Delaware network, but it was conducted after the original external results had been examined. It therefore provides supporting evidence

rather than an independent confirmation. Some ocean inputs have a lower time resolution than the hourly forecast grid, and several local weather records contain gaps. Historical alignment and interpolation may therefore be more favorable than the information available in real time. The Tier-A results should be understood as retrospective benchmark results. The 2025 H2 period was viewed during development, and the 84-day backtest overlaps earlier analyses. These experiments are therefore not a fully untouched temporal confirmation of C4. A longer new-period test with frozen preprocessing would provide stronger evidence. Testing C4 in a more distant coastal region would help assess whether its sequence gain extends beyond the two networks studied here.

8 Reproducibility

The experiment records contain the model settings, data splits, checkpoint and prediction hashes, calibration choices, and paired comparisons. Chronological refits estimate preprocessing statistics from their training periods, and historical and chronological results are analyzed separately. During the final audit, derived weather fields were checked before external scoring. The 84-day inputs were rebuilt with the preprocessing used by the frozen checkpoints, and earlier intermediate outputs were excluded. Confidence intervals were recomputed with consecutive time blocks to account for serial dependence. Checkpoints are selected with validation Gaussian NLL, VARX settings with validation MSE, and calibration uses validation data. Paired comparisons use the corresponding frozen HS-DT predictions and the same target array. The 84-day evaluation uses a common mask of observed targets and does not impute missing values. Tables distinguish averages across separate runs from scores computed after averaging predictions.

9 Conclusions

Different supervision targets produce modest differences between the two Graph WaveNet experts. Their fixed HS-DT combination improves sequence forecasts in several temporal evaluations. The gain over eta-only is clearer than the gain over multistate in the external network. C4 improves historical sequence accuracy and probabilistic scores. Residual correction accounts for most of its historical point-forecast improvement, while the contribution of the gate is limited. The component effects change across evaluation periods. VARX remains a strong baseline, and the matched physical-prior comparisons show small and mixed effects. Overall, the results support fixed expert combination and bounded residual correction under the tested settings. They do not show that one model is best for every forecast lead, period, or region.

Data availability

Water levels, tide predictions, and local observations are available from NOAA CO-OPS and NDBC. ERA5 is available from the Copernicus Climate Data Store, ocean and wave products from Copernicus Marine, and bathymetry from GEBCO [34, 35, 39–44]. Experiment records contain the result tables, figure sources, and revised bootstrap summaries. Source-product terms apply to derived data.

Declaration of competing interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Removed for anonymous review.

Declarations

Funding

The authors received no specific funding for this research.

Ethics Declaration

Ethics declaration: not applicable.

References

[1] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: International Conference on Learning Representations, 2017.

[2] Y. Li, R. Yu, C. Shahabi, Y. Liu, Diffusion convolutional recurrent neural network: Data-driven traffic forecasting, in: International Conference on Learning Representations, 2018.

[3] Z. Wu, S. Pan, G. Long, J. Jiang, C. Zhang, Graph WaveNet for deep spatial–temporal graph modeling, in: Proceedings of IJCAI, 2019: pp. 1907–1913. 10.24963/ijcai.2019/264.

[4] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686–707. 10.1016/j.jcp.2018.10.045.

[5] G.E. Karniadakis, I.G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, L. Yang, Physics-informed machine learning, Nature Reviews Physics 3 (2021) 422–440. 10.1038/s42254-021-00314-5.

[6] A. Zeng, M. Chen, L. Zhang, Q. Xu, Are transformers effective for time series forecasting?, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2023: pp. 11121–11128. 10.1609/aaai.v37i9.26317.

[7] R.M. Ponte, M. Carson, M. Cirano, et al., Towards comprehensive observing and modeling systems for monitoring and predicting regional to coastal sea level, Frontiers in Marine Science 6 (2019). 10.3389/fmars.2019.00437.

[8] A. Salighehdar, Z. Ye, M. Liu, I. Florescu, A.F. Blumberg, Ensemble-based storm surge forecasting models, Weather and Forecasting 32 (2017) 1921–1936. 10.1175/WAF-D-17-0017.1.

[9] W. Xie, G. Xu, H. Zhang, C. Dong, Developing a deep learning-based storm surge forecasting model, Ocean Modelling 182 (2023) 102179. 10.1016/j.ocemod.2023.102179.

[10] L.-H. Bai, H. Xu, Accurate storm surge forecasting using the encoder–decoder long short term memory recurrent neural network, Physics of Fluids 34 (2022). 10.1063/5.0081858.

[11] S. Giaremis, N. Nader, C. Dawson, C. Kaiser, E. Nikidis, H. Kaiser, Storm surge modeling in the AI era: Using LSTM-based machine learning for enhancing forecasting accuracy, Coastal Engineering 191 (2024) 104532. 10.1016/j.coastaleng.2024.104532.

[12] Y. Qin, C. Su, D. Chu, J. Zhang, J. Song, A review of application of machine learning in storm surge problems, Journal of Marine Science and Engineering 11 (2023) 1729. 10.3390/jmse11091729.

[13] V. Nieves, C. Radin, G. Camps-Valls, Predicting regional coastal sea level changes with machine learning, Scientific Reports 11 (2021). 10.1038/s41598-021-87460-z.

[14] N. Bruneau, J. Polton, J. Williams, J. Holt, Estimation of global coastal sea level extremes using neural networks, Environmental Research Letters 15 (2020) 074030. 10.1088/1748-9326/ab89d6.

[15] W. Jiang, J. Zhang, Y. Li, D. Zhang, G. Hu, H. Gao, Z. Duan, Advancing storm surge forecasting from scarce observation data: A causal-inference based spatio-temporal graph neural network approach, Coastal Engineering 190 (2024) 104512. 10.1016/j.coastaleng.2024.104512.

[16] B. Yu, H. Yin, Z. Zhu, Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018: pp. 3634–3640. 10.24963/ijcai.2018/505.

[17] Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, C. Zhang, Connecting the dots: Multivariate time series forecasting with graph neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2020: pp. 753–763. 10.1145/3394486.3403118.

[18] M. Jin, H.Y. Koh, Q. Wen, D. Zambon, C. Alippi, G.I. Webb, I. King, S. Pan, A survey on graph neural networks for time series: Forecasting, classification, imputation, and anomaly detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (2024) 10466–10485. 10.1109/TPAMI.2024.3443141.

[19] B. Lim, S.O. Arik, N. Loeff, T. Pfister, Temporal fusion transformers for interpretable multi-horizon time series forecasting, International Journal of Forecasting 37 (2021) 1748–1764. 10.1016/j.ijforecast.2021.03.012.

[20] R. Caruana, Multitask learning, Machine Learning 28 (1997) 41–75. 10.1023/A:1007379606734.

[21] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: pp. 7482–7491. 10.1109/CVPR.2018.00781.

[22] J.M. Bates, C.W.J. Granger, The combination of forecasts, Journal of the Operational Research Society 20 (1969) 451–468. 10.1057/jors.1969.103.

[23] R.T. Clemen, Combining forecasts: A review and annotated bibliography, International Journal of Forecasting 5 (1989) 559–583. 10.1016/0169-2070(89)90012-5.

[24] R.A. Jacobs, M.I. Jordan, S.J. Nowlan, G.E. Hinton, Adaptive mixtures of local experts, Neural Computation 3 (1991) 79–87. 10.1162/neco.1991.3.1.79.

[25] P. Montero-Manso, G. Athanasopoulos, R.J. Hyndman, T.S. Talagala, FFORMA: Feature-based forecast model averaging, International Journal of Forecasting 36 (2020) 86–92. 10.1016/j.ijforecast.2019.02.011.

[26] V. Cerqueira, L. Torgo, I. Mozetic, Evaluating time series forecasting models: An empirical study on performance estimation methods, Machine Learning 109 (2020) 1997–2028. 10.1007/s10994-020-05910-7.

[27] H. Hewamalage, K. Ackermann, C. Bergmeir, Forecast evaluation for data scientists: Common pitfalls and best practices, Data Mining and Knowledge Discovery 37 (2023) 788–832. 10.1007/s10618-022-00894-5.

[28] R.J. Hyndman, A.B. Koehler, Another look at measures of forecast accuracy, International Journal of Forecasting 22 (2006) 679–688. 10.1016/j.ijforecast.2006.03.001.

[29] S. Makridakis, E. Spiliotis, V. Assimakopoulos, The M4 competition: Results, findings, conclusion and way forward, International Journal of Forecasting 34 (2018) 802–808. 10.1016/j.ijforecast.2018.06.001.

[30] A. Karpatne, G. Atluri, J.H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2318–2331. 10.1109/TKDE.2017.2720168.

[31] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, Prabhat, Deep learning and process understanding for data-driven earth system science, Nature 566 (2019) 195–204. 10.1038/s41586-019-0912-1.

[32] J. Willard, X. Jia, S. Xu, M. Steinbach, V. Kumar, Integrating scientific knowledge with machine learning for engineering and environmental systems, ACM Computing Surveys 55 (2022) 1–37. 10.1145/3514228.

[33] Z. Zhu, Z. Wang, C. Dong, M. Yu, H. Xie, X. Cao, L. Han, J. Qi, Physics informed neural network modelling for storm surge forecasting: A case study in the bohai sea, china, Coastal Engineering 197 (2025) 104686. 10.1016/j.coastaleng.2024.104686.

[34] NOAA Center for Operational Oceanographic Products and Services, NOAA tide predictions and tides & currents, (2026). https://tidesandcurrents.noaa.gov/.

[35] H. Hersbach, B. Bell, P. Berrisford, et al., The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society 146 (2020) 1999–2049. 10.1002/qj.3803.

[36] K. Cho, B. van Merriënboer, C. Gulcehre, et al., Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Proceedings of EMNLP, 2014: pp. 1724–1734. 10.3115/v1/D14-1179.

[37] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681. 10.1109/78.650093.

[38] H.R. Künsch, The jackknife and the bootstrap for general stationary observations, The Annals of Statistics 17 (1989) 1217–1241. https://doi.org/10.1214/aos/1176347265.

[39] H. Hersbach et al., ERA5 hourly data on single levels from 1940 to present, Copernicus Climate Change Service, Climate Data Store (2018). https://doi.org/10.24381/cds.adbb2d47.

[40] Copernicus Marine Service, Global Ocean Physics Reanalysis, GLOBAL_MULTIYEAR_PHY_001_030. https://doi.org/10.48670/moi-00021.

[41] Copernicus Marine Service, Global Ocean Waves Reanalysis, GLOBAL_MULTIYEAR_WAV_001_032. https://doi.org/10.48670/moi-00022.

[42] Copernicus Marine Service, Global Ocean Physics Analysis and Forecast, GLOBAL_ANALYSISFORECAST_PHY_001_024. https://doi.org/10.48670/moi-00016.

[43] Copernicus Marine Service, Global Ocean Waves Analysis and Forecast, GLOBAL_ANALYSISFORECAST_WAV_001_027. https://doi.org/10.48670/moi-00017.

[44] GEBCO Bathymetric Compilation Group, The GEBCO_2025 Grid, a continuous terrain model for oceans and land at 15 arc-second intervals (2025). https://doi.org/10.5285/37c52e96-24ea-67ce-e063-7086abc05f29.