Online Resource 1 for the article in GPS Solutions:

Forecasting precipitable water vapor at unmonitored coastal sites using a sparse GNSS network and NWP forecasts

Peng Min[1,2], **Jie Zhao**[1], **Linrui Liu**[1], **You Yin**[1], **Ganzhi Luo**[1], **Han Su**[2,*]

[1] *Department of Electronic Information Engineering, Chengdu Jincheng College, Chengdu 611731, China*
[2] *School of Computer Science, Sichuan Normal University, Chengdu 610101, China*
[*] *Corresponding author: Han Su (E-mail: jkxy_sh@sicnu.edu.cn)*

S1. Samples and station information

The 30-station China coastal source backbone contains: BCST, BHLD, BLHT, BRZH, BTGU, BXCS, BXMD, BZFD, BZMW, DBJO, DCHM, DCWU, DDSH, DDSN, DLCG, DLSI, DLYG, DSPU, DWZH, DXMN, DZHI, DZJJ, NBHI, NDFG, NFCN, NHKO, NHZH, NQLN, NWZU, and NZUH. Station NFCN has no scored labels during the independent evaluation period, yielding 29 scored stations for source-network and station-withholding evaluations.

The 19 independent coastal target stations are: BBYQ, BLKO, BPLI, BQLY, DDCN, DDJS, DKMN, DNJI, DPTN, DSSA, DXQS, NDWS, NGZH, NNZU, NSTO, NSYA, NSZN, NZLG, and NZPO. Target stations are evaluated individually after appending them to the 30-station network; target locations do not exchange information with one another, strictly reflecting unmonitored deployment conditions.

All deployment routes—direct numerical prediction (GEFS; Zhou et al. 2022; Guan et al. 2022), calibrated NWP (MOS), a single-station neural model without exchange, frozen correction transfer, and NEP without local history (Level 1 regional coverage)—are evaluated across the unified cohort of 19 external targets (as well as the 29 withheld stations) using identical evaluation forecast–target pairs, ensuring strictly comparable, unconfounded comparisons across all methods.

Station withholding uses five region-stratified groups of six stations. During fitting and validation, held histories are replaced with the missing state, NWP channels are neutralized, and target labels are masked. Coordinates and the original graph geometry remain available. At evaluation, real NWP is restored and the target history follows the designated condition with or without local history. The local encoder's held-station offset is replaced by the mean of the fitting-station offsets.

The NWP-supported dataset is chronologically partitioned into 47,373 training windows (16 November 2010 to 23 April 2016), 1,095 validation windows (24 April 2016 to 1 February 2017), an intermediate development window set (to 22 August 2018), and 1,942 independent evaluation windows in the out-of-sample test period (23 August 2018 to 31 December 2019). The split dates refer to calendar boundaries; scored origins additionally require 168 h of context and usable numerical guidance. Model checkpoints were selected exclusively on validation loss, and all reported spatial-transfer benchmarks represent strictly frozen, out-of-sample evaluations.

S2. Training and no-history implementation

The temporal encoder uses 24 h patches, stride 6 h, two Transformer layers (Vaswani et al. 2017), and dimension 64. The exchange input concatenates numerical PWV, its tendency from the prediction origin, and the local forecast, each of length 24 (totaling 72 feature dimensions). Two four-head graph attention layers (Veličković et al. 2018)

operate on 64-dimensional station representations. The correction head is initialized to zero. The model contains 236,137 parameters.

In earlier tests with dedicated models, appending the raw observation-minus-NWP departure as an input feature gave no additional gain (source network 2.057 vs. 2.056 mm; 19 external targets 2.443 vs. 2.445 mm), consistent with the network extracting forecast deviations from the concatenated representation.

Models are optimized using observed-target MSE loss, Adam optimizer, initial learning rate 0.001, weight decay 0.0001, batch size 128, up to 100 epochs, and validation patience of 15 epochs. Learning-rate warm-up and cosine decay, gradient clipping at 1.0, and exponential moving-average weights with decay 0.999 are applied. Five independent random seeds (42–46) are evaluated for all neural models.

For Level 1 regional coverage (NEP without local history), each fitting station's entire 168 h history is independently masked with probability $p = 0.2$ in each training sample. The missing state sets standardized values to zero, observation validity to zero, and the transformed age channel to one. NWP and observed labels remain available. Validation masks are deterministic by origin (seed 20261002). The no-history experiment contains three model variants, five station-withholding folds, and one full-network setting, each evaluated across five seeds, totaling 90 completed fits.

The no-history normalization proxy is fitted on source backbone stations. Mean GNSS PWV is regressed on an intercept, training-period NWP mean, and station elevation. GNSS standard deviation is regressed on an intercept, NWP standard deviation, and NWP mean. The fitted relations provide the target scaling parameters without requiring target GNSS observations. In the external-target evaluation, replacing the proxy with ground-truth target GNSS statistics yields an MAE of 2.682 mm versus 2.680 mm with the proxy; in the withholding experiment, the corresponding values are 2.457 and 2.462 mm, verifying that proxy estimation incurs negligible performance penalty.

Training-free correction transfer uses the three nearest source stations, with weights proportional to $\max(\text{distance}, 1\ \text{km})^{-2}$. Source forecasts are averaged over seeds before corrections are formed and scored. Other neural scores average the five individual model scores. Comparisons between deployment routes retain these aggregation choices.

S3. Component and auxiliary comparisons

Table S1 documents the exhaustive baseline benchmark across 15 models evaluated at the 30-station source backbone network during the independent evaluation period. This suite spans observation-only time-series architectures (PatchTST, Nie et al. 2023; TimesNet, Wu et al. 2023; Crossformer, Zhang and Yan 2023; iTransformer, Liu et al. 2024; STGCN, Yu et al. 2018; and anchored attention variants), numerical calibration and local fusion baselines (Calibrated NWP and Local Fusion), and hybrid deep learning models fusing GNSS and NWP guidance.

Table S1. Comprehensive baseline comparison at the source backbone network during the independent evaluation period (30 stations, 1,942 forecast origins). Neural models are means of five seeds (± standard deviation over seeds); the last column gives the relative MAE reduction of NEP with respect to each model. † Simplified re-implementation trained under the same protocol. Anchor: lead-dependent coefficient on the latest causally filled observation. Own-station NWP fusion: the fusion head sees only the station's own NWP forecast and departure.

Model	Inputs	Parameters	MAE (mm)	RMSE (mm)	NEP MAE reduction (%)
Observation-only models	—	—	—	—	—
Persistence	GNSS	–	4.290	6.395	52.3
PatchTST	GNSS	95,128	4.057 ± 0.011	5.883	49.5
iTransformer	GNSS	332,696	3.851 ± 0.027	5.502	46.8
STGCN	GNSS	142,232	3.784 ± 0.018	5.381	45.9
Crossformer†	GNSS	360,456	3.769 ± 0.028	5.414	45.7
TimesNet†	GNSS	339,934	4.328 ± 0.059	5.965	52.7
STGCN + anchor	GNSS	142,256	3.756 ± 0.014	5.390	45.5
Crossformer† + anchor	GNSS	360,480	3.736 ± 0.024	5.388	45.2
Station attention + anchor	GNSS	229,753	3.680 ± 0.009	5.313	44.4
NWP-based references	—	—	—	—	—
Calibrated NWP (MOS)	NWP	–	2.579	3.823	20.6
Operational local fusion	GNSS + NWP	–	2.212	3.142	7.5
GNSS + NWP neural models	—	—	—	—	—
STGCN core + own-station NWP fusion	GNSS + NWP	148,576	2.155 ± 0.008	3.049	5.0
Attention core + own-station NWP fusion	GNSS + NWP	236,073	2.137 ± <0.001	3.033	4.2
No-exchange control	GNSS + NWP	211,233	2.131 ± 0.002	3.017	3.9
NEP (Level 2 / Full History)	GNSS + NWP	236,137	2.047 ± 0.003	2.878	–

Figure S1 details the no-history forecast skill for each individual tropical cyclone in China during the independent evaluation period. Across 12 storms with per-storm results (755 of 798 forecast pairs), NEP without local history outperforms calibrated NWP in 9 storms, whereas correction transfer degrades below NWP in 8 storms. The pooled estimate across all 18 storms achieves an 8.7% error reduction [95% CI: 4.8, 12.9]%.

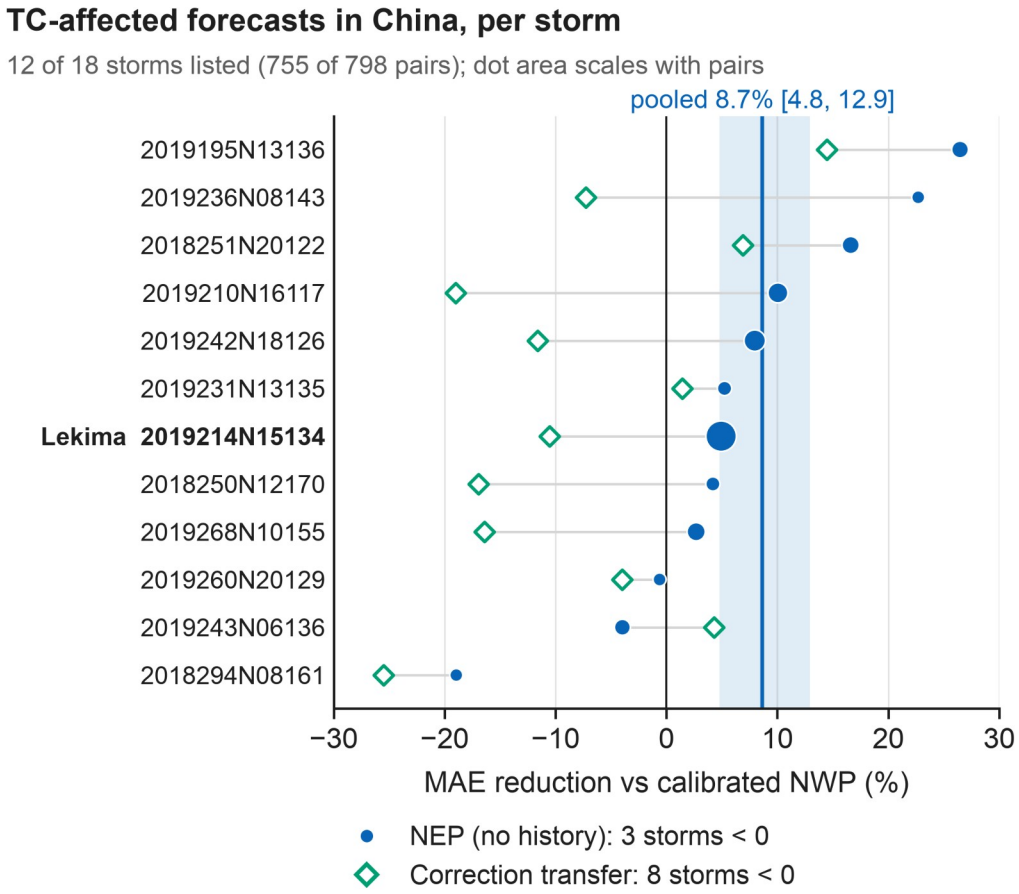


Fig. S1. *No-history skill for each tropical cyclone in China. MAE reduction relative to calibrated NWP for NEP without local history (dots; area proportional to the number of forecast pairs) and for correction transfer (diamonds), for the 12 storms with per-storm results (755 of 798 forecast pairs). The vertical line and shading show the pooled estimate for all 18 storms, 8.7% [4.8, 12.9]%. Storms are identified by IBTrACS (Knapp et al. 2010) serial number.*

The simple distance-weighted observed-departure correction yields MAE 2.189 mm compared with 2.212 mm for local fusion and 2.047 mm for NEP. This provides a geographic statistical comparison in addition to the neural no-exchange control.

Table S2. Forecast error without local history across comparators and deployment routes (MAE, mm). Withholding includes 29 scored stations and external-target evaluation includes 19 independent coastal stations.

Method	Withheld stations	External targets
Direct Numerical NWP (GEFS)	2.848	3.074
Pooled NWP calibration (MOS)	2.648	2.881
Neighbor persistence	4.513	4.646
Neighbor local fusion interpolation	2.999	3.088
Observation-only model (no local history)	4.468	4.669
No-exchange control (no local history)	2.628	2.834
Frozen correction transfer	2.592	2.715
NEP without local history (Level 1)	2.462	2.680

Neighbor persistence interpolates the latest source observations and holds the estimate through the forecast. Neighbor local fusion interpolation applies the spatial procedure to source local fusion forecasts. The trained observation-only model without local history omits NWP. These comparisons reinforce the role of numerical guidance, the viability of spatial correction transfer, and the value of a learned regional network correction.

Table S3. Complete performance comparison of principal deployment routes across the 19 independent coastal target stations, including the routes without local history and the reference with local history.

Method	Target GNSS history	MAE (mm)	RMSE (mm)
Direct Numerical NWP (GEFS)	No	3.074	4.391
Pooled NWP calibration (MOS)	No	2.881	4.185
No-exchange control (no local history)	No	2.834	4.083
Frozen correction transfer	No	2.715	3.942
NEP without local history (Level 1 regional coverage)	No	2.680	3.910
NEP with local history (Level 2 reference)	Yes	2.421	3.505

For spatial transfer variants across the 19 external targets, direct interpolation of source NEP forecasts without NWP departure subtraction gives 2.914 mm, while transferring no-exchange corrections gives 2.823 mm. A six-neighbor variant gives 2.762 mm. The correction primarily improves the mean forecast PWV level; the lead-1-to-lead-24 change MAE is 3.517 mm for direct GEFS and 3.730 mm after correction transfer.

Table S4 lists the station-wise results at the 19 external targets. NEP without local history has lower MAE than pooled calibration at every station (19/19), with the smallest reductions at the northern Bohai–Yellow Sea stations (0.4–2.6%) and the largest along the South China coast (up to 17.2%). The two stations with the largest absolute errors (NSYA and NSZN) also have the largest errors under pooled calibration.

Table S4. Station-wise MAE (mm) at the 19 external targets without target GNSS histories during the independent evaluation period, ordered by latitude (north to south). Nearest source: distance to the nearest of the 30 source stations. Level 2 reference reflects local instrumentation mode (aggregate MAE 2.421 mm across all evaluation windows).

Station	Lon (°E)	Lat (°N)	Nearest source (km)	Pooled calibration	NEP Level 1 (no local history)	NEP Level 2 (with local history)	Level 1 gain vs MOS (%)
BBYQ	122.09	40.28	106	1.454	1.448	1.389	0.4
BPLI	120.74	37.83	63	1.742	1.717	1.596	1.5
BLKO	120.29	37.68	98	1.616	1.574	1.526	2.6
BQLY	121.39	36.27	73	1.867	1.830	1.805	2.0
DDJS	122.17	30.81	32	3.067	2.992	2.928	2.4
DXQS	122.26	30.52	29	2.592	2.499	2.347	3.6
DDCN	121.90	28.49	79	2.896	2.766	2.379	4.5
DKMN	121.28	28.09	20	2.522	2.379	2.240	5.7
DNJI	121.07	27.47	53	2.730	2.612	2.318	4.4
DSSA	120.21	26.92	66	2.833	2.558	2.409	9.7
DPTN	119.85	25.46	80	2.623	2.504	2.294	4.5
NSTO	116.76	23.35	89	2.924	2.465	2.329	15.7
NGZH	113.69	22.66	49	2.926	2.429	2.306	17.0
NZLG	115.57	22.66	104	2.814	2.456	2.297	12.7
NSYA	114.25	22.56	35	6.485	6.064	4.835	6.5
NDWS	113.72	21.93	37	2.665	2.398	2.208	10.0
NZPO	111.84	21.57	193	2.911	2.423	3.230	16.8
NSZN	109.12	21.04	42	5.290	5.034	3.942	4.8
NNZU	110.56	20.90	103	2.904	2.404	2.406	17.2

Table S5 gives the MAE of the four routes by forecast lead at the China coastal target stations. There, NEP without local history is more accurate than correction transfer at 1 and 3 h (by 12.6% and 3.6%), the two routes agree within 1% from 6 h onward, and NEP with local history is the most accurate route at every lead, with the largest margin at short leads. The US pattern differs (Table S7): correction transfer is more accurate at 1 and 3 h (by 10.3% and 4.0%), and NEP without local history is 0.5–1.6% more accurate from 6 h onward.

Table S5. MAE (mm) by forecast lead across the China coastal target stations. Neural scores average five seeds; correction transfer uses ensemble-mean source forecasts.

Method	1 h	3 h	6 h	12 h	18 h	24 h
Pooled NWP calibration (MOS)	2.62	2.62	2.72	2.87	3.02	3.17
Correction transfer	2.70	2.49	2.54	2.69	2.85	2.99
NEP without local history (Level 1)	2.36	2.40	2.52	2.69	2.84	2.99
NEP with local history (Level 2)	1.21	1.77	2.22	2.56	2.77	2.93

S4. Observation context and spatial sensitivity

Source-network NEP improves local fusion at all 29 scored stations. Seasonal relative MAE reductions range from approximately 7.3% to 8.4%, and improvements remain across forecast-origin hours and recent-change strata. These results support the aggregate comparisons without altering the principal spatial-transfer questions.

Table S6 reports the reduction in MAE of NEP relative to local fusion and to the no-exchange control within strata defined from information available at forecast issuance. The reduction relative to local fusion lies between 7.3% and 8.4% in every season and varies little with issuance time, the size of the preceding 6-h PWV change, or the PWV anomaly. It is substantially larger when the local observation is missing for more than 6 h (19.9%), reaches 9.9% in tropical-cyclone situations, and increases from 1.3% in the Bohai–Yellow Sea group to 10.7% in the South China Sea group.

Table S6. Relative MAE reduction (%) of NEP at the source backbone network during the independent evaluation period, by stratum, with 95% time-block bootstrap (Künsch 1989) intervals (storm-clustered for the tropical-cyclone stratum). Strata are defined from information available at forecast issuance.

Stratum	Level	Scored pairs	Reduction vs Local Fusion (%) [95% CI]	Reduction vs no-exchange (%) [95% CI]
Season	Dec–Feb	257,502	8.1 [6.6, 9.7]	4.7 [3.3, 6.1]
	Mar–May	211,906	7.3 [5.6, 9.0]	4.4 [2.9, 6.1]
	Jun–Aug	180,344	8.4 [6.7, 9.9]	3.6 [2.9, 4.3]
	Sep–Nov	360,550	7.5 [6.3, 8.5]	3.7 [2.8, 4.6]
Coastal group	Bohai–Yellow Sea (B)	322,822	3.6 [2.7, 4.5]	1.3 [0.7, 2.0]
	East China Sea (D)	472,074	6.6 [5.7, 7.4]	1.9 [1.3, 2.5]
	South China Sea (N)	215,406	13.0 [11.6, 14.3]	10.7 [9.3, 12.0]
6-h PWV change at issuance	Q1 (smallest)	178,152	7.5 [6.2, 8.9]	3.7 [2.5, 5.0]

Stratum	Level	Scored pairs	Reduction vs Local Fusion (%) [95% CI]	Reduction vs no-exchange (%) [95% CI]
	Q2	205,927	6.8 [5.6, 8.0]	3.5 [2.5, 4.4]
	Q3	208,833	7.3 [6.1, 8.3]	3.5 [2.8, 4.2]
	Q4	210,731	7.5 [6.4, 8.4]	3.2 [2.4, 3.8]
	Q5 (largest)	206,659	8.1 [7.2, 9.0]	4.1 [3.5, 4.8]
PWV anomaly at issuance	Low	322,971	7.2 [5.9, 8.4]	4.0 [3.0, 5.0]
	Middle	360,803	8.7 [7.7, 9.7]	4.2 [3.3, 5.1]
	High	326,528	6.6 [5.5, 7.6]	2.7 [1.8, 3.5]
Local observation at issuance	Available	943,735	7.0 [6.2, 7.7]	3.7 [3.1, 4.3]
	Gap of 1–6 h	30,142	8.3 [6.8, 9.7]	4.6 [3.5, 5.7]
	Gap > 6 h	36,425	19.9 [15.0, 24.6]	8.6 [6.6, 10.8]
Tropical cyclone	Within 500 km	12,329	9.9 [4.5, 14.4]	4.9 [0.7, 8.9]
	Otherwise	997,973	7.7 [6.9, 8.4]	4.0 [3.4, 4.6]
Issuance time	05 UTC	252,392	7.2 [6.4, 8.1]	3.6 [3.0, 4.2]
	11 UTC	252,424	7.3 [6.4, 8.2]	3.5 [2.7, 4.2]
	17 UTC	252,452	8.0 [7.2, 8.9]	3.9 [3.2, 4.5]
	23 UTC	253,034	7.4 [6.5, 8.3]	3.5 [2.9, 4.1]

An exploratory radius experiment also masks source GNSS histories within a prescribed distance of each external target without local history, while retaining all NWP inputs. NEP without local history MAE rises gradually from 2.680 mm with no additional masking to 2.765 mm when source histories within 800 km are masked. The latter removes histories at approximately 12 source stations per target on average and remains below pooled NWP calibration at 2.881 mm.

The radius experiment indicates that nearby GNSS histories contribute to accuracy, while numerical context and learned calibration retain substantial value when those histories are unavailable. Associations between natural nearest-source distance and gain over the no-exchange control do not have intervals excluding zero. The evaluated external targets have nearest-source distances of approximately 20–193 km.

S5. Precipitation diagnostic

The precipitation analysis uses CMORPH satellite precipitation estimates (Joyce et al. 2004). Heavy-rain events are defined as accumulated precipitation of at least 30 mm in 12 h, averaged across a 3×3-grid neighborhood around each station. The independent out-of-sample evaluation contains 481 events. Maximum forecast PWV is used as a predictor, and discrimination thresholds and combined predictors are fitted using the development period.

Forecast PWV substantially improves discrimination over the current PWV value, with ROC AUC reaching approximately 0.902 for both NEP and local fusion compared with 0.871 for current PWV. Raw GEFS precipitation alone gives an ROC AUC of 0.929; combining GEFS with NEP or local fusion forecast PWV elevates the ROC AUC to 0.948 and 0.947, respectively (Fig. S2). This diagnostic demonstrates that forecast PWV provides actionable supplementary moisture guidance, improving heavy convective rain discrimination and reducing false alarms.

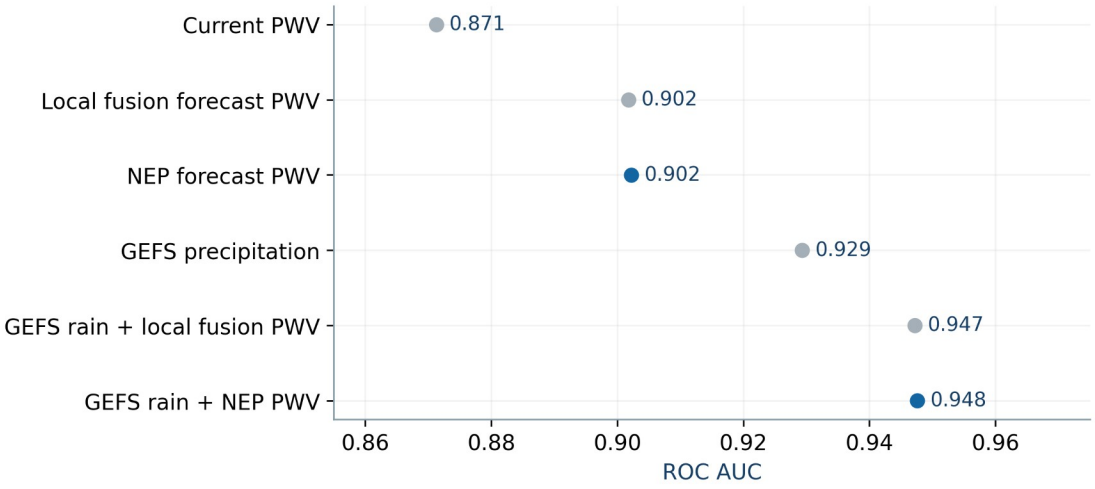


Fig. S2. *Heavy-rain discrimination using PWV and numerical precipitation predictors. All values refer to the same event definition (accumulated precipitation ≥ 30 mm in 12 h) across the independent evaluation sample. Blue identifies predictors using NEP.*

S6. US coastal benchmark and regional scaling

Candidate stations along the US Gulf and Atlantic coast were selected from the Nevada Geodetic Laboratory (NGL) archive within 50 km of the coastline (24–45°N) covering the 2011–2019 period (Blewitt et al. 2018). Thirty stations with data availability exceeding 85% were designated as the US training backbone (median spacing 148 km), while 108 non-overlapping coastal stations served as regional target sites. During observation outages exceeding 6 h, NEP reduced error by 30.7% (95% bootstrap interval 26.0–34.7%) relative to local fusion. Station-by-station evaluations confirm that NEP outperformed local fusion at 106 of 108 external target stations, and NEP without local history outperformed MOS at 103 of 108 stations. Under complete observational absence, NEP without local history reduced MAE by 0.8% (0.2–1.3%) relative to correction transfer across the 1–24 h evaluation horizon. Lead-dependent forecast errors across these 108 unmonitored target stations are tabulated in Table S7.

Table S7. MAE (mm) by forecast lead across the 108 US coastal target stations without local observation histories during the independent evaluation period (August 2018 to December 2019).

Method	1 h	3 h	6 h	12 h	18 h	24 h
Correction transfer	1.56	1.66	1.85	2.08	2.26	2.46
NEP without local history (Level 1)	1.74	1.73	1.84	2.05	2.23	2.42

S7. Station spacing sensitivity test

To test whether station spacing controls the benefit of inter-station exchange, we constructed three 30-station synthetic networks around the same ten target stations, with neighbors located at median distances of 34 km (dense), 70 km (medium), and 184 km (sparse) from the targets. NEP and the no-exchange control were trained independently on each network configuration. At the target stations, exchange reduced forecast MAE by 0.32% (95% bootstrap interval 0.02–0.63%), 0.33% (−0.02–0.68%), and 0.55% (0.18–0.94%) across the dense, medium, and sparse networks, respectively (Fig. S3). Closer station proximity did not increase the benefit of exchange (difference between the densest and sparsest networks +0.23 percentage points, 95% interval −0.10 to +0.54 percentage points), ruling out geometric network sparsity as the primary explanation for regional differences.

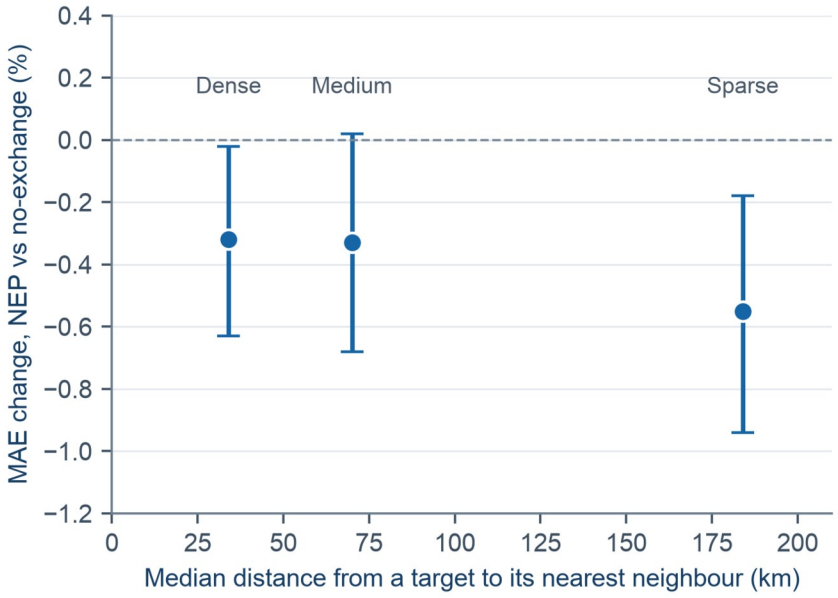


Fig. S3. *Relative MAE reduction (%) by inter-station exchange (NEP vs. no-exchange control) across synthetic networks with dense (median 34 km), medium (70 km), and sparse (184 km) station spacing around ten fixed target stations. Error bars denote 95% moving-block bootstrap intervals.*

S8. Distance decay, network siting sensitivity, and outage dynamics

To analyze the physical spatial limits of regional forecasting, Table S8 details forecast error reductions stratified by the distance from each target to the nearest backbone station across the US (600 target × chain evaluations) and China (48 target stations). Training-free correction transfer (inverse-distance weighting) decays sharply beyond 100 km, falling to 2.1% in the US and 2.3% in China. In contrast, NEP without local history sustains robust improvements (4.8–4.9% in the US and 7.9% in China) across 150–300 km footprints, demonstrating that geographically biased graph attention anchored by continuous NWP tendencies successfully preserves synoptic-scale error coherence where geometric interpolation collapses.

Table S8. Forecast improvement (MAE reduction vs. calibrated NWP, %) binned by distance to the nearest backbone station across US and China coastal domains.

Distance bin (km)	US NEP w/o history (%)	US Correction Transfer (%)	US No-exchange (%)	China NEP w/o history (%)	China Correction Transfer (%)	China No-exchange (%)
< 40 km	+6.2	+5.3	+1.4	+5.4	+6.6	+1.7
40–70 km	+6.7	+7.1	+2.2	+5.7	+4.8	+1.4
70–100 km	+5.5	+3.7	+1.5	+7.9	+2.4	+1.4
100–150 km	+4.8	+2.2	+2.2	+7.9 (>100 km)	+2.3	+0.4
> 150 km	+4.9	+2.1 [−0.2, 3.4]	+2.1	–	–	–

Table S9 evaluates backbone siting sensitivity by comparing the designed farthest-point station placement strategy against random station selection across 40 fixed US regional targets. At a sparse backbone density of 10 stations, designed farthest-point siting provides a meaningful advantage over random placement (3.6% vs. 2.7% error reduction relative to MOS). However, once the network scales to 20 stations, the two configurations achieve nearly identical performance (4.7% vs. 4.6%), confirming that geometric optimization is primarily beneficial in extremely sparse deployment regimes.

Table S9. Backbone siting sensitivity: designed farthest-point placement versus random station selection across 40 fixed US targets.

Backbone size	Siting configuration	Median distance to target (km)	Level 1 gain vs. MOS (%)	Level 2 upgrade value (%)
10 stations	Designed (farthest-point)	131	3.6	7.7
10 stations	Random selection	149	2.7	7.6
20 stations	Designed (farthest-point)	79	4.7	7.1
20 stations	Random selection	86	4.6	7.0

Under legacy dedicated architectures (models fitted strictly on complete observation records), sensor outages exceeding 6–12 h lead to severe forecast deterioration (+49% in US, +10% in China worse than MOS). An operational switching rule (transitioning to the no-history mode when local observation age exceeds an empirical threshold: 1 h in the US, 6 h in China) eliminates this divergence and recovers a 6.0% error reduction relative to MOS during prolonged outages.

In the single-network architecture trained with whole-station history dropout ($p = 0.2$), explicit threshold switching is rendered unnecessary in China, where the unified network gracefully transitions from Level 2 to Level 1 and outperforms dedicated models by 1.0% [95% CI: 0.6, 1.4]% across all windows. In the US, the single network remains 1.1–11.2% superior to MOS across all outage durations $k \in \{0, ..., 72\}$ h, while the switching rule provides a modest supplementary benefit (6.0% vs. 1.1% to 3.3% at 6–24 h gaps). Across natural outages (>6 h gap), the single network delivers an extraordinary 19.9% (China) and 30.7% (US) accuracy advantage over local fusion.

S9. Cross-ocean zero-shot transfer of China-trained model

To evaluate cross-continental portability without retraining, the NEP model trained strictly on the 30-station China coastal network was deployed directly onto the 30 US training stations without fine-tuning or updating neural network parameters. Local standardization parameters (mean and standard deviation) were calculated from training-period observations, while the local-encoder station offsets were substituted with the source-network mean. Across the independent evaluation period (August 2018 to December 2019), the zero-shot transferred model achieved an aggregate MAE of 2.21 mm—representing a 7.6% error reduction over US-fitted MOS (2.39 mm) and matching the performance of locally calibrated local fusion (2.22 mm; error reduction 0.5%, 95% bootstrap interval −0.7% to 1.7%). While localized retraining on the US network yielded an additional 4.9% improvement (2.10 mm), the direct zero-shot transfer underscores the architecture's inherent capacity to generalize across independent geodetic processing regimes and regional atmospheric dynamics.

S10. Synoptic case study: Typhoon Mangkhut

During the passage of Super Typhoon Mangkhut across the northern South China Sea on 15–16 September 2018, PWV observed across coastal stations exhibited a pronounced spatio-temporal succession (Fig. S4a, b). PWV peaked first at Huizhou and Zhuhai and approximately 15 h later at Haikou, Beihai, and Dongfang. At Xisha island station, observed PWV experienced a sharp anomalous dip from 52.5 mm to 45.3 mm between 15:00 and 18:00 UTC on 15 September before surging to 61.0 mm within 6 h; ERA5 reanalysis fields (Hersbach et al. 2020) confirm this rapid transient depression, establishing that the feature reflected authentic mesoscale dynamics rather than sensor error. Forecasts initialized at the trough of the dip (15 September 17:00 UTC) with single-station models (local fusion and no-exchange control) followed the falling local autoregressive trend, resulting in severe under-prediction (forecast errors 13.9 mm and 14.6 mm, respectively; Fig. S4c). In contrast, NEP leveraged upstream spatial graph attention to capture broader moisture inflow across the network, accurately anticipating the abrupt moisture surge (forecast error 4.0 mm). Across the broader South China Sea station array during Mangkhut, average performance between NEP and the no-exchange control was comparable, and attention weights remained stable throughout the storm period (mean temporal coefficient of variation of 5.7% for top-5 attended stations; Fig. S4d).



Fig. S4. *Synoptic analysis of Typhoon Mangkhut (September 2018). (a) Best track and station peak PWV timing. (b) Temporal sequence of observed GNSS PWV peaks across northern South China Sea stations. (c) Forecast issued at Xisha station at 17:00 UTC on 15 September 2018 during the transient PWV dip. (d) Evolution of layer-2 graph attention weights for top-5 attended stations relative to Xisha target during September 2018.*

Supplementary references

Blewitt G, Hammond WC, Kreemer C (2018) Harnessing the GPS data explosion for interdisciplinary science. Eos 99. https://doi.org/10.1029/2018EO104623

Guan H, Zhu Y, Sinsky E, Fu B, Li W, Zhou X, Xue X, Hou D, Peng J, Nageswararao MM, Tallapragada V, Hamill TM, Whitaker JS, Bates G, Pegion P, Frederick S, Rosencrans M, Kumar A (2022) GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. Mon Weather Rev 150:647–665. https://doi.org/10.1175/MWR-D-21-0245.1

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

Joyce RJ, Janowiak JE, Arkin PA, Xie P (2004) CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J Hydrometeorol 5:487–503. https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2

Knapp KR, Kruk MC, Levinson DH, Diamond HJ, Neumann CJ (2010) The International Best Track Archive for Climate Stewardship (IBTrACS). Bull Am Meteorol Soc 91:363–376. https://doi.org/10.1175/2009BAMS2755.1

Künsch HR (1989) The jackknife and the bootstrap for general stationary observations. Ann Stat 17:1217–1241. https://doi.org/10.1214/aos/1176347265

Liu Y, Hu T, Zhang H, Wu H, Wang S, Ma L, Long M (2024) iTransformer: Inverted transformers are effective for time series forecasting. In: International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2310.06625

Nie Y, Nguyen NH, Sinthong P, Kalagnanam J (2023) A time series is worth 64 words: Long-term forecasting with Transformers. In: International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2211.14730

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems 30, pp 5998–6008

Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph attention networks. In: International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1710.10903

Wu H, Hu T, Liu Y, Zhou H, Wang J, Long M (2023) TimesNet: Temporal 2D-variation modeling for general time series analysis. In: International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=ju_Uqw384Oq

Yu B, Yin H, Zhu Z (2018) Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI), pp 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Zhang Y, Yan J (2023) Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In: International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=vSVLM2j9eie

Zhou X, Zhu Y, Hou D, Fu B, Li W, Guan H, Sinsky E, Kolczynski W, Xue X, Luo Y, Peng J, Yang B, Tallapragada V, Pegion P (2022) The development of the NCEP Global Ensemble Forecast System version 12. Weather Forecast 37:1069–1084. https://doi.org/10.1175/WAF-D-21-0112.1