
Calibrating AI Subseasonal to Seasonal (S2S) Forecasts with Lightweight Regional Adapters

Ayush Raj
IISER Pune, India
Ashoka University, India

Manmeet Singh
Western Kentucky University, USA

Saptarishi Dhanuka
Ashoka University, India

Bedartha Goswami
IISER Pune, India

Parthasarathi Mukhopadhyay
Ashoka University, India

Sandeep Juneja
Ashoka University, India

Mrutyunjay Mohapatra
India Meteorological Department, India

Abstract

Global AI models provide broad-domain forecasts, but global skill does not guarantee reliable regional uncertainty. Using weekly rainfall over India as a single-variable case study, we find that raw FuXi-S2S members are strongly underdispersed, so we develop a lightweight probabilistic adapter that post-processes the fixed ensemble rather than retraining its global generator. A shared member encoder, permutation-invariant pooling, and a regional convolutional network estimate the spatially varying, lead-dependent location and spread corrections while maintaining member-index correspondence. In the primary retrospective evaluation, the adapter reduces the continuous ranked probability score by 16.21% relative to raw FuXi and by a further 4.26% relative to structured moment calibration. Supporting sensitivities and an independent scoring reconstruction retain this improvement, although coverage remains subnominal and bias improvement is unresolved. The evidence is specific to this model, rainfall variable, and India domain. Additional physically relevant predictors may provide richer regional context, while transfer to other regions requires explicit testing rather than assuming generalization.

1 Why regional uncertainty calibration matters

Weeks 1–6 rainfall guidance could extend the planning window for reservoir operations, agricultural advisories, and preparation for anomalously wet or dry periods. These applications need calibrated uncertainty, not only a more accurate ensemble mean: decision makers must know how much probability to assign to outcomes that matter locally. FuXi-S2S is a promising global AI ensemble with a 42-day horizon [1], yet its standardized rainfall members are strongly underdispersed over India. We therefore ask whether a small downstream model can adapt a fixed global ensemble to a regional observational record without retraining the global generator.

Ensemble post-processing is well established [2–4], and recent neural set models can return member-level ensembles [5, 6]; India-specific U-Net work already calibrates tercile probabilities through combined Weeks 3–4 [7]. Our contribution is bounded and ordered as follows: (1) a 42,434-parameter permutation-equivariant probabilistic adapter; (2) a primary retrospective evaluation with common-valid-midpoint, IMERG, regional, and independent artifact/scoring verification checks; and (3) complementary appendix evidence from a residual deterministic adapter. We claim improved

retrospective probabilistic quality for this FuXi/India protocol, not first-of-kind status, universal transfer, operational readiness, or decision value.

2 Regional probabilistic adapter

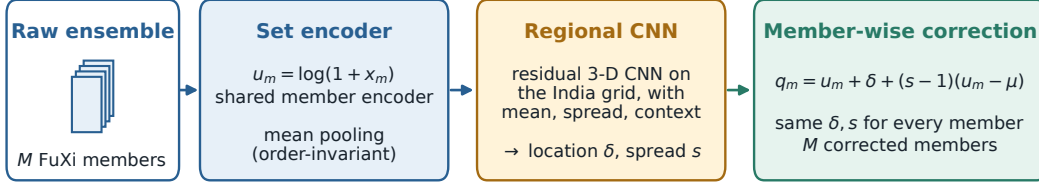


Figure 1: **Probabilistic regional adapter.** Shared encoding, symmetric pooling, and a regional Conv3D predict location and spread fields for every indexed member. Here “member-preserving” means only member-index identity correspondence under a permutation-equivariant transformation: clipping can create ties, and full spatial, temporal, or cross-lead dependence preservation is not established. Details are in Appendix B.

Data and fixed protocol. Fitting and checkpoint selection use the native 51-member FuXi reforecast archive from 2002–2021; strict retrospective ERA5 initialization supplies 50 locally generated members for the evaluation years. After unit conversion and India-domain standardization, native days $+1, \dots, +42$ form six seven-day means on the common 1.5° grid. IMD gauge rainfall is the fitting and primary verification target [8]. One of seeds 42–44 is selected only by 2018–2019 full-member validation CRPS and then applied without checkpoint averaging.

Transformation and member handling. For rainfall member x_m , let $u_m = \log(1 + x_m)$ and $\mu = M^{-1} \sum_m u_m$. The model predicts location δ and positive spread factor s and forms

$$q_m = u_m + \delta + (s - 1)(u_m - \mu), \quad s > 0. \quad (1)$$

Shared encoding followed by symmetric pooling makes the predicted fields invariant to input ordering; applying them to every member preserves each member’s identity under a permutation-equivariant transformation. Physical clipping can create ties, however, and preservation of full spatial, temporal, or cross-lead dependence has not been established. Training minimizes the area-weighted empirical finite-ensemble CRPS [9].

Structured moment comparator. The comparator is fitted only on 2002–2017 data. It applies a lead/month/grid-specific log-location correction, shrunk toward the weighted regional mean with ten pseudo-cases, and a lead/month regional spread scaling clipped to $[0.25, 4]$. Rainfall, including zero rainfall, is treated with $\log(1 + x) = \log_{1p}(x)$ before correction and mapped back with $\exp_{m1}$. It is deliberately structured but not an advanced precipitation-distribution model.

Evaluation and disclosure. The primary retrospective evaluation pools six lead-specific 2022–2024 JJAS cohorts, each with 100 initializations. CRPS, anomaly correlation coefficient (ACC), and signed bias use a fixed protocol; paired descriptive 95% intervals use 10,000 year-stratified circular-block resamples. The selected checkpoint was applied unchanged in 2022–2024, but the retrospective evidence may have influenced later analysis development; 2025 remains the first prospective, fully sealed evaluation within this study. No 2025 target or prospective result is reported here.

The evaluation design separates three questions. First, the raw comparison asks whether regional post-processing improves the supplied ensemble at all. Second, the moment comparator asks whether the neural model adds value beyond train-only correction of first and second moments. Third, the fixed sensitivities ask whether the result depends on lead-specific date composition or on using the same gridded reference for fitting and verification. None is a universal baseline ranking: the experiment concerns one global model, one regional domain, one rainfall aggregation, and one selected checkpoint. ACC uses anomalies relative to the fixed 2002–2017 IMD climatology, while bias is the signed area-weighted mean error. These complementary metrics distinguish distributional improvement from pattern agreement and mean offset.

3 Results across probabilistic and deterministic tasks

Table 1: **Probabilistic evaluation and fixed sensitivities.** CRPS is in mm day^{-1} and skill is relative CRPS reduction. The primary row is the 2022–2024 retrospective evaluation; common-midpoint and IMERG rows change one verification choice without refitting.

Evidence	Raw	Moment	Adapter	Skill vs. raw	Skill vs. moment
Primary IMD	3.1570	2.7630	2.6453	16.21%	4.26%
Common midpoint	3.1262	2.7416	2.6302	15.87%	4.06%
IMERG verification	3.1515	2.8256	2.7669	12.20%	2.08%
ACC, raw $\rightarrow$ adapter	ΔACC [95%]		Δbias [95%]		
0.2959 $\rightarrow$ 0.3220	+0.0261 [+0.0087,+0.0436]		+0.0201 [-0.2070,+0.2461] (unresolved)		

The adapter lowers pooled CRPS from 3.1570 to 2.6453 mm day^{-1} , with an additional reduction beyond moment calibration. ACC increases from 0.2959 to 0.3220; the paired interval for the +0.0261 change excludes zero. The signed-bias change is +0.0201 [-0.2070,+0.2461] mm day^{-1} and remains unresolved. A common set of 85 valid-period midpoints and an IMERG V07 rescoring sensitivity [10, 11] retain positive adapter CRPS skill; details are in Appendix C. Because both comparisons reuse the unchanged predictions, they isolate verification choices rather than model retuning. The common-midpoint check holds all leads to 85 identical dates and yields 15.87% CRPS reduction relative to raw. Rescoring against IMERG, which supplied no fitted quantities, yields 12.20% relative to raw. Agreement in direction is reassuring but does not make either rainfall product error-free, nor does it create an independent training sample.

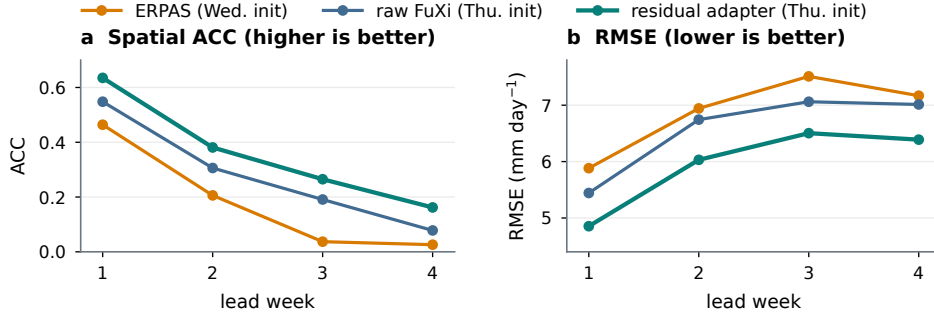


Figure 2: **Matched-valid-time operational comparison.** ERPAS initializes Wednesday; raw and corrected FuXi initialize Thursday. The 26-case Week 1–4 sensitivity therefore combines forecast-system differences with a 24-hour information-age difference; Table 10 gives the paired intervals.

The unchanged residual adapter has higher ACC and lower RMSE than ERPAS at all four shared leads; all eight descriptive pointwise intervals favor the adapter. This post-hoc, two-season sensitivity is an operational reference rather than an equal-information-age superiority claim, and ERPAS has no Week 5–6 product in this archive (Appendix I). Development-exposed lead-wise diagnostics (Figure 3; Appendix D) are consistent in direction across all six leads, but coverage remains below nominal 0.90. Artifact and scoring verification reproduced 45,450 stored case-metric rows to a maximum difference of 2.3×10^{-13} without rerunning neural inference (Appendix C.1).

Table 2: **Complementary deterministic mean correction.** Pooled 2022–2024 retrospective results use the separate day-0–41 mean-field contract. Error and bias are mm day^{-1} ; change is adapter minus raw except RMSE/MAE skill, where positive is better.

Metric	Raw FuXi mean	Residual adapter	Change
RMSE	5.723	5.239	8.45% skill
MAE	3.822	3.453	9.65% skill
ACC	0.276	0.364	+0.0889
Signed bias	-0.225	-0.858	-0.633

The deterministic companion reduces RMSE from 5.723 to 5.239 and MAE from 3.822 to 3.453, while ACC rises from 0.276 to 0.364. It predicts only the ensemble mean under a separate contract, and signed bias worsens by -0.633; therefore it supports error and pattern correction, not distributional or bias calibration (Appendix F).

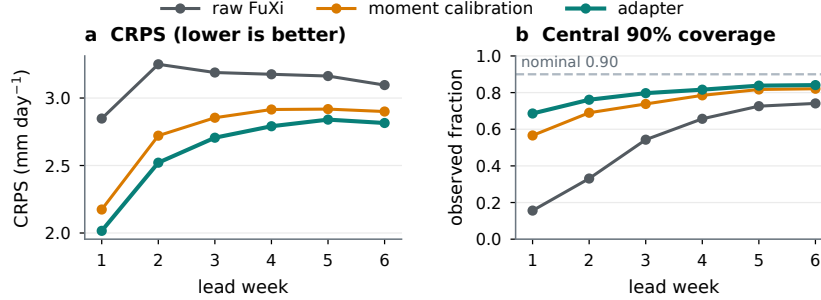


Figure 3: **Lead-wise CRPS and central 90% coverage.** Supplementary diagnostics for the development-exposed 2020–2024 JJAS cohorts (169 cases per lead); these are not an independent holdout. The dashed line marks nominal 0.90 coverage.

Figure 3 makes the underdispersion concrete: the raw central 90% member interval covers only 0.156 of observed Week 1 values and 0.741 at Week 6, whereas the adapter raises coverage to 0.686 and 0.841, with the largest CRPS reductions at the shortest leads. Coverage therefore moves toward but remains below 0.90 at every supplementary lead, so lower CRPS does not establish calibrated threshold probabilities. Wet-week, dry-week, and accumulated-rainfall products still require reliability diagrams, threshold Brier scores, and event-specific discrimination because the same location–spread map can affect thresholds differently and residual dependence errors can compound across weeks.

Spatial consistency is encouraging but similarly bounded. Pointwise regional intervals favor adapter CRPS across the four reported regions and six leads, but those cells share dates, large-scale weather regimes, and model errors. They are not 24 independent replications, and no multiplicity adjustment was applied. Figure 4 localizes the pooled gains descriptively: the largest coarse-cell ensemble-mean error reductions lie near Uttarakhand and along the Western Ghats, some cells worsen, and the median and lower-decile cases show how unevenly the All-India gain is distributed. The available evidence supports a regional India case study; it does not determine how much data, retraining, or architectural change would be needed in another hydroclimate or observation system.

Efficiency and limitations. The probabilistic network has 42,434 parameters; prepared-data training records span 52.0–82.7 seconds per seed on one NVIDIA A30. Inference latency and peak-memory requirements were not recorded; the supported broader claim is parameter efficiency, not low operational cost. Interpretation is limited by three evaluation years, overlapping periods, coarse resolution, subnominal coverage, unresolved bias, different training/scoring reconstructions, and no out-of-India or member-count sensitivity test. The block bootstrap models serial dependence but cannot make three years independent. Neither climatology-referenced CRPSS nor a fitted precipitation-distribution benchmark was evaluated.

Scope and next steps. The adapter consumes only a fixed ensemble array and a regional observation grid, so the same protocol can be reapplied to other global AI or NWP ensembles and to other monsoon regions; whether the gains transfer must be tested rather than assumed. A companion benchmark compares FuXi-S2S with operational NWP and other AI/hybrid systems over India [12], and applying the adapter to those systems is the most direct next test. Stronger comparators should include fitted precipitation-distribution models such as EMOS and censored shifted-gamma regression [2, 4] and learned distributional or generative calibrators [13]. The subnominal coverage calls for threshold reliability verification and possibly an explicit spread penalty, and the sealed 2025 season will extend the three-year evaluation.

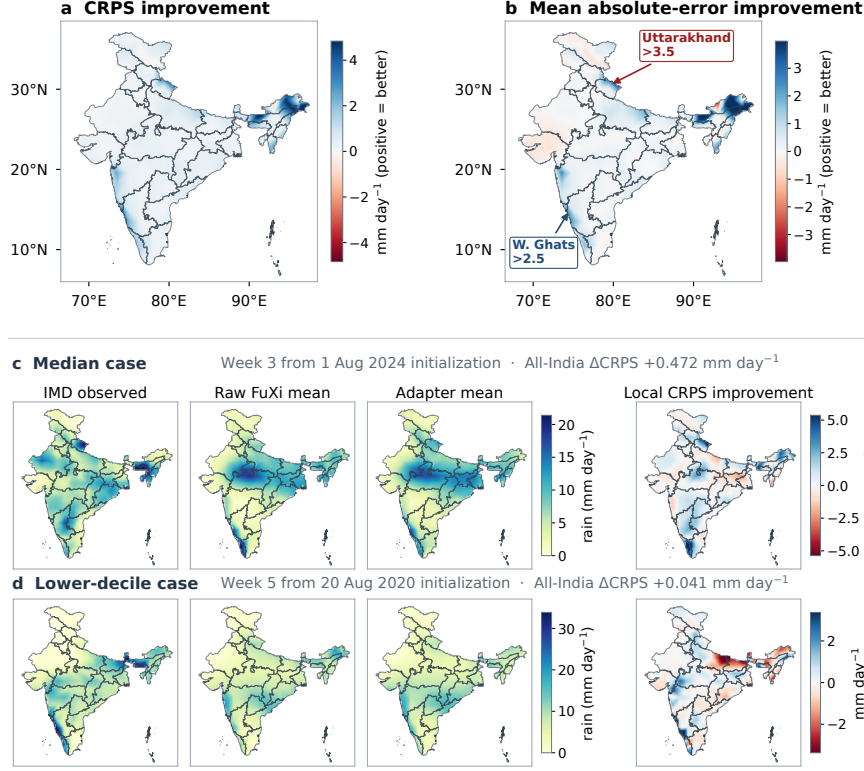


Figure 4: **Descriptive spatial footprint and diagnostic cases.** (a)–(b) Cellwise raw-minus-calibrated CRPS and ensemble-mean absolute-error reductions, pooled over six leads of the development-exposed 2020–2024 JJAS cohorts (169 cases per lead); positive is better, and (b) scores the probabilistic adapter’s ensemble mean. (c)–(d) Median and lower-decile cases ranked by paired All-India CRPS reduction. Shading linearly interpolates the native 1.5° cells within the India outline for display only; all scores use the native grid. Maps are descriptive: no pixelwise significance, event, or physical-attribution claim is made (Appendix J).

4 Conclusion

Across two complementary tasks, regional adapters improve both ensemble-distribution CRPS and ensemble-mean error and pattern skill for fixed global AI rainfall forecasts over the evaluated India case study. The corrected probabilistic members could support wet-week, dry-week, and accumulated-rainfall probability products, but those products require threshold-based verification and prospective testing with forecasting agencies and sectoral users. The evidence supports regional calibration as a promising downstream strategy, not operational or decision value. Prospective work should pre-register thresholds, aggregation windows, references, missing-data rules, and relevant scores, then test whether the probabilities are useful in agency and sectoral workflows. Future work will also explore richer computer-vision architectures for the adapter, such as U-Nets, vision transformers, and diffusion-based generative calibrators [7, 13, 14], and develop it into a generalizable framework that can be retrained for any global ensemble, variable, and region from a reforecast archive and local observations. Code and evaluation artifacts will be publicly released in a future version.

Acknowledgments and Disclosure of Funding

We thank the India Meteorological Department for the gridded rainfall observations and the FuXi-S2S developers for making their reforecasts available. We acknowledge the High Performance Computing facility at the Indian Institute of Science Education and Research (IISER) Pune for computational resources. The authors declare no competing interests.

References

- [1] Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu, Bo Lu, Deliang Chen, Shang-Ping Xie, Libo Wu, Qingchen Chao, Chensen Lin, Zixin Hu, and Yuan Qi. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1):6425, 2024.
- [2] Tilmann Gneiting, Adrian E. Raftery, Anton H. Westveld, and Tom Goldman. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118, 2005.
- [3] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.
- [4] Michael Scheuerer and Thomas M. Hamill. Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11):4578–4596, 2015.
- [5] Kevin Höhle, Benedikt Schulz, Rüdiger Westermann, and Sebastian Lerch. Postprocessing of ensemble weather forecasts using permutation-invariant neural networks. *Artificial Intelligence for the Earth Systems*, 3(1):e230070, 2024.
- [6] Zied Ben Bouallègue, Jonathan A. Weyn, Mariana C. A. Clare, Jesper Dramsch, Peter Dueben, and Matthew Chantry. Improving medium-range ensemble weather forecasts with hierarchical ensemble transformers. *Artificial Intelligence for the Earth Systems*, 3(1):e230027, 2024.
- [7] Emile D. Esmaili, Andrew W. Robertson, Muhammad Azhar Ehsan, Bohar Singh, Aaron Kaplan, Avijit Dey, and Susmitha Joseph. Improving subseasonal Indian Summer Monsoon rainfall forecasts with U-Net calibration. *Climate Dynamics*, 64(7):326, 2026.
- [8] D. S. Pai, Latha Sridhar, M. Rajeevan, O. P. Sreejith, N. S. Satbhai, and B. Mukhopadhyay. Development of a new high spatial resolution ($0.25^\circ \times 0.25^\circ$) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region. *MAUSAM*, 65(1):1–18, 2014.
- [9] Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.
- [10] George J. Huffman, David T. Bolvin, Dan Braithwaite, Kuo-Lin Hsu, Robert J. Joyce, Christopher Kidd, Eric J. Nelkin, Soroosh Sorooshian, Erich F. Stocker, Jackson Tan, David B. Wolff, and Pingping Xie. Integrated multi-satellite retrievals for the global precipitation measurement (GPM) mission (IMERG). In *Satellite Precipitation Measurement*, volume 67 of *Advances in Global Change Research*, pages 343–353. Springer, 2020.
- [11] George J. Huffman, Erich F. Stocker, David T. Bolvin, Eric J. Nelkin, and Jackson Tan. GPM IMERG final precipitation L3 1 day 0.1 degree x 0.1 degree V07, 2023. Dataset.
- [12] Ayush Raj, Saptarishi Dhanuka, Manmeet Singh, Parthasarathi Mukhopadhyay, Bedartha Goswami, and Sandeep Kumar Juneja. IndiaS2S-Bench: A comprehensive benchmark of subseasonal to seasonal forecasting over India. In *Tackling Climate Change with Machine Learning: workshop at NeurIPS 2026*, 2026.
- [13] Wen Shi, Baoxiang Pan, Jianbin Huang, Tingfeng Dou, Jie Feng, Huihui Yuan, Anran Wang, Xin Xia, Lei Huang, and Yong Luo. Distribution-guided ensemble postprocessing for S2S precipitation forecasts: A seamless pathway using deep generative models. *Journal of Geophysical Research: Machine Learning and Computation*, 3(3):e2025JH000993, 2026.
- [14] Jihun Ryu, Hisu Kim, Shih-Yu (Simon) Wang, and Jin-Ho Yoon. Increasing resolution and accuracy in sub-seasonal forecasting through 3D U-Net: the western US. *Geoscientific Model Development*, 19(1):27–39, 2026.
- [15] S. Abhilash, A. K. Sahai, N. Borah, S. Joseph, R. Chattopadhyay, S. Sharmila, M. Rajeevan, Brian E. Mapes, and A. Kumar. Improved spread–error relationship and probabilistic prediction from the CFS-based grand ensemble prediction system. *Journal of Applied Meteorology and Climatology*, 54(7):1569–1578, 2015.

A Archive, splits, and fixed evaluation protocol

Fitting and checkpoint selection use the native 51-member 2002–2021 FuXi reforecast archive after lossless extraction, unit conversion, India-domain subsetting, grid standardization, and aggregation of native days $+1, \dots, +42$. The retained rainfall is already on the common 1.5° grid. Evaluation uses 50 locally generated stochastic members from the official FuXi ONNX checkpoint and inference entry point with strict retrospective ERA5 initialization. No external perturbations are added and no control member is assumed.

The all-season reforecast splits contain 1,652 fitting starts from 2002–2017 and 196 checkpoint-selection starts from 2018–2019; boundary-crossing targets are purged. Mini-batches sample 16 members jointly without replacement, validation uses all 51 reforecast members, and evaluation uses all 50 locally generated members. Seed 43 is the only checkpoint used because it minimizes full-member validation CRPS among seeds 42–44; no checkpoint or output averaging is used.

The primary retrospective evaluation uses six lead-specific valid-midpoint JJAS cohorts from 2022–2024 with 100 initializations per lead (600 case–lead rows). They are not one common 100-date set. Paired descriptive intervals use 10,000 year-stratified circular-block draws of 16 starts, synchronized across leads by within-year ordinal. This accounts operationally for serial dependence but does not make the three source years independent. Comparator-relative CRPS reduction is $100(C_b - C_a)/C_b$ for named baseline b ; it is not climatology-referenced CRPSS.

The selected checkpoint was applied unchanged in 2022–2024, but the retrospective evidence may have influenced later analysis development. The 2025 evaluation remains the first prospective, fully sealed evaluation within this study; no 2025 target was opened or scored for this manuscript.

B Adapter and reconstruction details

For physical rainfall member x_m , let $u_m = \log(1 + x_m)$ and $\mu = M^{-1} \sum_m u_m$. A two-channel pointwise encoder receives u_m and $u_m - \mu$; its 8 features are sorted and mean-pooled. The pooled summary, log-location, RMS spread, and 7 spatial–seasonal context fields enter a residual three-dimensional CNN of width 24. Its heads predict location δ and $\log s = 2.0 \tanh(a)$, yielding

$$q_m = u_m + \delta + (s - 1)(u_m - \mu). \quad (2)$$

The network has 42,434 trainable parameters. Shared encoding and symmetric pooling make (δ, s) invariant to member order in exact arithmetic, while applying the same fields to each indexed member makes the set-to-set transformation permutation equivariant. This preserves each member’s identity through the transformation, but clipping can create ties; full spatial, temporal, and cross-lead dependence preservation is not established.

The fitting path sorts members before its float32 reduction. The retrospective scorer instead accumulates archive-order float32 log members in float64, casts the mean back to float32, clips stored log spread to $[-2, 2]$, and reconstructs

$$\mu_{\text{score}} = \text{float32} \left(M^{-1} \sum_m^{\text{float64}} u_m \right), \quad (3)$$

$$x'_{m,\text{score}} = \text{expm1}[\text{clip}\{\mu_{\text{score}} + \delta + s(u_m - \mu_{\text{score}}), 0, 20\}]. \quad (4)$$

The two paths are algebraically related but their numerical equivalence has not been established. Archive-order accumulation can be order-dependent in finite precision. Reported scores refer to the scoring equations above.

Area-weighted empirical finite-ensemble CRPS is

$$\text{CRPS}(\mathbf{x}, y) = \frac{1}{M} \sum_m |x_m - y| - \frac{1}{2M^2} \sum_{m,n} |x_m - x_n|. \quad (5)$$

AdamW uses learning rate 2×10^{-4} , weight decay 10^{-4} , batch size 8, gradient clipping at 5, and early stopping. The selected run is best at epoch 12.

Moment calibration. This comparator is estimated only from 2002–2017. At each lead, valid-midpoint month, and grid cell it estimates a log-location correction, with ten pseudo-cases shrinking the cell estimate toward the weighted regional mean. A lead/month regional spread factor is estimated from corrected error and raw member variance, then clipped to $[0.25, 4]$. The method uses log1p so zero rainfall maps to zero in log space and applies expm1 after correction. No evaluation-year observations enter either component.

C Fixed sensitivities and scoring verification

The common-midpoint sensitivity holds all six leads to the same 85 valid-period midpoints. CRPS is 3.1262 for raw, 2.7416 for moment calibration, and 2.6302 for the adapter, giving 15.87% and 4.06% relative reductions. The ACC change is +0.0313 [+0.0123,+0.0506], while the signed-bias interval crosses zero. This removes the lead-specific date-composition difference without changing the fitted model.

The IMERG Final V07 sensitivity changes only the verification reference; IMD still supplies all fitted quantities. CRPS is 3.1515, 2.8256, and 2.7669, corresponding to 12.20% versus raw and 2.08% versus moment calibration. The alternative reference has different retrieval and sampling errors, so agreement in direction is not an independent observational truth.

Table 3: Paired descriptive 95% interval detail for the probabilistic results in Table 1. Skill is relative CRPS reduction; ΔACC and Δbias are adapter minus raw. The latter two rows are fixed sensitivities, not additional fitted models.

Evidence	Skill vs. raw [95%]	Skill vs. moment [95%]	ΔACC [95%]	Δbias [95%]
Primary IMD	16.21 [14.81,17.69]	4.26 [2.99,5.70]	+0.0261 [+0.0087,+0.0436]	+0.0201 [-0.2070,+0.2461]
Common midpoint	15.87 [14.35,17.52]	4.06 [2.61,5.71]	+0.0313 [+0.0123,+0.0506]	+0.0455 [-0.2018,+0.2839]
IMERG verification	12.20 [10.51,14.02]	2.08 [0.67,3.66]	+0.0136 [-0.0057,+0.0335]	+0.0178 [-0.2134,+0.2474]

The CRPS intervals exclude zero for all three evidence rows. Primary and common-midpoint ACC changes also exclude zero, whereas the IMERG ACC interval and every adapter-minus-raw signed-bias interval cross zero. Thus the robust claim is improved distributional score, not uniformly resolved pattern or bias improvement across references.

C.1 Artifact and scoring verification

Independent table reconstruction checked all 45,450 stored case-metric rows, joins, weights, and interval inputs. A second implementation reproduced the saved scoring tables to a maximum absolute difference of 2.3×10^{-13} . These checks verify artifacts and scoring from saved predictions; they do not rerun training, recreate neural forward passes, or prove equivalence between the fitting and retrospective reconstruction paths.

D Supplementary lead-wise analysis using development-exposed 2020–2024 cohorts

Table 4 reports the 169-case-per-lead values plotted in Figure 3. Because 2020–2021 were available during development, these values are supplementary diagnostics rather than an independent final evaluation. Coverage is the empirical fraction within the central 90% member-quantile interval.

Table 4: Supplementary lead-wise CRPS and central 90% coverage. CRPS is in mm day^{-1} .

Week	CRPS			Coverage		
	Raw	Moment	Adapter	Raw	Moment	Adapter
1	2.848	2.174	2.016	0.156	0.566	0.686
2	3.250	2.721	2.521	0.331	0.690	0.761
3	3.189	2.854	2.706	0.543	0.738	0.797
4	3.176	2.915	2.791	0.657	0.785	0.816
5	3.163	2.918	2.840	0.726	0.817	0.838
6	3.096	2.900	2.815	0.741	0.821	0.841

The adapter has lower CRPS than both comparators at every lead, but its coverage remains below 0.90 throughout. Pointwise regional intervals also favor adapter CRPS in all 24 region–lead cells; no multiplicity adjustment was applied. These results characterize where the primary pooled direction arises and should not be read as six independent confirmation tests.

E Parameter and compute record

The probabilistic adapter has 42,434 parameters. Its three prepared-data training-loop records span 52.0– 82.7 seconds per seed on one NVIDIA A30. These timings exclude FuXi generation, data preparation, target construction, scoring, and artifact verification. They are not end-to-end runtime or energy estimates. Inference latency and peak-memory requirements were not recorded.

The deterministic companion comprises 3 separately trained 144,689-parameter networks. Their prepared-data training loops span 20.9–24.8 seconds per seed on one NVIDIA A30. The recorded values support the description “parameter-efficient”; they do not establish low operational latency or memory use.

F Residual deterministic adapter: complementary evidence

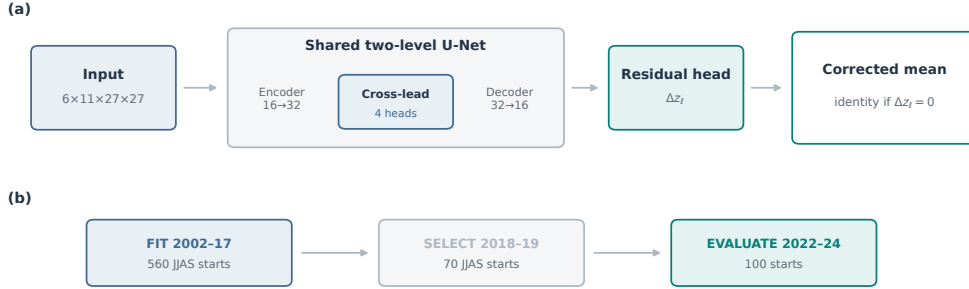


Figure 5: Residual deterministic adapter under its separate contract. Three networks learn cross-lead corrections to the FuXi ensemble mean. This architecture is appendix evidence only and does not produce corrected ensemble members.

This companion fits 560 JJAS starts from 2002–2017 and selects architecture and checkpoints on 70 starts from 2018–2019. Its evaluation uses 100 starts from 2022–2024. Weeks 1–6 use accumulation-start days 0–41, whereas the probabilistic experiment uses end-labelled days +1– +42; scores from the two contracts must not be ranked or combined.

The effective 11 channels per lead are normalized log FuXi weekly ensemble mean and spread, train-only IMD calendar climatology, latitude, longitude, valid-midpoint day-of-year sine and cosine, normalized lead, support, the FuXi minus climatology anomaly, and weekly ensemble-mean temperature. A two-level U-Net combines spatial scales and a four-head Transformer mixes six leads at each bottleneck location. Three residual fields are averaged before one reconstruction. There is no attention-disabled ablation.

Table 5: Detailed deterministic results under the separate contract. Errors and signed bias are mm day^{−1}.

Method	RMSE	MAE	Bias	ACC
Raw FuXi mean	5.723	3.822	-0.225	0.276
Train-only log-bias	5.365	3.543	-0.824	0.337
Learned residual	5.239	3.453	-0.858	0.364
Mean-constrained diagnostic	5.227	3.538	-0.216	0.360

Relative to raw FuXi, the learned residual reduces RMSE by 8.45% [7.27,9.51] and MAE by 9.65%, while ACC changes by +0.0889 [+0.0622,+0.1160]. Its signed-bias change is -0.633 mm day^{−1}, making the already negative bias more negative. The observation-free mean constraint restores the fixed-area regional mean to raw by construction; it is a diagnostic, not a selected forecasting model, and does not show that raw FuXi bias has been removed.

Thus the deterministic branch supports only a complementary statement: a small regional network can improve mean-field error and pattern skill, but it does not calibrate ensemble uncertainty and does not resolve bias.

G Additional regional and observation-target robustness

G.1 Regional CRPS detail

Table 6 resolves the spatial summary into four IMD homogeneous meteorological regions. These are development-exposed 2020–2024 diagnostics with 169 lead-specific cases, pointwise intervals, and no multiplicity adjustment; they are not independent regional replications.

Adapter CRPS skill is positive in every displayed region and lead. In the full four-region by six-lead analysis, CRPS and ensemble-mean RMSE intervals favor the adapter in 24/24 and 24/24 cells,

Table 6: Regional raw CRPS and adapter skill over raw FuXi at four representative leads. CRPS is mm day^{-1} and skill is percent reduction.

Region	Week 1		Week 2		Week 4		Week 6	
	Raw	Skill	Raw	Skill	Raw	Skill	Raw	Skill
Northwest	2.187	27.85%	2.343	19.90%	2.282	10.22%	2.253	8.02%
Central	3.035	24.16%	3.635	18.29%	3.606	8.19%	3.526	5.58%
South Peninsula	2.546	28.56%	3.011	22.55%	2.906	11.12%	2.832	7.82%
East & Northeast	4.309	39.46%	4.770	32.56%	4.596	22.08%	4.389	17.63%
All India	2.848	29.23%	3.250	22.45%	3.176	12.12%	3.096	9.07%

respectively; ACC does so in 13/24. Shared dates and weather regimes mean these counts describe consistency, not 24 independent confirmations.

G.2 Independent station-target sensitivity

The residual deterministic adapter was also scored without retraining against an independent quality-controlled station network: 30 JJAS 2024 initializations, approximately 319 stations, and approximately 96 supported common-grid cells. Table 7 reports the fixed secondary comparison.

Table 7: Independent station-target sensitivity under the separate deterministic contract. Positive RMSE/MAE effects are error reductions; ACC and signed-bias effects are adapter minus raw.

Metric	Raw	Adapter	Paired effect [95%]
RMSE (mm day^{-1})	7.942	7.493	+0.449 [+0.389,+0.507] reduction
MAE (mm day^{-1})	—	—	+0.391 [+0.360,+0.424] reduction
ACC	0.403	0.431	+0.0276 [+0.0150,+0.0390]
Signed bias (mm day^{-1})	—	—	-0.593 [-0.630,-0.562]

The station target retains the RMSE and ACC direction but also confirms a more negative signed bias. Point-to-cell representativeness and a three-hour rain-day convention mismatch limit interpretation. This is a one-season sensitivity check for the deterministic branch, not a second independent evaluation of the probabilistic adapter.

G.3 Additional lead-wise probabilistic diagnostics

Table 8 complements Table 4 with ensemble-mean error, pattern agreement, and dispersion. It uses the same development-exposed 169-case-per-lead cohorts. Spread/error is the ratio of pooled ensemble spread to pooled ensemble-mean error; for 50 exchangeable members under the implemented variance convention, the finite-ensemble reference is approximately 0.98.

Table 8: Additional lead-wise diagnostics. RMSE is mm day^{-1} ; skill is adapter RMSE reduction relative to raw. Spread/error columns are raw, moment, and adapter.

Week	RMSE			ACC		Spread/error		
	Raw	Adapter	Skill	Raw	Adapter	Raw	Moment	Adapter
1	4.956	4.238	14.5%	0.618	0.658	0.140	1.536	0.745
2	5.988	5.260	12.2%	0.404	0.436	0.299	1.160	0.959
3	6.237	5.631	9.7%	0.306	0.334	0.492	0.898	1.163
4	6.389	5.766	9.8%	0.219	0.246	0.612	0.909	1.019
5	6.486	5.875	9.4%	0.174	0.177	0.681	0.908	0.965
6	6.371	5.810	8.8%	0.143	0.179	0.690	0.936	0.927

RMSE is lower and ACC is higher for the adapter at every lead, while spread/error moves substantially toward its finite-ensemble reference. The adapter is not uniformly closest to 0.98 and coverage remains subnominal, so these diagnostics support distributional improvement rather than complete reliability.

H Additional comparator and prior-work context

H.1 Fixed multi-model mean sensitivity

On the same development-exposed 169-case-per-lead 2020–2024 cohorts, the calibrated FuXi ensemble mean has lower RMSE than a fixed equal-weight mean of six forecast systems at every lead. Table 9 reports the paired result.

All six RMSE intervals exclude zero, and the calibrated-mean ACC advantage is resolved at 5 of six leads. This is ancillary deterministic context, not a probabilistic baseline: the multi-model

Table 9: Calibrated FuXi ensemble-mean RMSE reduction relative to the fixed multi-model mean. Intervals are descriptive pointwise 95% intervals from the primary block-16 analysis; positive values favor calibrated FuXi.

Week	RMSE reduction (%)	95% interval
1	9.64	[7.63,11.54]
2	4.15	[1.84,6.31]
3	4.92	[2.67,7.12]
4	6.21	[3.85,8.34]
5	6.84	[4.88,8.79]
6	7.55	[5.65,9.36]

mean contains raw FuXi, does not return exchangeable calibrated members, and has no matched CRPS comparison. Shared cases and model errors also prevent reading the six leads as independent confirmations.

H.2 Positioning and remaining comparison gap

The closest India-specific neural comparator applies U-Net correction to three S2S systems and calibrates rainfall tercile probabilities through combined Weeks 3–4 [7]. Earlier Indian extended-range work used grand ensembles to improve monsoon dispersion and probabilistic prediction [15]. The present output contract is complementary: one fixed global AI ensemble is transformed into continuous corrected members through Week 6, rather than into tercile probabilities or a multi-system mean.

Cross-lead U-Nets and distribution-guided generative calibration provide stronger neural precedents than the structured moment comparator used here [14, 13]. Together with permutation-invariant post-processing and ensemble transformers cited in the main text, they make a broad priority claim inappropriate. The defensible contribution is the parameter-efficient member-equivariant map, India evaluation, and explicit alignment and scoring verification; it does not establish neural S2S postprocessing priority.

A stronger follow-up should compare censored precipitation regression and an ensemble-aware neural baseline on the identical fixed protocol, then test member count, capacity, reconstruction consistency, threshold reliability, and genuinely prospective agency use. Those are scientific comparison gaps, not claims supplied by the present appendix.

I Operational ERPAS matched-valid-time sensitivity

ERPAS is the most relevant India-specific operational reference available in the archive. We therefore compare it with raw FuXi and the unchanged residual deterministic adapter on the strict intersection of 26 cases shared at every Week 1–4 lead (14 in 2023 and 12 in 2024). ERPAS initializes Wednesday; FuXi and its adapter initialize the following Thursday, giving them 24 hours newer initial-condition information. Valid periods are identical, but information age is not.

Table 10: Operational matched-valid-time sensitivity against ERPAS. ACC effects are adapter minus ERPAS; RMSE effects are positive error reductions relative to ERPAS in mm day^{-1} . Brackets are descriptive pointwise 95% intervals from 10,000 paired year-stratified circular-block draws of 4 starts, without multiplicity adjustment.

ACC				
Week	ERPAS	Raw FuXi	Residual adapter	Δ vs. ERPAS [95%]
1	0.464	0.548	0.635	+0.171 [+0.120,+0.226]
2	0.206	0.306	0.381	+0.175 [+0.068,+0.270]
3	0.037	0.191	0.265	+0.229 [+0.141,+0.312]
4	0.026	0.078	0.162	+0.136 [+0.018,+0.260]
RMSE (mm day^{-1})				
Week	ERPAS	Raw FuXi	Residual adapter	Reduction vs. ERPAS [95%]
1	5.883	5.443	4.854	1.030 [0.781,1.308]
2	6.943	6.744	6.031	0.913 [0.548,1.293]
3	7.511	7.061	6.504	1.006 [0.598,1.388]
4	7.168	7.013	6.389	0.780 [0.511,1.055]

The residual adapter was neither retrained nor retuned for this comparison. All eight adapter-versus-ERPAS intervals in Table 10 lie on the favorable side of zero, establishing a consistent result for this fixed small cohort. It does not establish universal superiority: the evidence is post-hoc, contains only two seasons, and gives FuXi a one-day information-age advantage. The comparison also uses end-labelled +1–+28 targets and a common 1991–2019 IMD anomaly normal, so its absolute values should not be combined with the main deterministic table. ERPAS provides no Week 5–6 product in this archive. No 2025 target was opened.

J Descriptive spatial footprint

Figure 4 in the main text reuses the fixed, reviewed spatial diagnostic from the development-exposed 2020–2024 cohorts. It localizes the aggregate score changes and illustrates two ranked cases; it is not part of the primary retrospective evaluation and supplies no pixelwise significance test.