
IndiaS2S-Bench: A Comprehensive Benchmark of Subseasonal to Seasonal Forecasting over India

Ayush Raj
IISER Pune, India
Ashoka University, India

Saptarishi Dhanuka
Ashoka University
Sonipat, India

Manmeet Singh
Western Kentucky University
Bowling Green, USA

Bedartha Goswami
IISER Pune
Pune, India

Parthasarathi Mukhopadhyay
Ashoka University
Sonipat, India

Sandeep Juneja
Ashoka University
Sonipat, India

Abstract

Subseasonal-to-seasonal forecasting over India is critical for many fields, but cannot be evaluated uniformly since usefulness changes with lead, season, region, target, metric, and verification data. We present IndiaS2S-Bench, a unified comparison of archived operational numerical weather prediction (NWP) forecasts and retrospective ERA5-initialized AI and hybrid forecasts. We score precipitation and temperature over 2020-2024 at Weeks 1-6 lead times. All learned models are evaluated after their documented training cutoff, so the benchmark is out of training sample but not wholly untouched or operationally equivalent. We also compute a simple average, and find it scores the highest on the Anomaly Correlation Coefficient (ACC) metric for Week 1 across two different ground truth targets. By Week 6, FuXi-S2S has the highest precipitation ACC and the European model the highest temperature ACC. Root-mean-square error (RMSE), mean absolute error (MAE), and the native fair continuous ranked probability score (CRPS) give us different orderings, which underscores the need to take a comprehensive approach to evaluations. Surprisingly for precipitation, the multi-model ensemble is able to beat the climatology only at Weeks 1-2. We also use India Meteorological Department (IMD) gauge data, which agrees on evaluations at early leads but changes the long-lead rainfall picture. The benchmark's contribution is a practical selection map showing which configuration fits a target, lead, region, season, and forecast objective, without any operational or otherwise claim on real-world usefulness or performance.

1 Why India needs a regional S2S benchmark

Subseasonal-to-seasonal (S2S) prediction bridges weather and seasonal horizons. This matters for India because farmers, reservoir operators, and disaster planners need time to prepare for unusually wet, dry, hot, or cool weeks. In a warming climate, these risks do not affect every region or season in the same way, so a forecast that looks good only in the national average can still fail where a decision is made [1]. Forecast quality also depends on lead and target [2–4]. Global benchmarks standardize evaluation [5–7], and India-focused studies document distinct regional and monsoon behavior [8, 1, 9]. Modern systems span NWP, AI, and hybrid designs [10–13]. IndiaS2S-Bench complements them with shared dates, weekly windows, spatial supports, observations, and metrics over India, while keeping retrospective AI runs distinct from operational services. Table 1 shows that the contribution is not geographic alone: the benchmark adds homogeneous-region and seasonal stratification, independent gauge verification, and an explicit operational or retrospective label for every system. Appendix N lists all abbreviations.

Table 1: Scope of IndiaS2S-Bench relative to the closest global benchmarks, WeatherBench 2 [6] and ChaosBench [7].

Aspect	WeatherBench 2	ChaosBench	IndiaS2S-Bench
Domain	global	global	All India and four IMD homogeneous regions
Horizon	medium range, 1–14 days	subseasonal to seasonal	Weeks 1–6
Main verifying data	reanalysis and analyses	ERA5 with ocean, ice, and land reanalyses	IMERG and ERA5, plus IMD gauge rainfall
Metrics	deterministic and probabilistic	deterministic, probabilistic, and physics-based	ACC, RMSE, MAE, bias, native fair CRPS

Contributions.

- A frozen, common-start protocol that scores nine archived operational NWP and retrospective AI/hybrid systems over India for TP and T2M at Weeks 1–6.
- A released target–region–season–lead score cube with deterministic (ACC, RMSE, MAE, bias) and native-member probabilistic (fair CRPS) metrics, climatology and multi-model baselines, and an independent IMD gauge check.
- A selection map rather than a leaderboard. The cube screens candidates for later local calibration and trials. It measures forecast quality, not decision value.

2 Methodology

Forecast systems and evaluation status. Table 2 prevents architecture from being conflated with forecast status. Native ensemble sizes are unequal, so no equal-compute conclusion is available.

Table 2: Fixed comparison setup. TP: total precipitation, T2M: 2-m temperature, MME: multi-model mean.

Systems	Family	Forecast status	Target / native members
CMA, ECMWF, NCEP, UKMO	physics-based	archived operational	TP/T2M; 4, 51–101, 16, 4 ^a
DLESyM-v0/v1, FCN3, FuXi	AI	retrospective, ERA5-init.	target-specific; 1/4, 3, 50
NeuralGCM	hybrid	retrospective, ERA5-init.	TP; 10
MME ₆	equal-system mean	derived benchmark	TP/T2M; six system means

^aNCEP T2M unavailable; ECMWF TP Week 3 is missing from the archive.

Common cohort, weekly aggregation, and references. On common starts (517 TP and 516 T2M), forecast days 1–42 form six weekly means on a 27×27 grid over All India and four IMD homogeneous regions in January–February (JF), March–May (MAM), June–September (JJAS), and October–December (OND). The satellite–gauge IMERG V07B product and ERA5 reanalysis are the primary references [14, 15]. Independently rescored IMD gauge rainfall remains separate [16]. Weekly averaging asks the practical S2S question—whether a week is wetter, drier, hotter, or cooler than usual—while reducing day-to-day noise. Each system’s members are averaged first. MME₆ then gives each system one equal vote, so a model does not receive extra weight merely because it has more members. Averaging can cancel errors when systems fail differently, but it can also smooth extremes, which is why we report several complementary metrics.

Protocol details. Starts are 00 UTC Monday and Thursday initializations, kept only where every system in the cohort has an exact common start. Lead day 1 is the first 24 h after initialization, and Week w averages lead days $7(w - 1) + 1$ to $7w$. Accumulated precipitation is differenced to daily totals in mm day^{-1} . Physics-based archives and FuXi-S2S are produced directly on the 1.5° grid. Other AI outputs use exact node selection or conservative remapping, and observations use spherical cell-overlap averaging. ACC anomalies use each system’s own lead-dependent, leave-one-year-out normal and pre-2020 observation normals (IMERG 2001–2019, IMD 1991–2019, ERA5 1990–2017). Every learned model’s documented weight-training period ends by 2018 (Table 6). Appendix B gives the full details and score formulas.

(a) Precipitation · IMERG (N=517)

CMA	0.393	0.197	0.074	0.047	0.025	0.024
DLESyM-v0	0.517	0.270	0.140	0.082	0.053	0.001
ECMWF	0.569	0.382	—	0.175	0.141	0.096
FuXi-S2S	0.586	0.354	0.213	0.151	0.153	0.109
NCEP	0.434	0.225	0.113	0.072	0.050	0.021
NeuralGCM	0.591	0.382	0.202	0.133	0.063	0.028
UKMO	0.553	0.313	0.161	0.054	0.051	0.045
MME	0.622	0.399	0.228	0.136	0.098	0.057
	W1	W2	W3	W4	W5	W6

(b) 2-m temperature · ERA5 (N=516)

CMA	0.442	0.280	0.129	0.071	0.046	0.040
DLESyM-v0	0.574	0.357	0.190	0.094	0.059	0.034
DLESyM-v1	0.586	0.399	0.229	0.130	0.085	0.073
ECMWF	0.512	0.393	0.311	0.223	0.168	0.142
FCN3	0.577	0.406	0.241	0.105	0.076	0.053
FuXi-S2S	0.559	0.409	0.269	0.190	0.139	0.123
UKMO	0.437	0.310	0.177	0.108	0.080	0.072
MME	0.596	0.471	0.329	0.224	0.168	0.142
	W1	W2	W3	W4	W5	W6

Figure 1: Mean All-India ACC for TP verified against IMERG and T2M against ERA5. Bold values are lead-wise maxima within each target. The dash for ECMWF TP Week 3 marks a field missing from the archive, not a zero or an imputed value. MME is the fixed six-system mean.

Metrics, baselines, and uncertainty. ACC asks whether wet/dry or warm/cool anomalies appear in the right places. RMSE measures amount error, and signed bias measures a systematic tendency. Native-member fair CRPS evaluates the full ensemble with the finite-ensemble correction [17]. Deterministic DLESyM-v0 and the MME are ineligible rather than assigned pseudo-members. For selected paired MME contrasts, 95% intervals use 2,000 ordinary moving-block resamples with a primary block length of 16 ordered starts, with block lengths 8 and 24 as sensitivities. Intervals are pointwise, descriptive, and not multiplicity-adjusted.

3 What the benchmark reveals

Table 3: All-India scorecard at Weeks 1, 3, and 6: ACC $\uparrow$, RMSE $\downarrow$, and native fair CRPS $\downarrow$. Bias and all leads are in Appendix D. Bold marks the metric-optimal unrounded point estimate, not statistical separation. *n/e*: ineligible, —: unavailable.

Model	Week 1			Week 3			Week 6		
	ACC	RMSE	fCRPS	ACC	RMSE	fCRPS	ACC	RMSE	fCRPS
(a) TP-IMERG (mm day ⁻¹)									
MME (no ECMWF)	0.622	2.429	<i>n/e</i>	0.228	3.181	<i>n/e</i>	0.057	3.363	<i>n/e</i>
CMA	0.393	3.973	1.865	0.074	4.621	1.841	0.024	4.451	1.711
ECMWF	0.569	3.117	1.270	—	—	—	0.096	3.368	1.496
UKMO	0.553	3.461	1.514	0.161	4.047	1.622	0.045	4.151	1.670
NCEP	0.434	3.470	1.767	0.113	3.947	1.902	0.021	3.926	1.844
FuXi-S2S	0.586	2.687	1.435	0.213	3.471	1.613	0.109	3.425	1.523
NeuralGCM	0.591	2.541	1.115	0.202	3.235	1.410	0.028	3.516	1.555
DLESyM-v0	0.517	3.287	<i>n/e</i>	0.140	4.236	<i>n/e</i>	0.001	4.523	<i>n/e</i>
(b) T2M-ERA5 (°C)									
MME	0.596	1.354	<i>n/e</i>	0.329	1.743	<i>n/e</i>	0.142	1.885	<i>n/e</i>
CMA	0.442	2.116	1.561	0.129	2.894	1.889	0.040	2.979	1.836
ECMWF	0.512	1.863	1.068	0.311	1.997	1.113	0.142	2.167	1.205
UKMO	0.437	1.780	1.167	0.177	2.240	1.259	0.072	2.389	1.308
FuXi-S2S	0.559	1.654	0.988	0.269	2.060	1.184	0.123	2.202	1.190
DLESyM-v0	0.574	1.453	<i>n/e</i>	0.190	2.137	<i>n/e</i>	0.034	2.366	<i>n/e</i>
DLESyM-v1	0.586	1.412	0.743	0.229	1.960	0.997	0.073	2.202	1.080
FCN3	0.577	1.700	0.903	0.241	2.170	1.152	0.053	2.249	1.160

The MME is strongest on magnitude, not always pattern. Figure 1 and Table 3 summarize All-India skill. The MME attains the lowest All-India RMSE point estimate in 11/12 target–lead

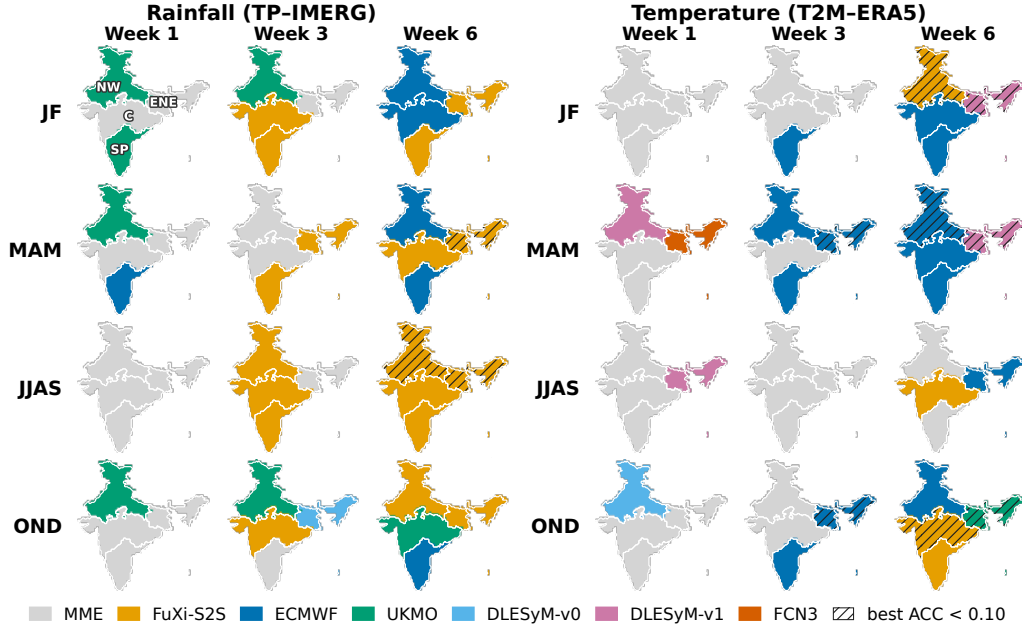


Figure 2: Regional selection map. Each IMD homogeneous region (NW: Northwest, C: Central, SP: South Peninsula, ENE: East & Northeast) is coloured by the system, including the fixed MME_6 , with the highest ACC point estimate for that season and lead over 2020–2024. Rainfall is verified against IMERG and temperature against ERA5. Hatching marks regions where even the highest ACC is below 0.10. Selections are descriptive point estimates without uncertainty or multiplicity control; all values are in Appendix F.

cells, and MAE shows the same pattern (Table 9). Selected paired MME-minus-best-individual ACC contrasts are +0.031 [+0.015, +0.045] for TP Week 1, +0.010 [+0.001, +0.022] for T2M Week 1. Both Week-3 intervals include zero. At TP Week 6 the contrast is -0.053 [-0.097, -0.017]. These pointwise interpretations persist under 8- and 24-start blocks. Model-selection uncertainty is omitted.

Probability scores favor different native ensembles. Among eligible and available native ensembles, the All-India minimum is NeuralGCM for TP Weeks 1–3, ECMWF for TP Weeks 4–6, and DLESyM-v1 for T2M at all leads. In 96 region–season–lead cells, NeuralGCM/ECMWF attain 48/28 TP minima and DLESyM-v1 attains 55/96 T2M minima (Appendix Fig. 4c). These are absolute point estimates, not CRPS skill score (CRPSS), equal-compute, or separation claims. ECMWF TP Week 3 is unavailable in all 16 regional–season cells.

Long-lead rainfall remains difficult. Climatology is the historical average for a location and time of year: a simple forecast of “usual conditions.” Beating it matters because a complex model should add information beyond what normally happens. TP- MME_6 has lower RMSE than this baseline at Weeks 1–2 and higher RMSE at Weeks 3–6. The MME_7 -available sensitivity has lower RMSE than MME_6 at every available lead and preserves that classification. Week 3 cannot be assessed because the ECMWF archive is unavailable. These are point estimates without intervals. Climatology ACC is undefined because its anomaly is always zero, and it is not a CRPS denominator.

4 Regional findings

No single system leads everywhere. Figure 2 maps the system with the highest ACC in each IMD homogeneous region for every season at Weeks 1, 3, and 6. At Week 1 the MME leads most regions for both targets. For JJAS rainfall at Weeks 3–6, FuXi-S2S leads almost everywhere. ECMWF leads most spring temperature regions from Week 3 and about half of the winter and spring rainfall regions at Week 6. Hatching shows where even the best system has ACC below 0.10, mostly at Week 6.

The monsoon rainfall result persists across rainfall references. FuXi-S2S attains the highest unrounded JJAS Week-3–6 ACC point estimate in 19/20 IMERG cells and 18/20 IMD cells (Appendix Fig. 4a, b). The identity of the maximizing model agrees in 19/20 cells; the only disagreement is East–Northeast at Week 5. These are 169-start seasonal point estimates without multiplicity control.

5 Using the benchmark

Table 4: Selection map from All-India point estimates (Sections 3–4). Entries name the configuration with the metric-optimal estimate, not a statistically separated winner.

Forecast objective	Configuration indicated by the benchmark
Week-1 pattern and amount, TP and T2M	MME ₆ (highest ACC; lowest RMSE and MAE)
TP amount at Weeks 3–6	observation climatology (lower RMSE than MME ₆)
Long-lead TP pattern (Week 6; JJAS Weeks 3–6)	FuXi-S2S, under both IMERG and IMD
Long-lead T2M pattern, single system	ECMWF at Week 6
TP ensemble (fair CRPS)	NeuralGCM at Weeks 1–3, ECMWF at Weeks 4–6
T2M ensemble (fair CRPS)	DLESym-v1 at all leads

Algorithm 1: Compare a candidate forecast system q

1. **Freeze the comparison.** Keep the released 2020–2024 cohort (517 TP or 516 T2M starts), grid, references, masks, weeks, and metrics. Record missing cases and never redefine the cohort.
2. **Build case-wise weekly fields.** For each start D , member m , and Week w ,

$$Z_{Dmw} = \frac{1}{7} \sum_{d=7(w-1)+1}^{7w} X_{Dmd}, \quad F_{Dw} = \frac{1}{M} \sum_{m=1}^M Z_{Dmw}.$$

3. **Score before pooling.** Compute $S_{qDw}^{(k)}$ for each start. ACC uses the candidate’s leave-one-year-out normal, RMSE, MAE, and bias use raw fields, and fair CRPS uses the native members Z_{Dmw} .
4. **Pool, then compare.** Report

$$\bar{S}_{qw}^{(k)} = \frac{\sum_D a_{Dw} S_{qDw}^{(k)}}{\sum_D a_{Dw}}$$

with its coverage, and compare systems only on identical available cases with paired moving blocks.

6 Limitations and availability

The record is retrospective through 2024. Regional winner counts are descriptive and carry no uncertainty estimate. A shorter matched ERPAS extension is reported separately in Appendix I. Future versions will add variables, a probabilistic-climatology CRPSS, and reliability diagnostics. The benchmark (evaluation code, result tables, figures, and instructions) is being prepared for public release and is available from the authors on request until then. Raw forecasts and model weights are not redistributed (Appendix L).

Acknowledgments and Disclosure of Funding

We thank the High Performance Computing facility at the Indian Institute of Science Education and Research (IISER) Pune for computational resources.

References

- [1] Iqura Malik and Vimal Mishra. Sub-seasonal to seasonal (S2S) prediction of dry and wet extremes for climate adaptation in India. *Climate Services*, 34:100457, 2024.
- [2] Frédéric Vitart, Christelle Ardilouze, Ana Bonet, Anca Brookshaw, Mingyue Chen, Cristian Codorean, Michel Déqué, Laura Ferranti, Edmond Fucile, Manuel Fuentes, et al. The subseasonal to seasonal (S2S) prediction project database. *Bulletin of the American Meteorological Society*, 98(1):163–173, 2017.
- [3] June-Yi Lee, Bin Wang, Matthew C. Wheeler, Xiouhua Fu, Duane E. Waliser, and In-Sik Kang. Real-time multivariate indices for the boreal summer intraseasonal oscillation over the Asian summer monsoon region. *Climate Dynamics*, 40:493–509, 2013.
- [4] Shang Zhou et al. Predictability limit of monsoon intraseasonal precipitation: An implication of central Indian Ocean mode. *Frontiers in Marine Science*, 8:809798, 2021.
- [5] Stephan Rasp, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11):e2020MS002203, 2020.
- [6] Stephan Rasp, Stephan Hoyer, Alexander Meroze, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Robert Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.
- [7] Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. ChaosBench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- [8] Xiangwen Liu, Song Yang, Jianglong Li, Weihua Jie, Liang Huang, and Weizong Gu. Subseasonal predictions of regional summer monsoon rainfall over tropical Asian oceans and land. *Journal of Climate*, 28(24):9583–9605, 2015.
- [9] Aheli Das, Dondeti Pranay Reddy, and Somnath Baidya Roy. Improving the skill of subseasonal to seasonal (S2S) wind speed forecasts over India using statistical and machine learning methods. *Journal of Geophysical Research: Machine Learning and Computation*, 1:e2024JH000187, 2024.
- [10] 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*, 15:6425, 2024.
- [11] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, et al. Neural general circulation models for weather and climate. *Nature*, 632:1060–1066, 2024.
- [12] Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. FourCast-Net 3: A geometric approach to probabilistic machine-learning weather forecasting at scale. *arXiv preprint arXiv:2507.12144*, 2025. Version 2, 18 July 2025.
- [13] Nathaniel Cresswell-Clay, Bowen Liu, Dale Durran, Andy Liu, Zachary I. Espinosa, Raul Moreno, and Matthias Karlbauer. A deep learning earth system model for stable and efficient simulation of the current climate. *arXiv preprint arXiv:2409.16247*, 2024.
- [14] G. J. Huffman, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan. GPM IMERG Final precipitation L3 1 day 0.1 degree x 0.1 degree V07, 2023. Digital science data, version 07, released 12 July 2023.

- [15] Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.
- [16] 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 degree x 0.25 degree) 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.
- [17] Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014.

A Scope, provenance, and model configurations

The physics-based inputs are archived operational NWP outputs from forecasting centres. In contrast, FuXi-S2S, DLESyM-v0/v1, FCN3, and NeuralGCM are locally generated retrospective forecasts initialized from ERA5 with fixed model versions. They are not provider-issued operational forecasts. Table 5 records target and native-member availability.

Table 5: Fixed model configurations. Physics-based rows are archived operational outputs, whereas AI and hybrid rows are retrospective forecasts.

System	Family	Comparable target	Evaluated members	Evaluation note
CMA	physics-based	TP,T2M	4	member mean
DLESyM-v0	AI	TP,T2M	1	deterministic
DLESyM-v1	AI	T2M	4	T2M MME member
ECMWF	physics-based	TP,T2M	51–101	TP W3 archive gap
FCN3	AI	T2M	3	one known missing start
FuXi-S2S	AI	TP,T2M	50	member mean
NCEP	physics-based	TP	16	native T2M unavailable
NeuralGCM	hybrid AI	TP	10	member mean
UKMO	physics-based	TP,T2M	4	member mean
MME ₆	equal system	TP,T2M	6 systems	fixed membership

FuXi-S2S uses 50 stochastic draws and two ERA5 daily states; DLESyM uses staged ERA5 states on its native HEALPix grid; FCN3 uses three fixed stochastic seeds; and NeuralGCM uses ten fixed seeds and persisted sea-surface forcing [10, 13, 12, 11]. All source-specific transformations precede ingestion into one standardized weekly evaluation dataset. Nothing here compares forecast latency, compute, initial-condition age, staffing, or service reliability.

B Reproducibility details

Model assets and training periods. Table 6 lists the model version, documented training period, and initial state of every learned configuration. Training periods are quoted from the cited papers or public model cards. Where two public sources disagree, both are reported. The developer evaluation or validation periods of FuXi-S2S (2017–2021) and FCN3 (2018–2021, paper) overlap 2020–2021, so those years are outside weight training but not untouched by model development. No documented training period extends into 2020–2024. Exact release identifiers are recorded with the benchmark code.

The physics-based systems are real-time operational forecasts from the S2S Prediction Project archive, obtained through the ECMWF Data Store and initialized at 00 UTC. Reforecasts are not mixed in.

Starts, leads, and units. Starts follow a Monday/Thursday calendar during 2020–2024 and are retained only where every system in the target cohort has an exact common initialization (517 TP and 516 T2M starts; FCN3 lacks 2021-03-08). For start D, lead day d is the half-open interval $[D + d - 1, D + d)$ and Week w averages lead days $7(w - 1) + 1$ to $7w$. IMERG pairs lead day 1 with the initialization day, whereas end-labelled IMD gauge days are paired with a one-day offset.

Table 6: Learned-model versions and training periods. D is the start date at 00 UTC.

System	Model version	Documented training period	Initial state	Members
FuXi-S2S	public release [10]	ERA5 1950–2016	ERA5 daily means, D–2 and D–1	50
DLESyM-v0	NVIDIA DLESyM v0 (ISCCP–ERA5)	ERA5 and ISCCP 1983–2016 (model card)	ERA5, D–48 h to D	1
DLESyM-v1	NVIDIA DLESyM v1 (ERA5), four paired checkpoints	ERA5 1980–2015 (current model card)	ERA5, D–48 h to D	4
FCN3	NVIDIA FourCastNet 3, three fixed seeds	1980–2016 (paper) or 1980–2015 (model card)	ERA5 at D	3
NeuralGCM	stochastic-precipitation model, 2.8° (v1.2.2)	ERA5 and IMERG 2001–2018	ERA5 at D	10

Accumulated NWP precipitation is converted to disjoint daily increments by first differences, and all TP is expressed in mm day^{-1} .

Common grid and regridding. All scores use one 27×27 grid at 1.5° spacing ($0\text{--}39^\circ\text{N}$, $60\text{--}99^\circ\text{E}$). Physics-based archives and FuXi-S2S are requested or produced directly on this grid and are not interpolated. FCN3 uses exact node selection. DLESyM is mapped from its native HEALPix (nside 64) grid to 0.25° with the earth2grid library. T2M then uses exact node selection and TP uses spherical cell-overlap conservative weights. NeuralGCM TP is regridded conservatively from its native Gaussian grid. IMERG, IMD, and ERA5 are remapped by first-order spherical cell-overlap averaging, which prevents ocean values from entering coastal India cells.

Climatologies. Each forecast system’s ACC normal is computed from its own ensemble mean, separately for every lead, with a centered 31-day window on the initialization day of year and five-fold leave-one-year-out (LOYO) over 2020–2024. Observation normals use pre-verification periods: IMERG 2001–2019, IMD 1991–2019 (both centered circular 31-day windows on a 366-day calendar), and the WeatherBench 2 ERA5 1990–2017 61-day-window climatology for T2M. The deterministic climatology baseline in Table 21 forecasts the weekly-averaged IMERG normal.

Scores. For grid cell i , the weight a_i is the spherical cell area times the India mask fraction times the represented observation fraction for the week. With forecast f_i , observation o_i , anomalies $f'_i = f_i - c_i^f$ and $o'_i = o_i - c_i^o$, and weighted means $\langle \cdot \rangle$,

$$\text{ACC} = \frac{\sum_i a_i (f'_i - \langle f' \rangle)(o'_i - \langle o' \rangle)}{\sqrt{\sum_i a_i (f'_i - \langle f' \rangle)^2 \sum_i a_i (o'_i - \langle o' \rangle)^2}}, \quad \text{RMSE} = \sqrt{\frac{\sum_i a_i e_i^2}{\sum_i a_i}},$$

with $e_i = f_i - o_i$, $\text{MAE} = \sum_i a_i |e_i| / \sum_i a_i$, and $\text{bias} = \sum_i a_i e_i / \sum_i a_i$. For $M \geq 2$ members sorted as $x_{(1)} \leq \dots \leq x_{(M)}$, the cellwise fair CRPS is

$$\text{fCRPS} = \frac{1}{M} \sum_{m=1}^M |x_m - o| - \frac{1}{M(M-1)} \sum_{k=1}^M (2k - M - 1) x_{(k)},$$

which is then area-weighted like MAE. Every score is computed per start and lead and then pooled as the arithmetic mean over starts.

Execution outline. The evaluation runs in a fixed order: (1) download or generate each system on the fixed start calendar; (2) standardize units, accumulations, grid, and ensemble members; (3) build weekly forecast and observation fields; (4) score each start, lead, region, and season; (5) pool the scores and run the paired bootstrap; and (6) regenerate every table, figure, and reported number from the pooled scores. Automated checks confirm coordinates, units, value ranges, and completeness at each step. Algorithm 1 (Section 5) gives the procedure for adding a new system.

C Metrics, supports, and uncertainty

Score formulas are given in Appendix B. Fair CRPS applies the finite-ensemble correction of Ferro [17]. Bias is printed with its sign, and any “closest to zero” formatting uses absolute bias before rounding.

All India and the four Northwest, Central, South Peninsula, and East–Northeast supports use the official IMD homogeneous-region masks. “IMD region” in this geometric sense does not mean that IMD rainfall is the verifying observation: the primary TP–IMERG and T2M–ERA5 tracks use the same masks, including the regional fair-CRPS atlas below.

Overlapping starts motivate paired ordinary moving-block resampling. The reported pointwise intervals use 2,000 repetitions and a primary block length of 16 ordered starts; 8 and 24 are prespecified sensitivities. Across all 496 available contrasts, 13 ancillary interval sign/zero decisions change with block length; the five main-paper contrasts retain their stated interpretations. The manuscript therefore makes no global robustness or familywise-significance claim.

D All-India results for every model and lead

Tables 7–8 print every comparable system at all six leads. ACC is maximized, RMSE and fair CRPS minimized, and signed bias judged by distance from zero. Fair CRPS is absolute rather than a skill score. The deterministic DLESyM-v0 and the equal-system MME are marked ineligible; ECMWF TP Week 3 is unavailable under the five-year evaluation.

Table 7: All-season TP–IMERG results for every model and lead. Metric-optimal cells are bolded from unrounded values.

Model	W1	W2	W3	W4	W5	W6
ACC ↑						
MME (no ECMWF)	0.622	0.399	0.228	0.136	0.098	0.057
CMA	0.393	0.197	0.074	0.047	0.025	0.024
ECMWF	0.569	0.382	—	0.175	0.141	0.096
UKMO	0.553	0.313	0.161	0.054	0.051	0.045
NCEP	0.434	0.225	0.113	0.072	0.050	0.021
FuXi-S2S	0.586	0.354	0.213	0.151	0.153	0.109
NeuralGCM	0.591	0.382	0.202	0.133	0.063	0.028
DLESyM-v0	0.517	0.270	0.140	0.082	0.053	0.001
RMSE ↓						
MME (no ECMWF)	2.429	2.955	3.181	3.282	3.346	3.363
CMA	3.973	4.566	4.621	4.532	4.520	4.451
ECMWF	3.117	3.382	—	3.302	3.332	3.368
UKMO	3.461	3.957	4.047	4.141	4.181	4.151
NCEP	3.470	3.878	3.947	3.964	3.955	3.926
FuXi-S2S	2.687	3.318	3.471	3.467	3.470	3.425
NeuralGCM	2.541	3.011	3.235	3.390	3.472	3.516
DLESyM-v0	3.287	3.893	4.236	4.351	4.462	4.523
 bias ↓						
MME (no ECMWF)	+0.077	-0.003	+0.005	-0.014	-0.045	-0.045
CMA	-0.120	+0.166	+0.358	+0.415	+0.395	+0.398
ECMWF	+0.389	+0.316	—	+0.191	+0.151	+0.135
UKMO	+0.665	+0.257	+0.208	+0.136	+0.134	+0.121
NCEP	-0.176	-0.379	-0.483	-0.533	-0.503	-0.437
FuXi-S2S	-0.202	-0.158	-0.076	-0.014	-0.026	-0.121
NeuralGCM	-0.064	-0.165	-0.137	-0.109	-0.166	-0.175
DLESyM-v0	+0.359	+0.261	+0.161	+0.019	-0.106	-0.055
fair CRPS ↓						
MME (no ECMWF)	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>
CMA	1.865	1.934	1.841	1.742	1.718	1.711
ECMWF	1.270	1.408	—	1.462	1.481	1.496
UKMO	1.514	1.654	1.622	1.671	1.679	1.670
NCEP	1.767	1.909	1.902	1.900	1.874	1.844
FuXi-S2S	1.435	1.653	1.613	1.565	1.544	1.523
NeuralGCM	1.115	1.309	1.410	1.492	1.535	1.555
DLESyM-v0	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>

Table 8: All-season T2M–ERA5 results for every model and lead. Metric-optimal cells are bolded from unrounded values.

Model	W1	W2	W3	W4	W5	W6
ACC $\uparrow$						
MME	0.596	0.471	0.329	0.224	0.168	0.142
CMA	0.442	0.280	0.129	0.071	0.046	0.040
ECMWF	0.512	0.393	0.311	0.223	0.168	0.142
UKMO	0.437	0.310	0.177	0.108	0.080	0.072
FuXi-S2S	0.559	0.409	0.269	0.190	0.139	0.123
DLESyM-v0	0.574	0.357	0.190	0.094	0.059	0.034
DLESyM-v1	0.586	0.399	0.229	0.130	0.085	0.073
FCN3	0.577	0.406	0.241	0.105	0.076	0.053
RMSE $\downarrow$						
MME	1.354	1.573	1.743	1.834	1.871	1.885
CMA	2.116	2.676	2.894	2.943	2.997	2.979
ECMWF	1.863	2.021	1.997	2.092	2.142	2.167
UKMO	1.780	2.026	2.240	2.329	2.385	2.389
FuXi-S2S	1.654	1.896	2.060	2.114	2.169	2.202
DLESyM-v0	1.453	1.862	2.137	2.283	2.324	2.366
DLESyM-v1	1.412	1.754	1.960	2.099	2.168	2.202
FCN3	1.700	1.901	2.170	2.216	2.248	2.249
 bias $\downarrow$						
MME	+0.020	+0.212	+0.298	+0.262	+0.254	+0.216
CMA	+0.756	+1.343	+1.519	+1.600	+1.638	+1.587
ECMWF	-0.501	-0.435	-0.399	-0.410	-0.391	-0.369
UKMO	-0.477	-0.108	-0.028	+0.000	+0.045	+0.059
FuXi-S2S	+0.022	+0.099	+0.120	+0.066	-0.115	-0.230
DLESyM-v0	+0.312	+0.292	+0.262	+0.224	+0.218	+0.161
DLESyM-v1	+0.085	-0.011	-0.047	-0.059	-0.056	-0.066
FCN3	+0.237	+0.383	+0.625	+0.378	+0.400	+0.317
fair CRPS $\downarrow$						
MME	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>
CMA	1.561	1.903	1.889	1.852	1.876	1.836
ECMWF	1.068	1.127	1.113	1.165	1.193	1.205
UKMO	1.167	1.215	1.259	1.278	1.304	1.308
FuXi-S2S	0.988	1.148	1.184	1.169	1.176	1.190
DLESyM-v0	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>	<i>n/e</i>
DLESyM-v1	0.743	0.919	0.997	1.058	1.075	1.080
FCN3	0.903	0.990	1.152	1.145	1.148	1.160

Table 9: All-season All-India MAE $\downarrow$ for every model and lead (pooled mean of case-wise MAE). Bold marks the lowest unrounded value. MAE ranks systems much as RMSE does, but differs from fair CRPS (Table 3).

Model	W1	W2	W3	W4	W5	W6
(a) TP-IMERG (mm day⁻¹)						
MME (no ECMWF)	1.510	1.876	2.058	2.145	2.183	2.197
CMA	2.288	2.723	2.857	2.821	2.819	2.820
ECMWF	1.759	2.068	—	2.181	2.195	2.214
UKMO	1.985	2.354	2.472	2.580	2.620	2.605
NCEP	2.137	2.433	2.504	2.521	2.510	2.495
FuXi-S2S	1.563	1.976	2.125	2.173	2.185	2.148
NeuralGCM	1.535	1.882	2.081	2.201	2.249	2.282
DLESyM-v0	2.038	2.421	2.663	2.765	2.819	2.839
(b) T2M-ERA5 (°C)						
MME	0.830	1.075	1.243	1.327	1.356	1.368
CMA	1.636	2.146	2.341	2.394	2.450	2.431
ECMWF	1.243	1.415	1.464	1.548	1.590	1.612
UKMO	1.274	1.476	1.664	1.742	1.800	1.800
FuXi-S2S	1.036	1.308	1.475	1.516	1.557	1.586
DLESyM-v0	0.962	1.351	1.591	1.730	1.754	1.791
DLESyM-v1	0.907	1.231	1.408	1.522	1.577	1.591
FCN3	1.116	1.333	1.606	1.632	1.650	1.665

E Observation-selected high-rainfall weeks

Figure 3 displays the two highest non-overlapping Week-2 valid windows among 517 paired starts, ranked solely by the area-weighted All-India IMERG weekly mean. No forecast field or forecast score enters the selection. The plate is descriptive rather than an event-conditioned skill estimate; the printed ACC and RMSE values reproduce the released same-case metrics.

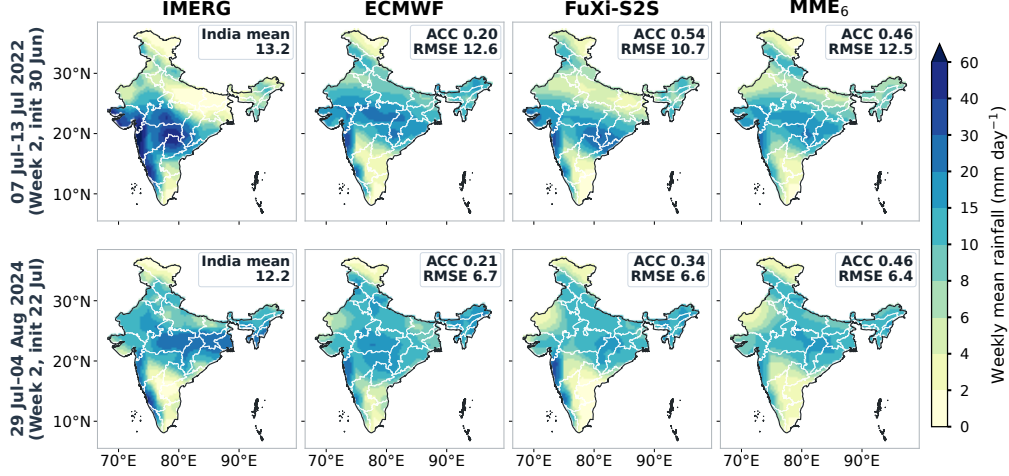


Figure 3: Spatial precipitation fields for two observation-selected high-rainfall weeks. Rows are the selected valid windows and columns are IMERG, ECMWF, FuXi-S2S, and the fixed MME₆. Fields are scored on the common 1.5° India support. For display only, each field is bilinearly interpolated to 0.05° and clipped to the Survey of India boundary, with state lines in white. All panels share one binned color scale, and the darkest bin holds all values above 60 mm day⁻¹. Badges give the IMERG All-India mean and each forecast’s same-case ACC and RMSE. ACC is anomaly-pattern correlation; RMSE is raw rainfall in mm day⁻¹. MME₆ excludes ECMWF. Two cases do not establish extreme-event skill or statistical superiority.

F Regional–seasonal all-model benchmark atlas

The next eight tables expose the benchmark dimension that an All-India leaderboard hides. They print every target-valid model over all four IMD homogeneous geographic masks, all four seasons, and the Week-3 and Week-6 milestones. In total, the atlas contains 256 model–region–season display rows and 2048 metric cells. ACC is maximized; RMSE and native-member fair CRPS are minimized; signed bias is bolded by distance from zero. Bolding uses unrounded values and is descriptive, not a claim of statistical separation.

“IMD mask” denotes geography only: precipitation remains verified against IMERG and temperature against ERA5. TP RMSE, bias, and fair CRPS are in mm day⁻¹; the corresponding T2M quantities are in °C. Values for all six weeks are included in the released result tables. Case partitions are JF/MAM/JJAS/OND = 85/131/169/132 for TP and 85/130/169/132 for T2M. Here *n/e* means probabilistically ineligible and — means unavailable under the fixed archive setup.

Figure 2 in the main text condenses the ACC part of this atlas into a regional selection map.

Table 10: JF TP-IMERG regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME (no ECMWF)	0.175	1.344	+0.354	<i>n/e</i>	0.047	1.415	+0.383	<i>n/e</i>
	CMA	0.088	1.758	+0.151	0.775	0.058	1.989	+0.374	0.752
	ECMWF	—	—	—	—	0.161	1.521	+0.561	0.573
	UKMO	0.267	1.378	+0.351	0.479	0.081	1.576	+0.436	0.550
	NCEP	0.111	1.544	+0.427	0.691	0.028	1.622	+0.414	0.703
	FuXi-S2S	0.119	1.590	+0.339	0.660	0.069	1.493	+0.385	0.587
	NeuralGCM	0.162	1.068	+0.022	0.436	0.122	1.137	+0.123	0.459
	DLESyM-v0	0.147	1.992	+0.836	<i>n/e</i>	0.115	1.820	+0.569	<i>n/e</i>
CI	MME (no ECMWF)	0.251	0.536	+0.107	<i>n/e</i>	0.102	0.530	+0.070	<i>n/e</i>
	CMA	0.141	0.525	-0.144	0.221	0.150	0.520	-0.089	0.192
	ECMWF	—	—	—	—	0.208	0.605	+0.180	0.185
	UKMO	0.200	0.737	+0.198	0.204	0.163	0.762	+0.186	0.210
	NCEP	0.204	0.600	+0.111	0.210	0.162	0.512	-0.010	0.195
	FuXi-S2S	0.271	0.465	-0.018	0.176	0.120	0.477	-0.022	0.187
	NeuralGCM	0.242	0.620	+0.181	0.214	0.082	0.678	+0.231	0.218
	DLESyM-v0	0.225	0.840	+0.316	<i>n/e</i>	0.123	0.638	+0.121	<i>n/e</i>
SP	MME (no ECMWF)	0.258	0.781	+0.162	<i>n/e</i>	0.081	0.869	+0.156	<i>n/e</i>
	CMA	0.200	1.126	+0.226	0.388	0.087	0.974	+0.037	0.309
	ECMWF	—	—	—	—	0.094	0.910	+0.220	0.297
	UKMO	0.205	0.854	+0.154	0.292	0.053	1.043	+0.220	0.331
	NCEP	0.241	0.790	+0.120	0.290	-0.018	0.779	-0.064	0.306
	FuXi-S2S	0.315	0.645	+0.004	0.263	0.169	0.702	-0.099	0.298
	NeuralGCM	0.195	0.902	+0.277	0.324	0.039	1.196	+0.537	0.364
	DLESyM-v0	0.149	1.016	+0.191	<i>n/e</i>	0.139	1.148	+0.305	<i>n/e</i>
E&NE	MME (no ECMWF)	0.248	1.076	+0.557	<i>n/e</i>	0.071	1.127	+0.500	<i>n/e</i>
	CMA	0.015	1.033	-0.199	0.523	0.069	1.074	+0.018	0.440
	ECMWF	—	—	—	—	0.095	1.113	+0.422	0.451
	UKMO	0.240	2.110	+1.343	0.872	0.054	2.018	+1.235	0.774
	NCEP	0.093	1.104	+0.092	0.536	0.034	0.979	-0.049	0.480
	FuXi-S2S	0.191	1.632	+0.636	0.712	0.133	1.633	+0.676	0.645
	NeuralGCM	0.086	0.926	+0.127	0.425	-0.059	0.956	+0.086	0.446
	DLESyM-v0	0.107	2.104	+1.343	<i>n/e</i>	0.008	2.016	+1.035	<i>n/e</i>

Table 11: MAM TP-IMERG regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME (no ECMWF)	0.225	1.621	+0.223	<i>n/e</i>	0.122	1.754	+0.333	<i>n/e</i>
	CMA	0.094	2.141	+0.092	0.973	0.098	2.266	+0.283	0.959
	ECMWF	—	—	—	—	0.222	1.597	-0.049	0.790
	UKMO	0.162	1.653	+0.118	0.729	0.110	1.706	+0.162	0.730
	NCEP	0.183	2.710	+0.866	1.288	0.082	2.770	+1.038	1.278
	FuXi-S2S	0.182	1.624	-0.113	0.806	0.124	1.655	-0.090	0.801
	NeuralGCM	0.197	1.287	-0.146	0.593	0.073	1.393	-0.105	0.637
	DLESyM-v0	0.147	2.003	+0.522	<i>n/e</i>	0.037	2.399	+0.712	<i>n/e</i>
CI	MME (no ECMWF)	0.202	1.284	-0.293	<i>n/e</i>	0.097	1.320	-0.308	<i>n/e</i>
	CMA	0.104	2.030	+0.240	0.778	0.025	2.217	+0.427	0.759
	ECMWF	—	—	—	—	0.108	1.328	-0.466	0.720
	UKMO	0.182	1.407	-0.233	0.624	0.089	1.429	-0.349	0.632
	NCEP	0.109	1.469	-0.637	0.782	0.088	1.423	-0.586	0.758
	FuXi-S2S	0.190	1.372	-0.614	0.741	0.142	1.333	-0.590	0.732
	NeuralGCM	0.167	1.256	-0.256	0.567	0.092	1.319	-0.383	0.630
	DLESyM-v0	0.178	1.500	-0.261	<i>n/e</i>	0.092	1.462	-0.366	<i>n/e</i>
SP	MME (no ECMWF)	0.218	1.799	-0.124	<i>n/e</i>	0.136	1.859	-0.252	<i>n/e</i>
	CMA	0.130	3.716	+1.443	1.399	0.075	3.800	+1.657	1.239
	ECMWF	—	—	—	—	0.216	1.868	-0.312	0.974
	UKMO	0.137	2.105	-0.357	1.009	0.129	2.098	-0.573	1.027
	NCEP	0.201	2.049	-0.544	1.086	0.077	1.944	-0.717	1.077
	FuXi-S2S	0.298	1.828	-0.489	1.041	0.188	1.809	-0.547	1.011
	NeuralGCM	0.279	1.745	-0.343	0.885	0.151	1.909	-0.809	1.058
	DLESyM-v0	0.158	1.997	-0.457	<i>n/e</i>	0.110	1.885	-0.525	<i>n/e</i>
E&NE	MME (no ECMWF)	0.208	4.318	+1.514	<i>n/e</i>	-0.002	4.648	+1.645	<i>n/e</i>
	CMA	0.111	5.895	+1.482	2.605	-0.005	6.140	+1.943	2.431
	ECMWF	—	—	—	—	0.029	3.834	+0.066	1.948
	UKMO	0.150	7.202	+3.358	3.277	0.030	7.054	+3.155	3.020
	NCEP	0.063	5.638	+0.536	2.837	-0.025	5.377	+0.444	2.650
	FuXi-S2S	0.216	4.551	+0.629	2.371	0.071	4.606	+0.918	2.187
	NeuralGCM	0.154	3.496	-0.636	1.816	-0.030	3.726	-0.630	1.969
	DLESyM-v0	0.155	7.316	+3.711	<i>n/e</i>	0.008	7.994	+4.038	<i>n/e</i>

Table 12: JJAS TP-IMERG regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME (no ECMWF)	0.178	3.845	-0.172	<i>n/e</i>	0.023	4.166	-0.337	<i>n/e</i>
	CMA	0.017	5.337	-0.133	2.516	-0.006	5.395	+0.422	2.449
	ECMWF	—	—	—	—	0.020	4.169	+0.434	2.124
	UKMO	0.152	4.579	-0.143	2.172	0.031	4.964	-0.184	2.385
	NCEP	0.047	5.108	-0.772	2.937	0.005	5.107	-1.213	2.928
	FuXi-S2S	0.226	3.927	-0.280	2.147	0.081	4.009	-0.483	2.101
	NeuralGCM	0.167	3.924	-0.411	1.975	0.009	4.227	-0.641	2.201
	DLESyM-v0	0.117	5.011	+0.710	<i>n/e</i>	0.028	5.457	+0.075	<i>n/e</i>
CI	MME (no ECMWF)	0.225	6.176	-0.932	<i>n/e</i>	0.034	6.550	-1.024	<i>n/e</i>
	CMA	0.074	9.596	+1.281	4.356	-0.012	8.780	+0.907	3.991
	ECMWF	—	—	—	—	0.073	6.577	-0.001	3.569
	UKMO	0.138	7.421	-0.776	3.831	0.007	8.046	-0.751	4.186
	NCEP	0.074	7.476	-2.776	4.451	-0.006	7.513	-1.520	4.226
	FuXi-S2S	0.253	6.381	-0.308	3.683	0.133	6.439	-0.798	3.519
	NeuralGCM	0.188	6.394	-0.305	3.401	0.037	7.130	-0.018	3.853
	DLESyM-v0	0.094	7.587	-2.708	<i>n/e</i>	0.008	7.946	-3.968	<i>n/e</i>
SP	MME (no ECMWF)	0.246	5.047	-0.738	<i>n/e</i>	0.097	5.101	-0.529	<i>n/e</i>
	CMA	0.072	7.906	+1.689	3.672	0.030	6.915	+0.795	3.180
	ECMWF	—	—	—	—	0.092	5.496	-0.266	2.807
	UKMO	0.143	6.194	-0.700	3.206	0.025	6.265	-1.247	3.195
	NCEP	0.067	6.284	-1.527	3.510	-0.033	6.267	-0.088	3.187
	FuXi-S2S	0.254	5.406	-0.999	3.145	0.163	5.250	-0.367	2.772
	NeuralGCM	0.212	5.443	+0.275	2.786	0.008	5.951	+0.502	3.063
	DLESyM-v0	0.135	5.881	-3.167	<i>n/e</i>	0.007	6.015	-2.768	<i>n/e</i>
E&NE	MME (no ECMWF)	0.204	6.189	+0.350	<i>n/e</i>	0.030	6.662	+0.037	<i>n/e</i>
	CMA	0.019	7.917	-2.621	4.533	0.009	8.103	-2.029	4.371
	ECMWF	—	—	—	—	0.068	7.345	+1.096	3.807
	UKMO	0.077	8.738	+2.703	4.359	-0.003	8.427	+2.105	4.052
	NCEP	0.124	8.548	-0.711	5.027	0.023	8.512	-2.321	5.063
	FuXi-S2S	0.196	7.722	+1.185	4.383	0.091	7.817	+1.202	4.141
	NeuralGCM	0.153	6.669	-1.358	3.707	0.016	7.268	-1.872	4.105
	DLESyM-v0	0.139	8.134	+2.904	<i>n/e</i>	0.027	9.144	+3.134	<i>n/e</i>

Table 13: OND TP-IMERG regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME (no ECMWF)	0.190	1.099	+0.197	<i>n/e</i>	0.037	1.123	+0.165	<i>n/e</i>
	CMA	0.091	1.578	+0.318	0.515	0.040	1.508	+0.256	0.474
	ECMWF	—	—	—	—	0.058	1.203	+0.245	0.419
	UKMO	0.199	1.225	+0.207	0.345	0.098	1.337	+0.215	0.412
	NCEP	0.134	1.133	+0.027	0.472	0.037	1.114	+0.043	0.429
	FuXi-S2S	0.173	1.145	+0.050	0.444	0.125	1.074	+0.014	0.412
	NeuralGCM	0.125	1.015	+0.049	0.375	0.022	0.991	-0.020	0.380
	DLESyM-v0	0.197	1.596	+0.531	<i>n/e</i>	0.076	1.672	+0.481	<i>n/e</i>
CI	MME (no ECMWF)	0.264	1.640	+0.234	<i>n/e</i>	0.056	1.691	+0.179	<i>n/e</i>
	CMA	0.210	2.363	+0.545	0.845	0.093	2.225	+0.457	0.736
	ECMWF	—	—	—	—	0.022	1.706	+0.232	0.671
	UKMO	0.145	1.884	+0.099	0.690	0.112	1.901	+0.045	0.701
	NCEP	0.114	1.757	-0.087	0.768	0.033	1.732	-0.095	0.705
	FuXi-S2S	0.306	1.607	-0.019	0.738	0.048	1.630	-0.048	0.676
	NeuralGCM	0.148	1.784	+0.389	0.704	0.026	1.890	+0.371	0.745
	DLESyM-v0	0.142	1.977	+0.476	<i>n/e</i>	0.041	2.232	+0.344	<i>n/e</i>
SP	MME (no ECMWF)	0.163	3.767	+0.082	<i>n/e</i>	0.100	3.960	-0.024	<i>n/e</i>
	CMA	0.089	5.376	+1.413	2.344	0.074	5.174	+0.774	2.177
	ECMWF	—	—	—	—	0.112	3.993	-0.194	1.906
	UKMO	0.120	4.166	-0.613	1.904	0.052	4.256	-0.536	1.853
	NCEP	0.135	3.942	-0.354	1.883	0.064	4.032	-1.097	2.004
	FuXi-S2S	0.117	4.090	-0.449	2.078	0.109	3.932	-0.328	1.920
	NeuralGCM	0.140	4.226	+0.388	1.956	0.052	4.537	+0.343	2.044
	DLESyM-v0	0.112	4.291	+0.108	<i>n/e</i>	0.048	4.903	+0.698	<i>n/e</i>
E&NE	MME (no ECMWF)	0.192	1.981	+0.354	<i>n/e</i>	0.070	2.062	+0.077	<i>n/e</i>
	CMA	0.063	2.424	-0.202	1.225	0.025	2.380	-0.201	1.103
	ECMWF	—	—	—	—	0.050	2.136	+0.183	1.019
	UKMO	0.123	2.591	+0.882	1.091	0.061	2.705	+0.751	1.132
	NCEP	0.033	2.253	-0.340	1.212	0.008	2.242	-0.328	1.178
	FuXi-S2S	0.198	2.450	+0.585	1.214	0.128	2.268	+0.182	1.065
	NeuralGCM	0.108	2.050	-0.149	1.011	0.063	2.112	-0.350	1.091
	DLESyM-v0	0.217	2.973	+1.345	<i>n/e</i>	0.076	2.742	+0.406	<i>n/e</i>

Table 14: JF T2M–ERA5 regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME	0.240	2.252	+0.377	<i>n/e</i>	0.025	2.369	+0.025	<i>n/e</i>
	CMA	0.124	3.346	+1.810	2.171	0.005	3.372	+1.842	1.988
	ECMWF	0.213	2.570	-0.655	1.474	0.077	2.745	-0.952	1.534
	UKMO	0.133	3.008	-0.495	1.700	0.002	3.173	-0.445	1.750
	FuXi-S2S	0.189	2.757	+0.103	1.628	0.086	3.084	-0.970	1.757
	DLESyM-v0	0.037	2.837	+0.499	<i>n/e</i>	-0.008	2.949	+0.205	<i>n/e</i>
	DLESyM-v1	0.132	2.511	+0.178	1.454	0.036	2.603	-0.263	1.393
	FCN3	0.092	2.921	+1.324	1.725	0.015	2.997	+0.936	1.652
CI	MME	0.323	1.343	+0.644	<i>n/e</i>	0.056	1.393	+0.374	<i>n/e</i>
	CMA	0.112	3.294	+2.916	2.549	-0.031	3.411	+3.108	2.492
	ECMWF	0.286	1.355	-0.129	0.797	0.225	1.462	-0.230	0.862
	UKMO	0.076	1.799	-0.810	1.045	-0.053	1.923	-1.040	1.094
	FuXi-S2S	0.234	1.441	+0.009	0.912	0.119	1.706	-0.832	1.054
	DLESyM-v0	0.130	1.909	+0.960	<i>n/e</i>	-0.026	1.937	+0.671	<i>n/e</i>
	DLESyM-v1	0.301	1.579	+0.904	0.956	0.096	1.538	+0.532	0.843
	FCN3	0.152	1.657	+0.975	0.939	-0.059	1.760	+0.705	0.996
SP	MME	0.330	0.836	+0.384	<i>n/e</i>	0.201	0.862	+0.262	<i>n/e</i>
	CMA	0.102	2.421	+2.217	1.942	0.088	2.600	+2.424	2.033
	ECMWF	0.356	0.991	-0.325	0.589	0.246	1.042	-0.346	0.618
	UKMO	0.194	1.267	-0.700	0.777	0.083	1.342	-0.755	0.760
	FuXi-S2S	0.211	1.038	-0.093	0.654	0.103	1.138	-0.513	0.684
	DLESyM-v0	0.060	1.203	+0.145	<i>n/e</i>	-0.006	1.320	-0.163	<i>n/e</i>
	DLESyM-v1	0.169	1.083	+0.449	0.624	0.196	1.106	+0.503	0.617
	FCN3	0.209	1.279	+0.758	0.735	-0.047	1.182	+0.261	0.656
E&NE	MME	0.107	2.571	-0.190	<i>n/e</i>	0.042	2.602	-0.387	<i>n/e</i>
	CMA	0.049	3.918	+1.744	2.797	-0.009	4.105	+2.166	2.765
	ECMWF	0.057	2.844	-0.989	1.575	0.016	2.855	-1.043	1.539
	UKMO	0.059	2.548	-0.996	1.438	0.034	2.487	-1.060	1.330
	FuXi-S2S	0.072	2.852	-0.527	1.726	0.006	3.108	-1.377	1.793
	DLESyM-v0	0.012	2.935	-0.355	<i>n/e</i>	0.002	2.985	-0.627	<i>n/e</i>
	DLESyM-v1	0.080	2.967	-0.607	1.625	0.045	3.192	-1.037	1.660
	FCN3	0.057	2.995	+0.235	1.806	0.016	2.998	+0.031	1.752

Table 15: MAM T2M–ERA5 regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME	0.197	2.293	+0.755	<i>n/e</i>	0.060	2.629	+0.734	<i>n/e</i>
	CMA	0.084	3.897	+2.917	2.750	0.007	4.413	+3.554	3.017
	ECMWF	0.222	2.538	+0.060	1.424	0.082	2.952	+0.000	1.706
	UKMO	0.049	2.983	+0.512	1.743	0.022	3.387	+0.906	1.982
	FuXi-S2S	0.167	2.786	+0.356	1.675	0.067	3.128	+0.155	1.834
	DLESyM-v0	0.150	2.745	+0.709	<i>n/e</i>	0.015	3.180	+0.430	<i>n/e</i>
	DLESyM-v1	0.147	2.431	-0.124	1.291	0.005	2.718	-0.547	1.434
	FCN3	0.178	2.713	+0.807	1.462	0.037	2.990	+0.336	1.613
CI	MME	0.376	1.371	+0.686	<i>n/e</i>	0.215	1.572	+0.668	<i>n/e</i>
	CMA	0.135	2.825	+2.062	2.004	0.042	3.037	+2.323	2.029
	ECMWF	0.372	1.424	+0.291	0.855	0.263	1.618	+0.498	0.989
	UKMO	0.245	1.745	+0.038	1.067	0.142	1.927	+0.120	1.150
	FuXi-S2S	0.280	1.623	+0.412	1.057	0.131	1.749	+0.208	1.075
	DLESyM-v0	0.207	1.845	+0.729	<i>n/e</i>	0.010	2.015	+0.736	<i>n/e</i>
	DLESyM-v1	0.249	1.584	+0.381	0.887	0.099	1.592	+0.237	0.863
	FCN3	0.293	1.680	+0.929	0.984	0.049	1.862	+0.620	1.040
SP	MME	0.412	1.029	+0.505	<i>n/e</i>	0.328	1.047	+0.469	<i>n/e</i>
	CMA	0.112	1.905	+0.874	1.315	0.088	1.918	+0.920	1.230
	ECMWF	0.417	1.144	+0.060	0.675	0.364	1.142	+0.349	0.695
	UKMO	0.331	1.346	+0.249	0.884	0.287	1.391	+0.392	0.886
	FuXi-S2S	0.301	1.257	+0.420	0.838	0.181	1.207	+0.193	0.733
	DLESyM-v0	0.234	1.542	+0.294	<i>n/e</i>	0.033	1.497	+0.239	<i>n/e</i>
	DLESyM-v1	0.282	1.376	+0.356	0.830	0.158	1.325	+0.475	0.756
	FCN3	0.304	1.530	+1.071	0.967	0.113	1.361	+0.485	0.754
E&NE	MME	0.084	2.657	-0.068	<i>n/e</i>	-0.022	2.748	-0.061	<i>n/e</i>
	CMA	0.012	3.770	+1.633	2.683	-0.030	3.745	+1.917	2.392
	ECMWF	0.092	2.971	-0.751	1.730	0.020	3.060	-0.629	1.786
	UKMO	0.049	2.775	+0.177	1.698	0.015	2.907	+0.318	1.759
	FuXi-S2S	0.067	2.981	-0.372	1.866	-0.004	3.061	-0.593	1.828
	DLESyM-v0	0.050	3.146	-0.558	<i>n/e</i>	-0.005	3.145	-0.627	<i>n/e</i>
	DLESyM-v1	0.057	3.078	-0.979	1.714	0.034	3.204	-1.249	1.727
	FCN3	0.049	3.092	-0.114	1.820	-0.040	3.178	-0.127	1.878

Table 16: JJAS T2M–ERA5 regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME	0.373	1.851	+0.402	<i>n/e</i>	0.159	1.982	+0.340	<i>n/e</i>
	CMA	0.121	3.014	+1.741	1.904	0.085	2.744	+1.281	1.542
	ECMWF	0.342	2.117	-0.249	1.239	0.124	2.298	-0.134	1.350
	UKMO	0.193	2.254	+0.826	1.366	0.061	2.503	+1.162	1.483
	FuXi-S2S	0.317	2.256	+0.262	1.354	0.122	2.387	+0.059	1.362
	DLESyM-v0	0.271	1.979	+0.315	<i>n/e</i>	0.047	2.412	+0.254	<i>n/e</i>
	DLESyM-v1	0.277	1.941	-0.415	1.005	0.076	2.597	-0.384	1.320
	FCN3	0.320	2.265	+0.245	1.191	0.086	2.309	+0.054	1.207
CI	MME	0.396	0.936	+0.207	<i>n/e</i>	0.120	1.084	+0.085	<i>n/e</i>
	CMA	0.124	1.525	+0.348	0.913	0.033	1.399	-0.194	0.802
	ECMWF	0.374	1.202	-0.480	0.768	0.109	1.340	-0.459	0.859
	UKMO	0.209	1.477	+0.629	0.929	0.032	1.663	+0.574	1.017
	FuXi-S2S	0.324	1.147	+0.152	0.739	0.149	1.238	-0.121	0.740
	DLESyM-v0	0.223	1.217	+0.183	<i>n/e</i>	0.067	1.606	+0.553	<i>n/e</i>
	DLESyM-v1	0.285	0.982	+0.106	0.540	0.040	1.352	+0.720	0.738
	FCN3	0.312	1.237	+0.485	0.688	0.096	1.281	-0.009	0.685
SP	MME	0.421	0.790	+0.107	<i>n/e</i>	0.192	0.881	-0.038	<i>n/e</i>
	CMA	0.161	1.306	-0.558	0.883	0.089	1.355	-0.794	0.901
	ECMWF	0.381	1.144	-0.604	0.738	0.110	1.171	-0.447	0.734
	UKMO	0.257	1.137	+0.288	0.724	0.071	1.288	+0.512	0.755
	FuXi-S2S	0.326	1.092	+0.277	0.725	0.182	1.073	-0.126	0.637
	DLESyM-v0	0.267	1.272	+0.562	<i>n/e</i>	0.093	1.411	+0.448	<i>n/e</i>
	DLESyM-v1	0.327	1.039	+0.272	0.602	0.131	1.207	+0.445	0.685
	FCN3	0.284	1.416	+0.965	0.836	0.085	1.253	+0.183	0.693
E&NE	MME	0.202	1.992	-0.625	<i>n/e</i>	0.103	2.066	-0.715	<i>n/e</i>
	CMA	0.072	2.272	-0.051	1.360	0.071	2.166	-0.344	1.185
	ECMWF	0.182	2.441	-1.338	1.590	0.124	2.437	-1.257	1.575
	UKMO	0.135	2.134	-0.102	1.248	0.060	2.159	-0.017	1.209
	FuXi-S2S	0.144	2.224	-0.467	1.389	0.088	2.362	-0.821	1.389
	DLESyM-v0	0.158	2.301	-1.006	<i>n/e</i>	0.059	2.638	-1.453	<i>n/e</i>
	DLESyM-v1	0.196	2.228	-1.131	1.252	0.069	2.431	-1.073	1.365
	FCN3	0.176	2.392	-0.662	1.368	0.036	2.446	-0.775	1.425

Table 17: OND T2M–ERA5 regional benchmark atlas at Weeks 3 and 6.

IMD mask	Model	Week 3				Week 6			
		ACC	RMSE	Bias	fCRPS	ACC	RMSE	Bias	fCRPS
NW	MME	0.330	2.084	+0.587	<i>n/e</i>	0.111	2.301	+0.656	<i>n/e</i>
	CMA	0.092	3.206	+1.813	1.986	-0.027	3.457	+2.205	2.072
	ECMWF	0.307	2.264	+0.024	1.224	0.118	2.556	+0.198	1.370
	UKMO	0.117	3.206	+0.134	1.861	0.073	3.191	+0.159	1.777
	FuXi-S2S	0.328	2.555	+0.473	1.472	0.079	2.643	+0.426	1.379
	DLESyM-v0	0.193	2.516	+0.480	<i>n/e</i>	0.060	2.736	+0.255	<i>n/e</i>
	DLESyM-v1	0.176	2.397	-0.004	1.193	0.021	2.622	-0.141	1.254
	FCN3	0.164	2.751	+1.084	1.489	0.056	2.875	+1.088	1.510
CI	MME	0.252	1.125	+0.291	<i>n/e</i>	0.037	1.225	+0.276	<i>n/e</i>
	CMA	0.063	2.840	+2.295	2.160	-0.038	3.050	+2.636	2.218
	ECMWF	0.241	1.469	-0.547	0.946	0.064	1.612	-0.640	1.025
	UKMO	0.123	1.705	-0.998	1.057	0.027	1.795	-1.036	1.090
	FuXi-S2S	0.180	1.324	+0.032	0.864	0.077	1.295	-0.364	0.771
	DLESyM-v0	0.083	1.570	+0.479	<i>n/e</i>	0.010	1.612	+0.550	<i>n/e</i>
	DLESyM-v1	0.119	1.217	+0.255	0.701	0.034	1.320	+0.510	0.750
	FCN3	0.186	1.406	+0.705	0.788	0.008	1.543	+0.549	0.875
SP	MME	0.288	0.825	+0.043	<i>n/e</i>	0.113	0.858	-0.044	<i>n/e</i>
	CMA	0.080	1.811	+0.954	1.322	0.046	1.859	+1.144	1.279
	ECMWF	0.292	1.340	-0.954	0.914	0.116	1.469	-1.065	0.979
	UKMO	0.146	1.269	-0.605	0.796	0.057	1.355	-0.630	0.807
	FuXi-S2S	0.214	1.023	-0.036	0.661	0.117	1.038	-0.312	0.608
	DLESyM-v0	0.162	1.065	-0.023	<i>n/e</i>	0.037	1.196	-0.096	<i>n/e</i>
	DLESyM-v1	0.158	0.979	+0.050	0.567	0.064	1.059	+0.242	0.598
	FCN3	0.233	1.278	+0.848	0.732	0.055	1.196	+0.355	0.634
E&NE	MME	0.062	2.222	-0.726	<i>n/e</i>	0.005	2.344	-0.797	<i>n/e</i>
	CMA	0.016	3.096	+0.518	2.093	-0.034	3.294	+0.904	2.078
	ECMWF	0.082	2.709	-1.495	1.618	0.015	2.808	-1.616	1.666
	UKMO	0.040	2.244	-1.317	1.325	0.035	2.268	-1.329	1.273
	FuXi-S2S	0.063	2.530	-0.558	1.491	0.013	2.663	-0.910	1.511
	DLESyM-v0	0.054	2.734	-1.201	<i>n/e</i>	0.018	2.993	-1.498	<i>n/e</i>
	DLESyM-v1	0.029	2.725	-1.183	1.383	-0.012	3.035	-1.427	1.535
	FCN3	0.047	2.680	-0.319	1.474	-0.015	2.679	-0.405	1.459

G Regional–seasonal native fair CRPS

The regional and seasonal fair-CRPS scores are computed from 185,940 case-level rows and agree exactly with the All-India results. Table 18 summarizes 192 cellwise point-estimate minima over four homogeneous regions, four seasons, six leads, and two targets. Each cell reports the eligible native ensemble attaining the minimum unrounded fair-CRPS point estimate and that estimate. The values for every model behind each cell are included in the released result tables.

Table 18: Regional–seasonal native fair-CRPS point-estimate minima. Each cell gives the eligible ensemble attaining the minimum unrounded fair CRPS and its absolute score. NGCM: NeuralGCM; D-v1: DLESyM-v1.

Target	IMD homogeneous mask	Season	W1	W2	W3	W4	W5	W6
TP-IMERG	NW	JF	NGCM 0.359	NGCM 0.417	NGCM 0.436	NGCM 0.467	NGCM 0.463	NGCM 0.459
		MAM	NGCM 0.478	NGCM 0.566	NGCM 0.593	NGCM 0.630	NGCM 0.605	NGCM 0.637
		JJAS	NGCM 1.693	NGCM 1.892	NGCM 1.975	ECMWF 2.089	FuXi 2.100	FuXi 2.101
		OND	NGCM 0.263	NGCM 0.341	UKMO 0.345	UKMO 0.381	UKMO 0.370	NGCM 0.380
	CI	JF	NGCM 0.141	ECMWF 0.175	FuXi 0.176	NCEP 0.187	NCEP 0.186	ECMWF 0.185
		MAM	NGCM 0.392	NGCM 0.503	NGCM 0.567	NGCM 0.578	NGCM 0.605	NGCM 0.630
		JJAS	NGCM 2.553	NGCM 3.116	NGCM 3.401	ECMWF 3.500	ECMWF 3.533	FuXi 3.519
		OND	ECMWF 0.506	NGCM 0.641	UKMO 0.690	ECMWF 0.669	FuXi 0.662	ECMWF 0.671
	SP	JF	FuXi 0.201	ECMWF 0.232	FuXi 0.263	FuXi 0.281	ECMWF 0.286	ECMWF 0.297
		MAM	NGCM 0.656	NGCM 0.790	NGCM 0.885	ECMWF 0.928	ECMWF 0.943	ECMWF 0.974
		JJAS	NGCM 2.259	ECMWF 2.583	NGCM 2.786	ECMWF 2.745	ECMWF 2.796	FuXi 2.772
		OND	ECMWF 1.400	UKMO 1.710	NCEP 1.883	ECMWF 1.862	UKMO 1.835	UKMO 1.853
	E&NE	JF	NGCM 0.320	NGCM 0.361	NGCM 0.425	NGCM 0.432	NGCM 0.445	CMA 0.440
		MAM	NGCM 1.433	NGCM 1.663	NGCM 1.816	ECMWF 1.853	ECMWF 1.912	ECMWF 1.948
		JJAS	NGCM 3.294	NGCM 3.564	NGCM 3.707	ECMWF 3.681	ECMWF 3.750	ECMWF 3.807
		OND	NGCM 0.706	NGCM 0.945	NGCM 1.011	ECMWF 1.004	ECMWF 0.985	ECMWF 1.019
T2M-ERA5	NW	JF	D-v1 1.207	ECMWF 1.408	D-v1 1.454	D-v1 1.512	D-v1 1.408	D-v1 1.393
		MAM	D-v1 0.841	D-v1 1.142	D-v1 1.291	D-v1 1.417	D-v1 1.401	D-v1 1.434
		JJAS	D-v1 0.653	D-v1 0.844	D-v1 1.005	D-v1 1.164	D-v1 1.269	FCN3 1.207
		OND	D-v1 0.900	D-v1 1.098	D-v1 1.193	D-v1 1.263	D-v1 1.320	D-v1 1.254
	CI	JF	FuXi 0.630	ECMWF 0.719	ECMWF 0.797	ECMWF 0.837	D-v1 0.822	D-v1 0.843
		MAM	D-v1 0.532	FCN3 0.700	ECMWF 0.855	ECMWF 0.931	D-v1 0.898	D-v1 0.863
		JJAS	D-v1 0.328	D-v1 0.454	D-v1 0.540	D-v1 0.652	FCN3 0.691	FCN3 0.685
		OND	D-v1 0.531	FCN3 0.639	D-v1 0.701	D-v1 0.703	D-v1 0.786	D-v1 0.750
	SP	JF	D-v1 0.524	ECMWF 0.568	ECMWF 0.589	FCN3 0.592	FCN3 0.577	D-v1 0.617
		MAM	D-v1 0.538	ECMWF 0.602	ECMWF 0.675	ECMWF 0.706	ECMWF 0.698	ECMWF 0.695
		JJAS	D-v1 0.467	D-v1 0.551	D-v1 0.602	FCN3 0.542	FCN3 0.634	FuXi 0.637
		OND	D-v1 0.430	D-v1 0.519	D-v1 0.567	FCN3 0.536	D-v1 0.586	D-v1 0.598
	E&NE	JF	D-v1 1.458	UKMO 1.457	UKMO 1.438	UKMO 1.333	UKMO 1.415	UKMO 1.330
		MAM	D-v1 1.407	UKMO 1.641	UKMO 1.698	UKMO 1.688	D-v1 1.710	D-v1 1.727
		JJAS	D-v1 1.085	D-v1 1.186	UKMO 1.248	UKMO 1.212	CMA 1.180	CMA 1.185
		OND	D-v1 1.214	D-v1 1.318	UKMO 1.325	UKMO 1.287	UKMO 1.287	UKMO 1.273

These unrounded point-estimate minima have no cellwise uncertainty or multiplicity control. In particular, small gaps between the lowest and next-lowest estimates must not be read as 192 independent superiority claims.

H Sensitivity to the rainfall reference: IMERG versus IMD

Figure 4 summarizes the two rainfall references and the regional fair-CRPS minima. Table 19 lists the values.

Both precipitation tracks use the same 517 initialization cohort and fixed six-system MME. They do not have identical observation timing or support: IMERG pairs forecast day 1 with initialization day, whereas end-labelled IMD gauge days pair initialization+1 through +7 for Week 1, with an unavoidable approximately three-hour displacement from UTC forecast-day intervals. IMD also uses its own 1991–2019 climatology. Agreement is therefore a reference sensitivity, not a pooled-observation estimate.

Native-member fair CRPS was not computed against IMD rainfall. We do not copy TP-IMERG fair CRPS into the TP-IMD track; regional fair CRPS elsewhere in this paper means IMD geographic masks under the primary IMERG/ERA5 references.

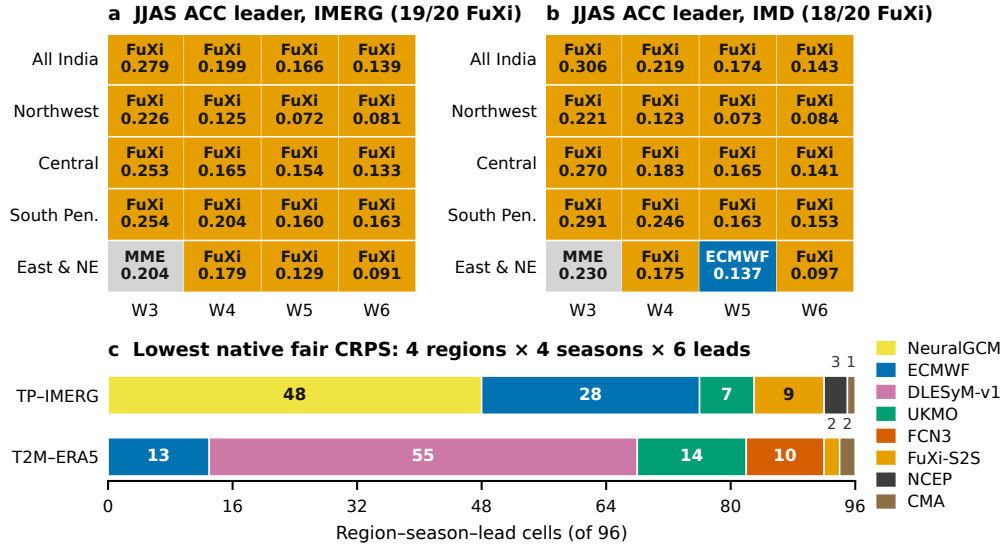


Figure 4: Structured specialization. (a, b) Models attaining the highest JJAS ACC point estimate at Weeks 3–6 across five spatial supports under separate IMERG and IMD verification. (c) Eligible native ensembles attaining the lowest fair-CRPS point estimate over 4 regions × 4 seasons × 6 leads. All selections are descriptive, use eligible and available systems with unrounded scores, and imply neither statistical separation nor equal compute.

Table 19: JJAS Weeks 3–6 highest-ACC point estimates under parallel IMERG and IMD precipitation references. Each entry identifies the model attaining the highest unrounded ACC point estimate for All India or one of four homogeneous-region masks; estimates are never pooled across references.

Region	Week	IMERG highest-ACC model (ACC)	IMD highest-ACC model (ACC)	Same model?
All India	W3	FuXi (0.279)	FuXi (0.306)	yes
	W4	FuXi (0.199)	FuXi (0.219)	yes
	W5	FuXi (0.166)	FuXi (0.174)	yes
	W6	FuXi (0.139)	FuXi (0.143)	yes
Northwest	W3	FuXi (0.226)	FuXi (0.221)	yes
	W4	FuXi (0.125)	FuXi (0.123)	yes
	W5	FuXi (0.072)	FuXi (0.073)	yes
	W6	FuXi (0.081)	FuXi (0.084)	yes
Central	W3	FuXi (0.253)	FuXi (0.270)	yes
	W4	FuXi (0.165)	FuXi (0.183)	yes
	W5	FuXi (0.154)	FuXi (0.165)	yes
	W6	FuXi (0.133)	FuXi (0.141)	yes
South Peninsula	W3	FuXi (0.254)	FuXi (0.291)	yes
	W4	FuXi (0.204)	FuXi (0.246)	yes
	W5	FuXi (0.160)	FuXi (0.163)	yes
	W6	FuXi (0.163)	FuXi (0.153)	yes
East & Northeast	W3	MME (0.204)	MME (0.230)	yes
	W4	FuXi (0.179)	FuXi (0.175)	yes
	W5	FuXi (0.129)	ECMWF (0.137)	no
	W6	FuXi (0.091)	FuXi (0.097)	yes

I Separate comparison with ERPAS over matched 2023–2024 weeks

ERPAS cannot enter the five-year leaderboard because its available record, issue schedule, lead range, and reference differ from the primary cohort. The bounded comparison below instead matches 101 forecast-valid windows from 2023–2024 at Weeks 1–4 and verifies both raw systems against IMD rainfall. ERPAS Wednesday issues are paired with preceding-Monday strict-00Z FuXi-S2S cycles, so FuXi is initialized 48 hours earlier despite identical valid windows. ACC uses a common IMD 1991–2020 calendar-day anomaly baseline; RMSE and bias use raw rainfall. ERPAS is a provider-precomputed mean with variable source count, while FuXi-S2S is uncalibrated.

Table 20: Separate raw-system ERPAS sensitivity, not a primary benchmark ranking. Values are pooled means of case-wise All-India scores over 101 matched forecast-valid windows; RMSE and bias are in mm day^{-1} . No bolding or statistical separation is claimed.

Lead	ERPAS			raw FuXi-S2S		
	ACC $\uparrow$	RMSE $\downarrow$	Bias	ACC $\uparrow$	RMSE $\downarrow$	Bias
W1	0.490	2.934	+0.176	0.514	2.886	-0.104
W2	0.261	3.498	+0.427	0.281	3.521	-0.022
W3	0.132	3.682	+0.638	0.160	3.469	+0.002
W4	0.094	3.744	+0.478	0.133	3.530	+0.181

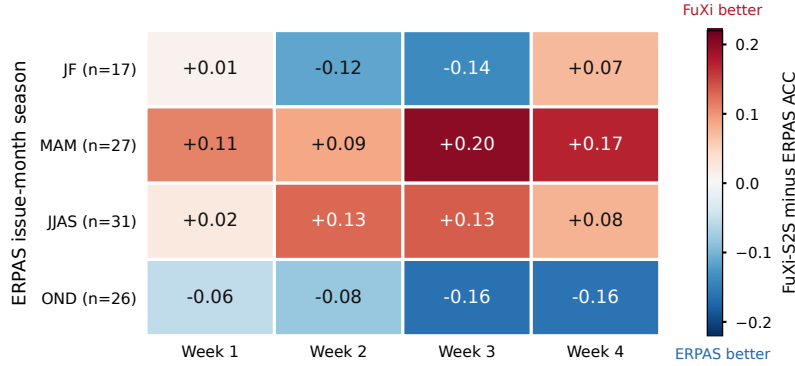


Figure 5: Seasonal heterogeneity in raw FuXi-S2S minus ERPAS mean spatial ACC for the same 101 matched forecast-valid windows. Seasons follow ERPAS issue month; positive values favor FuXi-S2S. FuXi-S2S is higher at every lead in MAM and JJAS, ERPAS is higher throughout OND, and JF is mixed. These two-year point estimates have no cellwise inference or operational-superiority interpretation.

J MME membership and deterministic baseline

TP-MME₆ averages CMA, DLESyM-v0, FuXi-S2S, NCEP, NeuralGCM, and UKMO. T2M-MME₆ uses CMA, DLESyM-v1, ECMWF, FCN3, FuXi-S2S, and UKMO. Membership is fixed over lead. TP-MME7-available adds ECMWF only where the archived field is valid. Week 3 remains missing because the archived ECMWF fields change resolution at forecast day 15, inside the days 15–21 window. It is never filled with a smaller or year-filtered substitute.

Observation climatology is a deterministic baseline with undefined ACC because its spatial anomaly forecast is zero. It is not a prespecified probability distribution and therefore supplies no CRPSS denominator.

Table 21: TP MME availability sensitivity and deterministic observation-climatology baseline. Lower RMSE is better.

Definition / metric	W1	W2	W3	W4	W5	W6
TP-MME6 ACC	0.622	0.399	0.228	0.136	0.098	0.057
TP-MME7-available ACC	0.631	0.415	—	0.153	0.112	0.067
TP-MME6 RMSE	2.429	2.955	3.181	3.282	3.346	3.363
TP-MME7-available RMSE	2.414	2.922	—	3.240	3.302	3.322
Climatology RMSE	3.161	3.153	3.129	3.128	3.136	3.119
MME6 RMSE reduction vs climatology (%)	+23.2	+6.3	-1.7	-4.9	-6.7	-7.8

K Released results

The released result tables give deterministic ACC, RMSE, MAE, and bias and native fair CRPS for every target, model, region, season, and lead. A status flag separates available scores from archive gaps, single-member systems, and the multi-model mean, for which fair CRPS is not defined. A companion table gives the full IMD-reference results. Every released file is listed with its size, row count, and checksum.

L Data availability and release conditions

The released benchmark contains aggregate scores and figures only. It does not redistribute raw IMERG, ERA5, IMD, S2S, or ERPAS data or AI model weights. Upstream data remain subject to their providers’ terms, and publishing aggregate scores does not grant permission to redistribute the underlying forecasts. Region masks and map boundaries follow the Survey of India.

M Interpretation limits

Supported conclusions are specific to this evaluation: deterministic ACC/RMSE/bias, native-member absolute fair CRPS, paired MME contrasts, regional and seasonal fair-CRPS point-estimate minima, and the stated precipitation-reference and climatology sensitivities, plus the separate two-year ERPAS extension. The paper makes no operational-superiority, universal-AI, CRPSS, extreme-event, or demonstrated-decision-value claim. Prospective multi-season validation, reliability diagnostics, common compute and information-age comparisons, and user-centered evaluation remain future work.

N Abbreviations

Table 22: Abbreviations used in the paper.

Term	Meaning	Term	Meaning
S2S	subseasonal to seasonal	NWP	numerical weather prediction
TP	total precipitation	T2M	2-metre air temperature
W1–W6	forecast Weeks 1–6	MME	multi-model (equal-system) mean
ACC	anomaly correlation coefficient	RMSE	root-mean-square error
MAE	mean absolute error	CRPS	continuous ranked probability score
fCRPS	fair CRPS (finite-ensemble corrected)	CRPSS	CRPS skill score
LOYO	leave one year out	n/e	not eligible (single member or MME)
JF	January–February	MAM	March–May
JJAS	June–September (monsoon)	OND	October–December
ERA5	ECMWF Reanalysis v5	IMERG	Integrated Multi-satellitE Retrievals for GPM
IMD	India Meteorological Department	GPM	Global Precipitation Measurement
ECMWF	European Centre for Medium-Range Weather Forecasts	NCEP	US National Centers for Environmental Prediction
UKMO	UK Met Office	CMA	China Meteorological Administration
FCN3	FourCastNet 3	DLESyM	Deep Learning Earth System Model
ERPAS	extended-range prediction product for India (provider mean)	HEALPix	Hierarchical Equal Area isoLatitude Pixelization