Supplemental Material for SeasonBench-EA

October 18, 2025

Contents

A Overview of Ensemble Forecast Systems

This section provides an overview of numerical models included in SeasonBench-EA. Seasonal prediction systems adopt ensemble-based approaches, generating multiple simulations with perturbed initial conditions or varying model configurations. The incorporation of additional components of the Earth system, such as ocean dynamics, land surface processes, is essential for capturing the slowly varying boundary conditions that influence atmospheric variability on seasonal timescales.

Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC) The CMCC-SPS3.5 is based on the atmospheric component CESM1.2-CAM5.3, coupled with land surface model CESM1.2-CLM4.5. The horizontal resolution is approximately 0.5° in latitude–longitude direction. The system is coupled with ocean model NEMOv3.4 and sea ice model CICE4.0, both operating at a horizontal resolution of 0.25°. The operational forecast system uses 50 ensemble members, while the hindcast system includes 40 members covering the years 1993–2016. Model details are available at Description of CMCC-CM2-v20191201 C3S contribution.

Deutscher Wetterdienst (DWD) The DWD GCFS2.1 forecast system is based on the atmospheric component ECHAM6.3.05, coupled with land model JSBACH3.20. The horizontal resolution is approximate 100 km on a regular Gaussian grid. The system is coupled with the ocean model MPIOM 1.6.3 and includes a sea ice component with thermodynamic and dynamic processes, both running on the TP04 tripolar grid. The operational forecast system contains 50 ensemble members, while the hindcast system includes 30 members covering the years 1993-2019. More information is available at Description of GCFS2.1-v20200320 C3S contribution.

Environment and Climate Change Canada (ECCC) The ECCC GEM5-NEMO system is based on the Global Environmental Multiscale Model (GEM), configured with 283 × 190 YY-grid points. The land scheme is ISBA [1]. It is coupled with the NEMO v3.6 ocean model and CICE4.0 sea ice model, both operating at a horizontal resolution of 1°. The operational forecast system includes 10 ensemble members, and the hindcast system also contains 10 members, covering the period from 1990 to 2020. Further details can be found at Description of GEM5-NEMO-v20211130 C3S contribution.

European Centre for Medium-Range Weather Forecasts (ECMWF) The ECMWF-SEAS5 system is based on the atmospheric model IFS Cycle 43r1, configured with a $\mathrm{T_{CO}}319$ cubic octahedral grid for dynamics and an O320 Gaussian grid for physics. The land surface processes are represented by the HTESSEL scheme [2], which includes vegetation, bare soil, snow, and open water. SEAS5 is coupled with the NEMO v3.4 ocean model, the LIM2 sea ice model, and the ECMWF wave model. The ocean and sea ice components operate on the ORCA 0.25° grid, while the wave model runs at 0.5° resolution. The forecast system includes 51 ensemble members, and the hindcast ensemble comprises 25 members spanning the period 1981–2016. Further details are available at Description of SEAS5-v20171101 C3S contribution.

Météo-France The Météo-France System 8 is based on atmospheric component ARPEGE v6.4, coupled with the land surface model SURFEX v8. The horizontal resolution is TL359, corresponding to a reduced Gaussian grid of approximately 0.5°. The system is coupled with the ocean model NEMO v3.6 and sea ice model GELATO v6, both operating on a 0.25° ORCA grid. The forecast ensemble consists of 51 members, while the hindcast ensemble includes 25 members covering the years 1993–2018. More details are available at Description of System8-v20210101 C3S contribution

B Metrics

B.1 Deterministic Metrics

Root Mean Squared Error (RMSE), Eq. 1 is a widely used metric that emphasizes large deviations by squaring the error terms, providing a measure of the overall magnitude of prediction errors. Since SeasonBench-EA focuses on the East Asia region, we do not apply latitude-based weighting to this or any of the following metrics.

$$\text{RMSE} = \sqrt{\frac{1}{H \cdot W} \sum_{j=1}^{H} \sum_{k=1}^{W} \left(\hat{y}^{(j,k)} - y^{(j,k)}\right)^2} \tag{1}$$

Bias, Eq. 2 measures the average difference between predicted and observed values, indicating the systematic error of the model for a certain variable.

$$\text{Bias} = \frac{1}{H \cdot W} \sum_{j=1}^{H} \sum_{k=1}^{W} \left(\hat{y}^{(j,k)} - y^{(j,k)}\right) \tag{2}$$

Willmott's Index of Agreement (WI), Eq. 3 is a standardized metric ranging from 0 to 1 that quantifies model performance by comparing predicted and observed values, accounting for both bias and variability. A value of 1 indicates perfect agreement between model predictions and observations, while a value of 0 indicates no agreement at all.

$$\text{WI} = 1 - \frac{\sum_{j=1}^{H} \sum_{k=1}^{W} (\hat{y}_{j,k} - y_{j,k})^2}{\sum_{j=1}^{H} \sum_{k=1}^{W} \left(|\hat{y}_{j,k} - \bar{y}| + |y_{j,k} - \bar{y}|\right)^2} \tag{3}$$

Anomaly Correlation Coefficient (ACC), Eq. 4 measures the similarity in spatial or temporal patterns between forecasts and observations after removing the climatological signal, and is widely used in weather and climate prediction to assess the skill of anomaly-based forecasts. $\Delta\hat{y}_{j,k} = \hat{y}_{j,k} - \text{clim}_{m,j,k}$, $\Delta y_{j,k} = y_{j,k} - \text{clim}_{m,j,k}$, $\text{clim}_{m,j,k}$ is the observed climatology at each month m.

$$\text{ACC} = \frac{\sum_{j=1}^{H} \sum_{k=1}^{W} \Delta\hat{y}_{j,k} \cdot \Delta y_{j,k}}{\sqrt{\sum_{j=1}^{H} \sum_{k=1}^{W} \Delta\hat{y}_{j,k}^2 \cdot \sum_{j=1}^{H} \sum_{k=1}^{W} \Delta y_{j,k}^2}} \tag{4}$$

Energy Spectrum, Eq. 5 is used to evaluate the scale-dependent performance of a model by quantifying how variance is distributed across spatial scales. This metric is particularly useful for diagnosing whether a model realistically captures the distribution of variability across scales, from planetary waves to synoptic and mesoscale features.

To calculate the zonal energy spectrum, first apply the real-valued Fast Fourier Transform along the longitude direction. Specifically, for an input field $x(i,j) \in \mathbb{R}^{H \times W}$, the Fourier coefficients are given by $\tilde{x}(i,k) = \sum_{j=0}^{W-1} x(i,j) \cdot e^{-2\pi i j k / W}, \quad k = 0, 1, \ldots, \lfloor W/2 \rfloor$. The power spectrum at each latitude i and zonal wavenumber k is computed as $E(i,k) = (\text{Re}[\tilde{x}(i,k)])^2 + (\text{Im}[\tilde{x}(i,k)])^2$. To obtain a representative energy spectrum across the domain, we average across the latitudes:

$E(k) = \frac{1}{H}\sum_{i=1}^{H} E(i,k)$. In summary, the zonal energy spectrum is given by

$$E(k) = \frac{1}{H}\sum_{i=1}^{H} \left|\mathcal{F}_{\text{rfft}}\left[x(i,\cdot)\right]_k\right|^2 \tag{5}$$

Critical Success Index (CSI), Eq. 6 is also known as the Threat Score, evaluates the accuracy of binary event prediction by measuring the fraction of correctly predicted events relative to the total number of predicted or observed events. In this study, we use CSI to assess the performance of total precipitation forecasts by applying thresholds at the 50th, 75th, 90th, 95th, and 99th percentiles of the observed precipitation distribution at each grid.

$$\text{CSI}(\tau) = \frac{\text{TP}\tau}{\text{TP}\tau + \text{FN}\tau + \text{FP}\tau}, \text{for each threshold } \tau \tag{6}$$

B.2 Probabilistic Metrics

Rank Histogram, Eq. 7 is a diagnostic tool used to assess the reliability of ensemble forecasts. It is constructed by ranking the ensemble members and determining the position of the observed value within the ensemble distribution for each forecast case. Systematic deviations from uniformity distribution indicate under ($\cup$-shape) or over ($\cap$-shape) dispersion of the ensemble members. The rank histogram is calculated by `xskillscore.rank_histogram` [3].

Given N ensemble members $\{f_n(i,j)\}_{n=1}^{N}$ and observation $o(i,j)$ at each grid point (i,j), the rank histogram is computed by sorting the ensemble and identifying the rank $r \in \{0,1,\ldots,N\}$ such that

$$f_{(r)}(i,j) \le o(i,j) < f_{(r+1)}(i,j)$$

where $f_{(r)}$ denotes the r-th order statistic ensemble value. The rank histogram counts the frequency of each rank over all grid points:

$$\text{RH}(r) = \sum_{i,j} \mathbf{1}\left[\text{rank}_{i,j} = r\right], \quad r = 0,1,...,N \tag{7}$$

Continuous Ranked Probability Score (CRPS), Eq. 8 evaluates the accuracy of probabilistic forecasts. Lower CRPS values indicate better probabilistic forecast performance, as they imply the forecast distribution is closer to the observed outcome. $F(y)$ is the cumulative distribution function of the ensemble forecasts. The CRPS is calculated by `xskillscore.rank_histogram` [3].

$$\text{CRPS}(F,x) = \int_{-\infty}^{\infty} \left[F(y) - \mathbf{1}(y \ge x)\right]^2 \, dy \tag{8}$$

Spread Skill Ratio (SSR), Eq. 9 is a diagnostic metric used to evaluate the reliability of ensemble forecasts by comparing the ensemble standard deviation to the root mean squared error of the ensemble mean. In ideal cases, a well-calibrated system should satisfy Spread $\approx$ RMSE.

$$\text{SSR} = \frac{1}{H \times W}\sum_{i,j}\sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(f_n(i,j) - \bar{f}(i,j)\right)^2}/\text{RMSE} \tag{9}$$

C Precipitation Patterns in Test Years

In this section, we illustrate the precipitation anomaly patterns for the prediction test set (2020–2024) and the post-processing test set (2013–2016). Precipitation anomaly percentage is used to highlight the interannual variability relative to the climatological baseline. As shown in Figure 1 and Figure 2, distinct precipitation patterns are observed across the Indochina Peninsula, India Peninsula, and southern China, where the spatiotemporal distribution of rainfall is particularly complex.

For instance, over the Indochina Peninsula, 2013, 2016, and 2022 exhibit above-normal precipitation, while significant deficits are observed in 2014 and 2015. In southern China, rainfall is notably above average in 2015, 2016, and 2024, but below normal in 2021 and 2023.

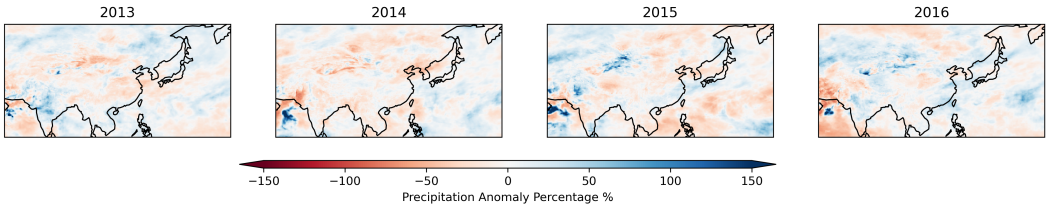


Figure 1: The precipitation anomaly percentage in 2013-2016.

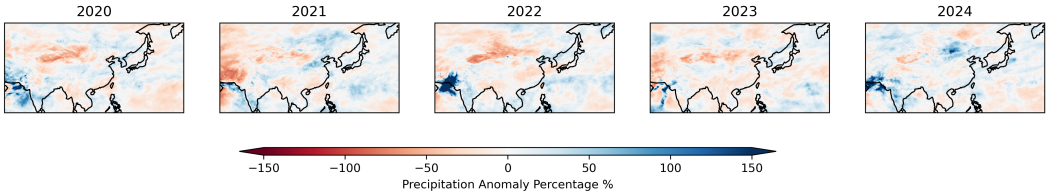


Figure 2: The precipitation anomaly percentage in 2020-2024.

D Additional Evaluation Results for Prediction

This section provides additional visualizations and evaluation results that are omitted in the paper due to space limitations.

D.1 Energy Spectrum Across Lead Times

Figure 3 shows the energy spectrum distribution in prediction task from lead month 1 to month 6.

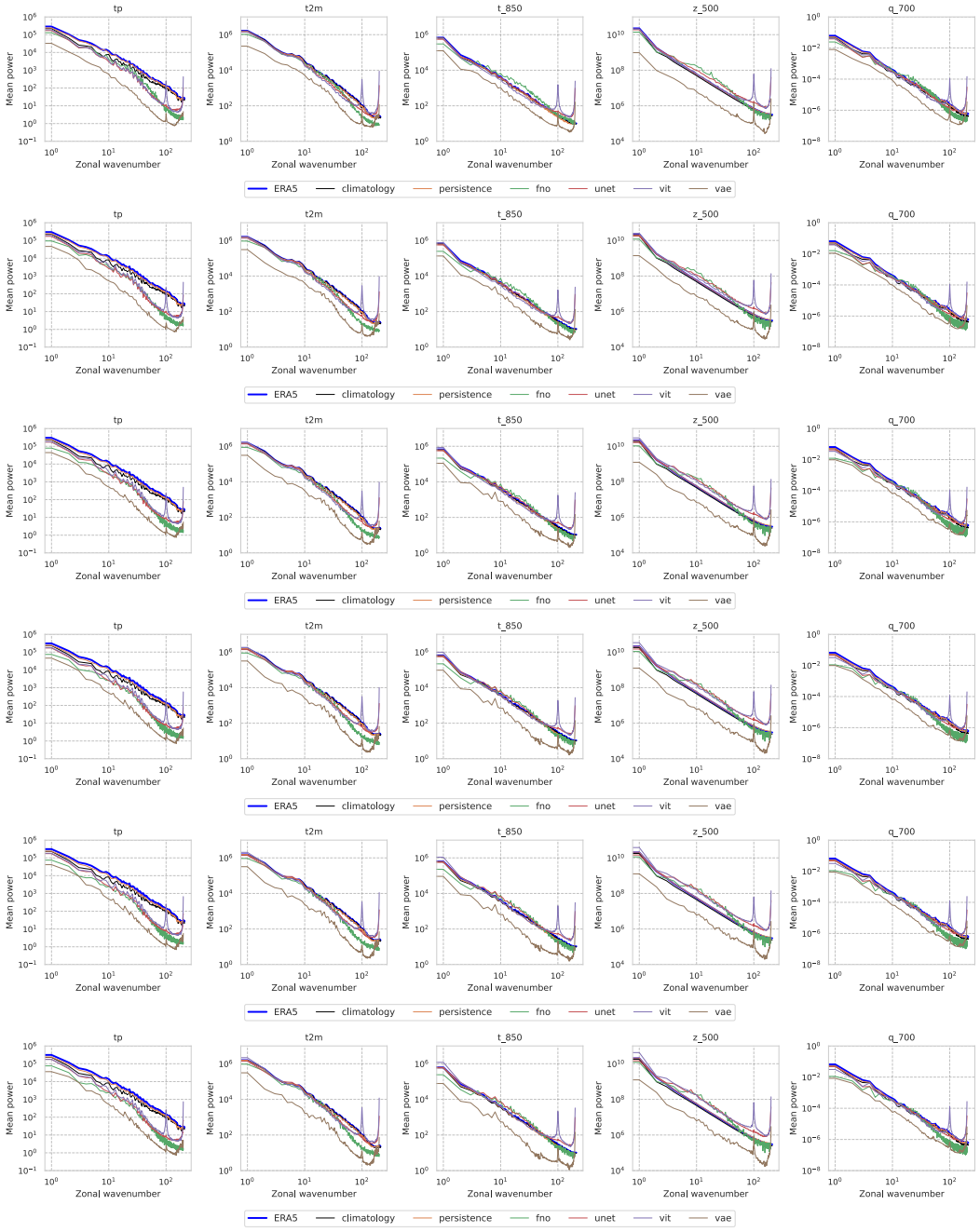


Figure 3: Energy spectrum of seasonal predictions from lead month 1 to 6 (from top to bottom).

D.2 Visualization of Precipitation Predictions

Figures 4 and 5 visualize the precipitation forecasts from different models, initialized in February 2024. Forecast steps 1–6 correspond to March to August 2024, covering the East Asian summer monsoon seasons. Figure 4 shows the precipitation anomalies relative to monthly climatology, while Figure 5 presents the corresponding anomaly percentages.

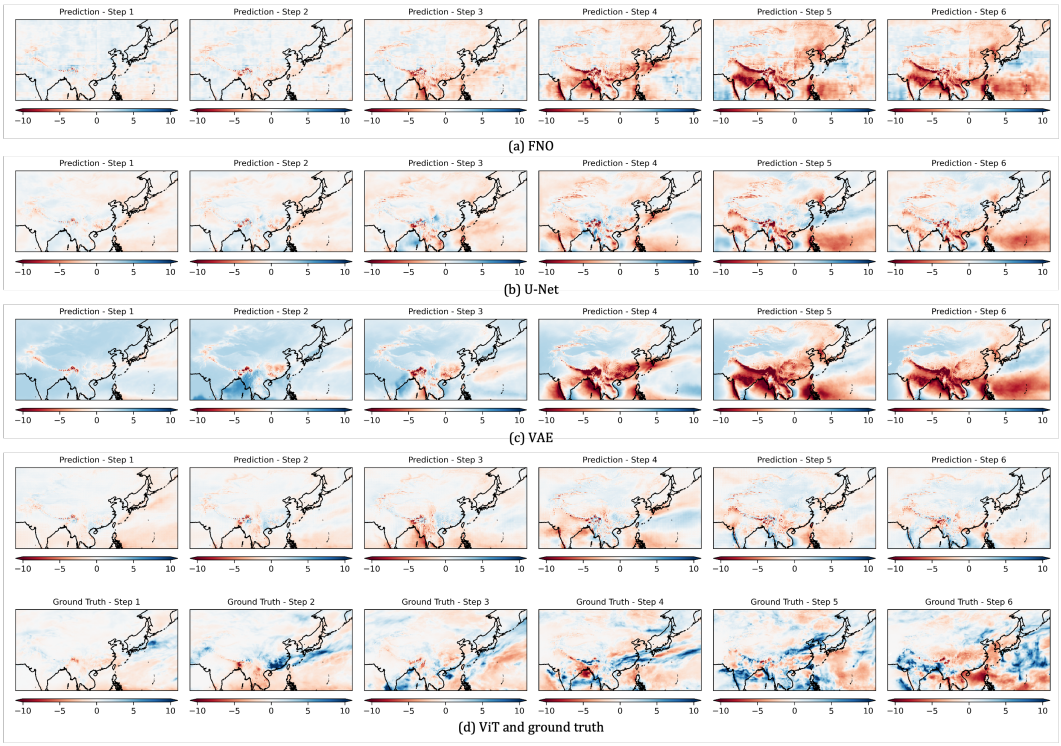


Figure 4: Predicted precipitation anomalies (in mm/day) from different models, initialized in February 2024. Forecast steps 1 to 6 correspond to March to August 2024.

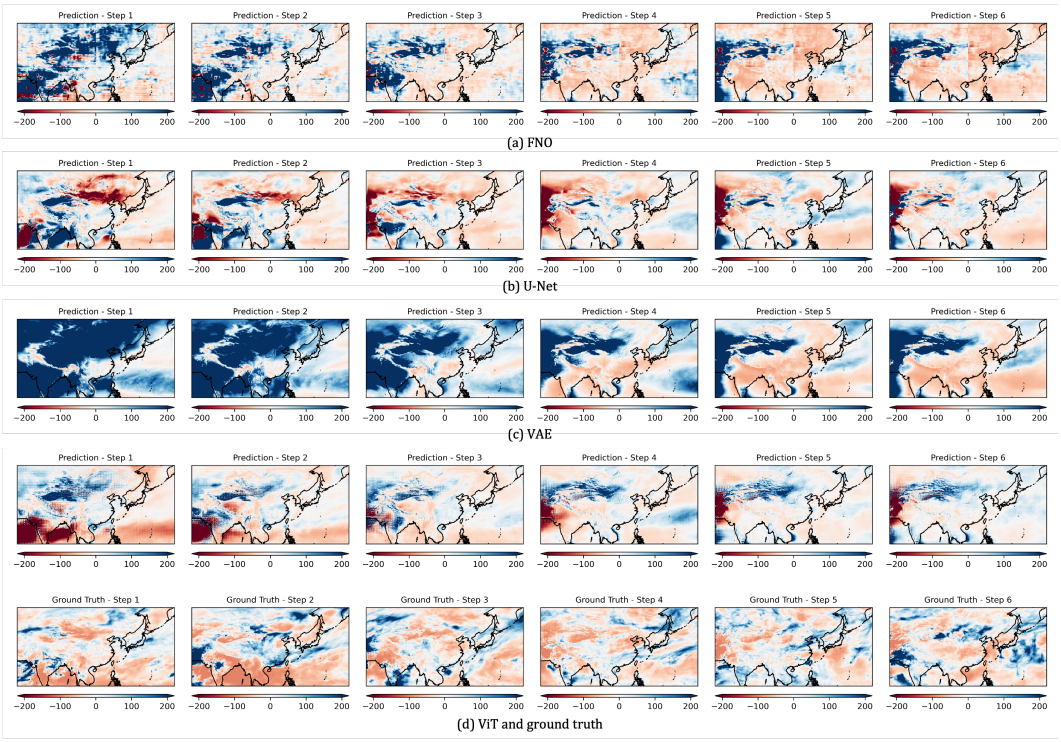
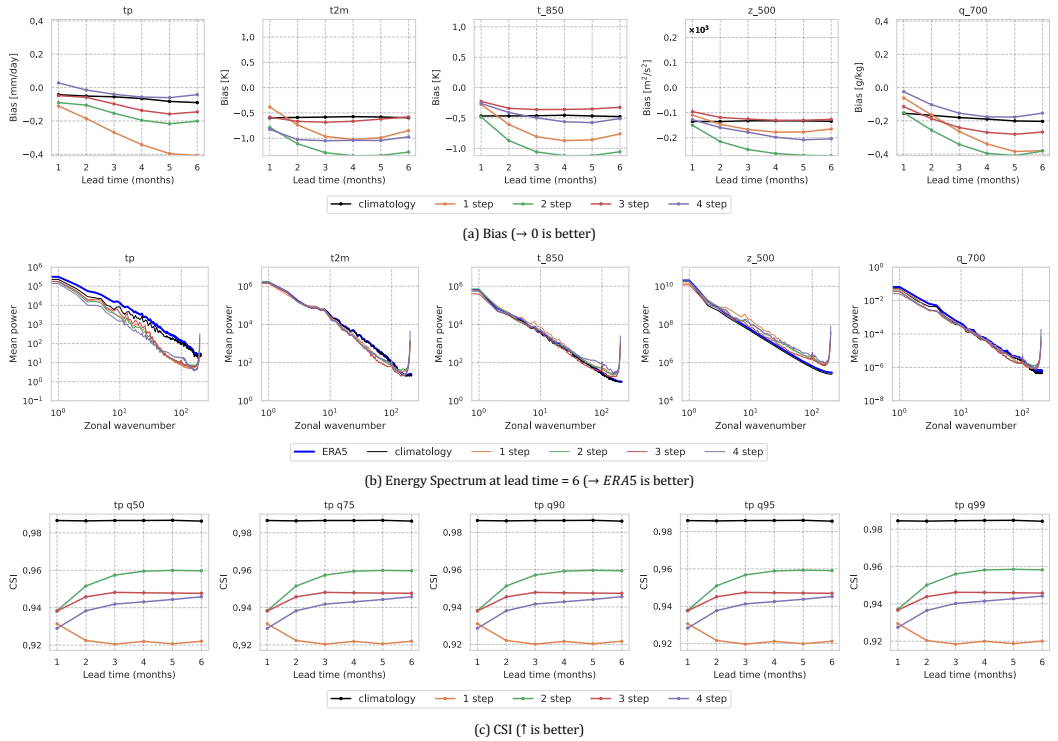


Figure 5: Same as Figure 4, but showing precipitation anomaly percentages relative to the monthly climatology.

D.3 Effects of Autoregressive Steps on Prediction

Figure 4 in the main paper presents RMSE and ACC comparisons of U-Net models trained with different autoregressive steps, highlighting how temporal context length influences model performance. In Figure 6, we further include Bias, Energy Spectrum, and CSI to provide a more comprehensive evaluation. In Figure 7, we visualize the prediction results of U-Net models trained with different autoregressive steps, using 2m temperature initialized in July 2021 as an example. The forecasts span lead months 1 to 6, corresponding to August 2021 to January 2022. Figure 8 shows the precipitation anomalies relative to climatology, initialized in February 2024, while Figure 9 presents the corresponding anomaly percentages.

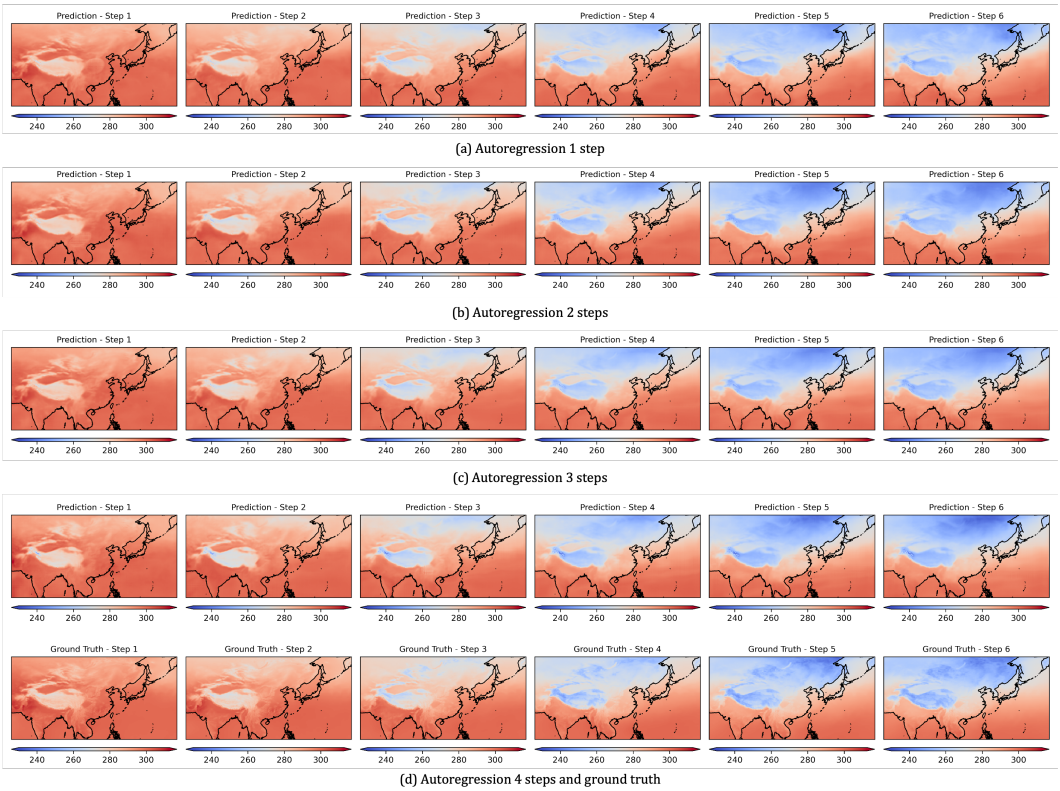


(a) Bias (→ 0 is better)

(b) Energy Spectrum at lead time = 6 (→ *ERA5* is better)

(c) CSI (↑ is better)

Figure 6: Bias, energy spectrum and CSI comparison of U-Net models trained with different autoregressive steps.



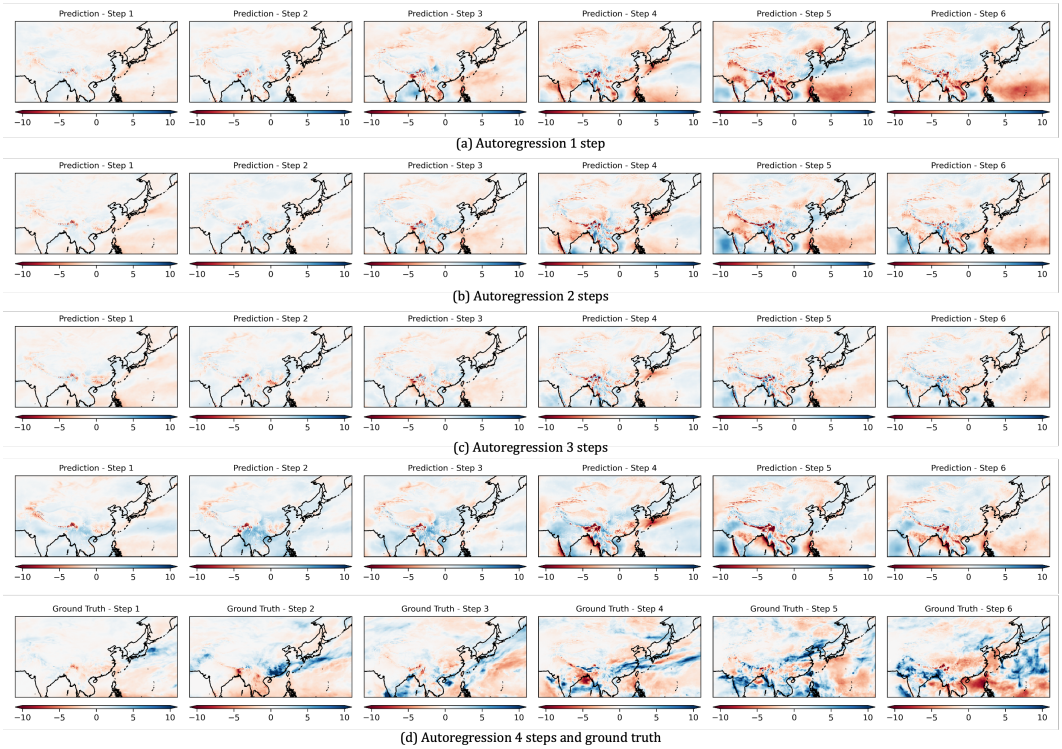
(a) Autoregression 1 step

(b) Autoregression 2 steps

(c) Autoregression 3 steps

(d) Autoregression 4 steps and ground truth

Figure 7: Predictions of 2m temperature from U-Net models with different autoregressive step lengths. The initial month is July 2021, and the predicted months range from August 2021 to January 2022.



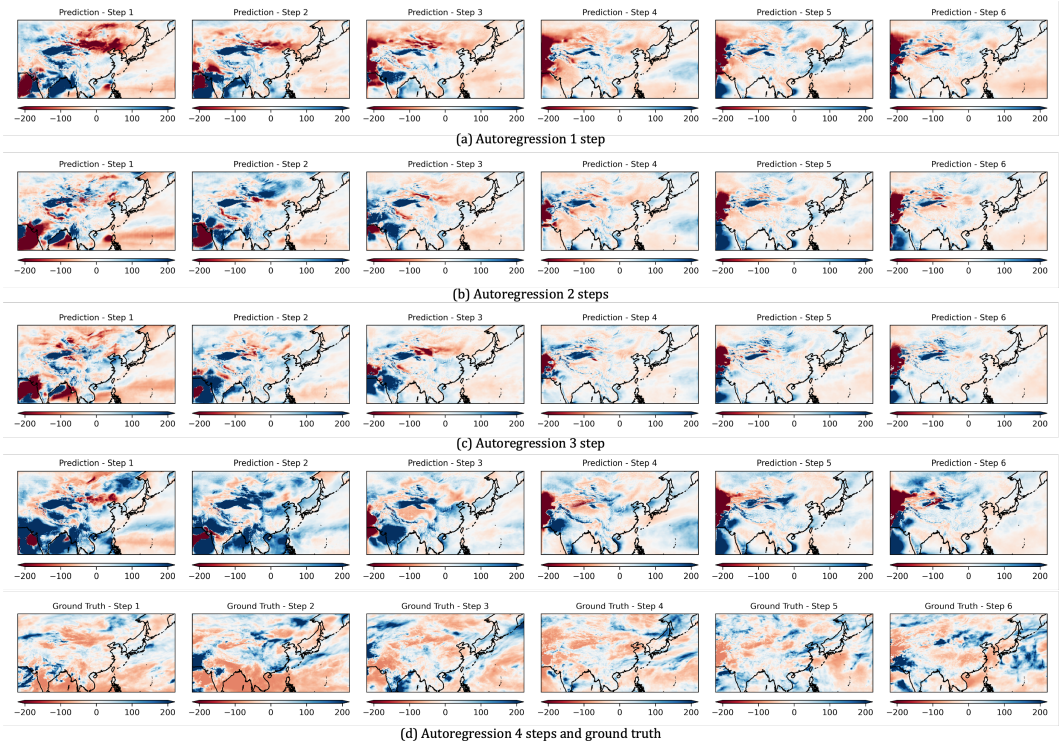
(a) Autoregression 1 step

(b) Autoregression 2 steps

(c) Autoregression 3 steps

(d) Autoregression 4 steps and ground truth

Figure 8: Predicted precipitation anomalies (in mm/day) from U-Net with different autoregressive steps, initialized in February 2024. Forecast steps 1–6 correspond to March–August 2024.



(a) Autoregression 1 step

(b) Autoregression 2 steps

(c) Autoregression 3 step

(d) Autoregression 4 steps and ground truth

Figure 9: Same as Figure 8, but showing precipitation anomaly percentages relative to monthly climatology.

D.4 Impact of shifts in long-term historical data on model performance

To assess whether using long-term historical data may affect model generalization, we examine the impact of climatic shifts over time on prediction skills. Specifically, we test whether training over an extended period (1940-2015) could compromise predictive performance for more recent meteorological variables, given that climatic characteristics in East Asia have evolved. To this end, we train the model with more recent observations (1979-2015), while keeping the same validation and test periods as in the original configuration. The comparative results are presented in Table 1.

Table 1: Impact of training data periods on model performance. Each results is reported as [training from 1940] / [training from 1979], respectively.

variable	Metric	Lead month 1	Lead month 2	Lead month 3	Lead month 4	Lead month 5	Lead month 6
				U-Net			
Z500	RMSE	363.46 / 383.12	392.71 / 420.72	431.39 / 486.44	476.75 / 568.73	505.98 / 637.46	516.84 / 690.30
	ACC	0.10 / 0.17	-0.03 / 0.06	-0.05 / 0.02	-0.08 / 0.01	-0.07 / 0.02	-0.03 / 0.03
t2m	RMSE	2.28 / 2.55	2.74 / 3.04	3.16 / 3.71	3.52 / 4.44	3.76 / 5.05	3.89 / 5.52
	ACC	0.12 / 0.13	-0.00 / 0.03	-0.03 / -0.02	-0.04 / -0.04	-0.01 / -0.03	0.02 / -0.01
tp	RMSE	2.03 / 2.07	2.14 / 2.21	2.25 / 2.41	2.36 / 2.61	2.45 / 2.77	2.48 / 2.88
	ACC	0.12 / 0.12	0.07 / 0.09	0.05 / 0.08	0.04 / 0.07	0.03 / 0.07	0.03 / 0.06
				ViT			
Z500	RMSE	362.20 / 333.27	486.07 / 396.23	636.17 / 488.80	805.31 / 643.31	964.11 / 787.97	1092.61 / 903.55
	ACC	0.07 / 0.28	-0.17 / 0.08	-0.28 / -0.04	-0.34 / -0.15	-0.38 / -0.24	-0.40 / -0.31
t2m	RMSE	2.32 / 2.57	3.08 / 3.28	4.06 / 4.15	5.22 / 5.38	6.34 / 6.55	7.23 / 7.38
	ACC	0.08 / 0.16	-0.11 / 0.03	-0.24 / -0.06	-0.29 / -0.16	-0.32 / -0.23	-0.34 / -0.27
tp	RMSE	2.00 / 2.09	2.09 / 2.28	2.15 / 2.45	2.25 / 2.58	2.32 / 2.66	2.37 / 2.70
	ACC	0.17 / 0.12	0.12 / 0.06	0.11 / 0.05	0.09 / 0.04	0.07 / 0.02	0.05 / 0.01
				FNO			
Z500	RMSE	407.58 / 624.46	476.74 / 801.21	592.03 / 988.54	711.96 / 1115.33	802.12 / 1180.62	842.05 / 1209.36
	ACC	0.10 / 0.12	-0.12 / -0.02	-0.25 / -0.09	-0.29 / -0.12	-0.31 / -0.13	-0.31 / -0.11
t2m	RMSE	3.36 / 5.58	3.90 / 7.00	4.55 / 8.34	5.23 / 9.27	5.70 / 9.75	5.84 / 9.95
	ACC	0.04 / 0.08	-0.12 / -0.02	-0.20 / -0.07	-0.23 / -0.10	-0.25 / -0.12	-0.25 / -0.10
tp	RMSE	2.09 / 2.63	2.19 / 2.86	2.34 / 3.15	2.52 / 3.37	2.67 / 3.48	2.73 / 3.52
	ACC	0.10 / 0.06	0.06 / 0.03	0.04 / 0.03	0.03 / 0.02	0.01 / 0.01	0.00 / 0.01
				VAE			
Z500	RMSE	1319.22 / 1339.89	1318.14 / 1322.90	1342.03 / 1352.56	1341.41 / 1364.74	1340.69 / 1367.13	1365.75 / 1381.68
	ACC	0.18 / 0.21	0.15 / 0.19	0.09 / 0.10	0.06 / 0.06	0.05 / 0.05	0.05 / 0.06
t2m	RMSE	9.67 / 9.81	10.08 / 10.06	10.31 / 10.46	10.31 / 10.50	10.33 / 10.55	10.56 / 10.66
	ACC	0.21 / 0.22	0.12 / 0.14	0.07 / 0.07	0.04 / 0.04	0.04 / 0.04	0.05 / 0.05
tp	RMSE	2.86 / 2.87	3.10 / 3.14	3.29 / 3.31	3.39 / 3.40	3.42 / 3.43	3.43 / 3.43
	ACC	0.05 / 0.05	0.02 / 0.04	0.03 / 0.04	0.03 / 0.03	0.04 / 0.03	0.03 / 0.04

Although the recent period may better represent contemporary climate conditions and benefit from improved observational quality, the substantially reduced training sample size leads to less stable predictions, particularly for models such as U-Net and FNO. In contrast, the ViT model shows improved skill in predicting Z500 during this period.

D.5 Rolling window evaluation

To make the evaluations more rigorous and less prone to data bias and anomalies, we further conduct a rolling window evaluation for U-Net, ViT, FNO, VAE, and a simple linear regression model using three additional data splits beyond the original setting. Specifically, we fix the validation and test period to 4 and 5 years, respectively. The following four configurations are used, with the results summarized in Table 2.

1. Training set: 1940-2015, validation set: 2016-2019, test set: 2020-2024 (in the manuscript)
2. Training set: 1940-2010, validation set: 2011-2014, test set: 2015-2019
3. Training set: 1940-2005, validation set: 2006-2009, test set: 2010-2014
4. Training set: 1940-2000, validation set: 2001-2004, test set: 2005-2009

Table 2: Rolling window evaluation results for U-Net, ViT, FNO, VAE and Linear Regression models on prediction tasks. Each score is reported as *mean* $\pm$ *standard deviation*.

Variable	Metrics	Lead month 1	Lead month 2	Lead month 3	Lead month 4	Lead month 5	Lead month 6
				U-Net			
Z500	RMSE	369.68 ± 7.30	416.95 ± 19.39	483.46 ± 50.00	560.92 ± 77.08	633.63 ± 109.68	690.48 ± 142.08
	ACC	0.05 ± 0.05	-0.03 ± 0.05	-0.03 ± 0.05	-0.03 ± 0.05	-0.03 ± 0.07	-0.04 ± 0.06
t2m	RMSE	2.40 ± 0.15	3.04 ± 0.27	3.83 ± 0.56	4.62 ± 0.86	5.30 ± 1.16	5.83 ± 1.44
	ACC	0.08 ± 0.04	-0.01 ± 0.05	-0.03 ± 0.04	-0.04 ± 0.03	-0.04 ± 0.03	-0.04 ± 0.04
tp	RMSE	2.01 ± 0.04	2.17 ± 0.11	2.32 ± 0.17	2.46 ± 0.20	2.57 ± 0.20	2.62 ± 0.19
	ACC	0.08 ± 0.03	0.03 ± 0.04	0.01 ± 0.03	0.01 ± 0.03	0.01 ± 0.02	0.01 ± 0.02
				ViT			
Z500	RMSE	360.96 ± 20.02	433.71 ± 50.57	516.65 ± 88.63	602.61 ± 144.18	672.39 ± 204.42	719.49 ± 258.28
	ACC	0.01 ± 0.04	-0.08 ± 0.07	-0.12 ± 0.13	-0.14 ± 0.15	-0.14 ± 0.18	-0.14 ± 0.19
t2m	RMSE	2.36 ± 0.04	2.85 ± 0.19	3.40 ± 0.50	3.95 ± 0.93	4.39 ± 1.38	4.70 ± 1.76
	ACC	0.05 ± 0.04	-0.05 ± 0.06	-0.11 ± 0.10	-0.13 ± 0.12	-0.13 ± 0.15	-0.13 ± 0.16
tp	RMSE	1.97 ± 0.02	2.08 ± 0.04	2.16 ± 0.04	2.25 ± 0.05	2.32 ± 0.07	2.35 ± 0.10
	ACC	0.10 ± 0.05	0.06 ± 0.05	0.04 ± 0.05	0.02 ± 0.05	0.02 ± 0.04	0.01 ± 0.03
				FNO			
Z500	RMSE	429.63 ± 18.97	455.84 ± 23.93	515.91 ± 61.40	593.86 ± 93.33	666.35 ± 106.10	710.47 ± 103.36
	ACC	0.03 ± 0.07	-0.04 ± 0.07	-0.09 ± 0.11	-0.11 ± 0.13	-0.12 ± 0.14	-0.13 ± 0.15
t2m	RMSE	3.39 ± 0.07	3.65 ± 0.19	4.04 ± 0.35	4.51 ± 0.49	4.93 ± 0.52	5.14 ± 0.49
	ACC	0.02 ± 0.05	-0.06 ± 0.05	-0.11 ± 0.07	-0.13 ± 0.08	-0.15 ± 0.07	-0.15 ± 0.07
tp	RMSE	2.06 ± 0.03	2.12 ± 0.07	2.22 ± 0.12	2.34 ± 0.16	2.45 ± 0.18	2.50 ± 0.17
	ACC	0.06 ± 0.02	0.02 ± 0.03	0.01 ± 0.03	-0.00 ± 0.03	-0.01 ± 0.02	-0.01 ± 0.01
				VAE			
Z500	RMSE	1349.51 ± 22.71	1345.48 ± 24.75	1345.51 ± 11.99	1342.81 ± 15.36	1347.89 ± 19.90	1374.90 ± 20.72
	ACC	0.06 ± 0.08	0.05 ± 0.07	0.04 ± 0.04	0.03 ± 0.03	0.02 ± 0.03	0.02 ± 0.03
t2m	RMSE	9.88 ± 0.20	10.08 ± 0.23	10.21 ± 0.25	10.28 ± 0.28	10.37 ± 0.28	10.61 ± 0.24
	ACC	0.10 ± 0.11	0.05 ± 0.09	0.03 ± 0.08	0.02 ± 0.08	0.02 ± 0.07	0.02 ± 0.07
tp	RMSE	2.85 ± 0.02	3.10 ± 0.07	3.27 ± 0.04	3.36 ± 0.04	3.39 ± 0.04	3.41 ± 0.03
	ACC	0.02 ± 0.03	0.00 ± 0.02	0.01 ± 0.02	0.01 ± 0.02	0.01 ± 0.02	0.01 ± 0.01
				Linear Regression			
Z500	RMSE	514.47 ± 7.17	824.53 ± 18.23	1124.37 ± 22.32	1362.87 ± 25.04	1527.78 ± 24.97	1626.61 ± 22.64
	ACC	0.05 ± 0.03	-0.02 ± 0.08	-0.04 ± 0.10	-0.05 ± 0.10	-0.05 ± 0.10	-0.05 ± 0.09
t2m	RMSE	3.49 ± 0.04	6.16 ± 0.09	8.45 ± 0.14	10.12 ± 0.17	11.17 ± 0.16	11.70 ± 0.11
	ACC	0.03 ± 0.06	-0.05 ± 0.07	-0.06 ± 0.09	-0.07 ± 0.09	-0.06 ± 0.08	-0.06 ± 0.07
tp	RMSE	2.39 ± 0.04	2.82 ± 0.04	3.11 ± 0.03	3.31 ± 0.03	3.41 ± 0.03	3.44 ± 0.03
	ACC	0.07 ± 0.04	0.04 ± 0.05	0.03 ± 0.03	0.02 ± 0.02	0.02 ± 0.02	0.01 ± 0.02

D.6 Training with different seeds

To access the robustness of model training, we conduct experiments using different random seeds to quantify the mean and standard deviation across different runs. For the prediction task, we perform multi-seed evaluation with the U-Net model, training it under three additional random seeds (four runs in total). The results are summarized in Table 3.

Table 3: Performance of U-Net model trained with different random seeds. Each value represents the *mean* $\pm$ *standard deviation* across different runs.

Variable	Metric	Lead month 1	Lead month 2	Lead month 3	Lead month 4	Lead month 5	Lead month 6
Z500	RMSE	353.83 ± 13.63	384.93 ± 20.16	422.42 ± 33.43	468.12 ± 46.01	503.05 ± 66.96	530.21 ± 81.58
	ACC	0.13 ± 0.05	-0.03 ± 0.05	-0.08 ± 0.05	-0.12 ± 0.05	-0.13 ± 0.07	-0.13 ± 0.09
t2m	RMSE	2.27 ± 0.09	2.66 ± 0.18	3.00 ± 0.28	3.29 ± 0.33	3.52 ± 0.36	3.69 ± 0.37
	ACC	0.15 ± 0.03	0.04 ± 0.05	-0.01 ± 0.06	-0.03 ± 0.04	-0.03 ± 0.05	-0.02 ± 0.06
tp	RMSE	2.01 ± 0.01	2.13 ± 0.04	2.24 ± 0.06	2.34 ± 0.09	2.40 ± 0.08	2.44 ± 0.06
	ACC	0.14 ± 0.02	0.10 ± 0.02	0.09 ± 0.03	0.07 ± 0.03	0.07 ± 0.03	0.06 ± 0.02

E Additional Evaluation Results for Post-processing

This section provides additional visualizations and evaluation results that are omitted in the paper due to space limitations.

E.1 More evaluation metrics on CMCC

Despite the metrics shown in the main paper, the deterministic metrics is shown in Figure 10. The energy spectrum is shown in Figure 11. The rank histograms of total precipitation, 2m temperature and geopotential at 500 hPa are shown in Figure 12 $\sim$ 14.

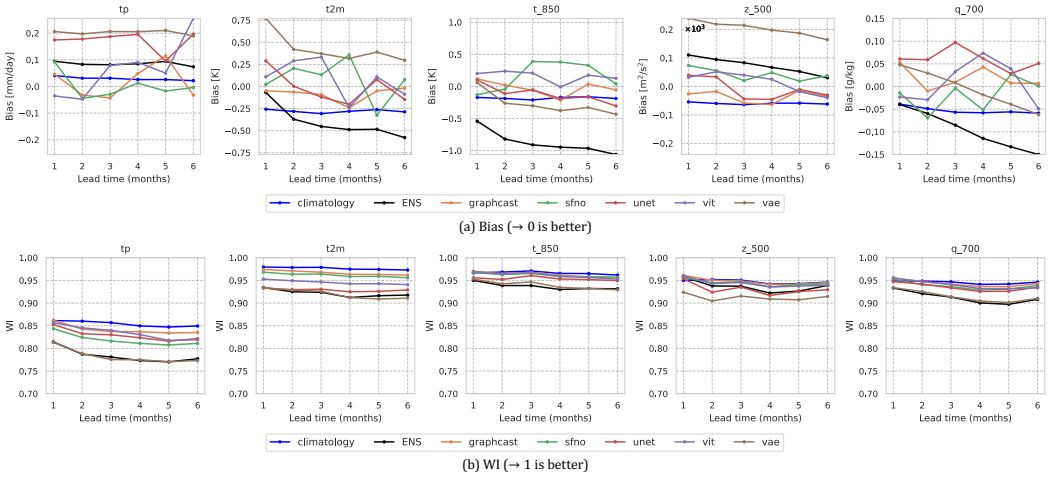


Figure 10: Bias and WI comparison between different models for post-processing.

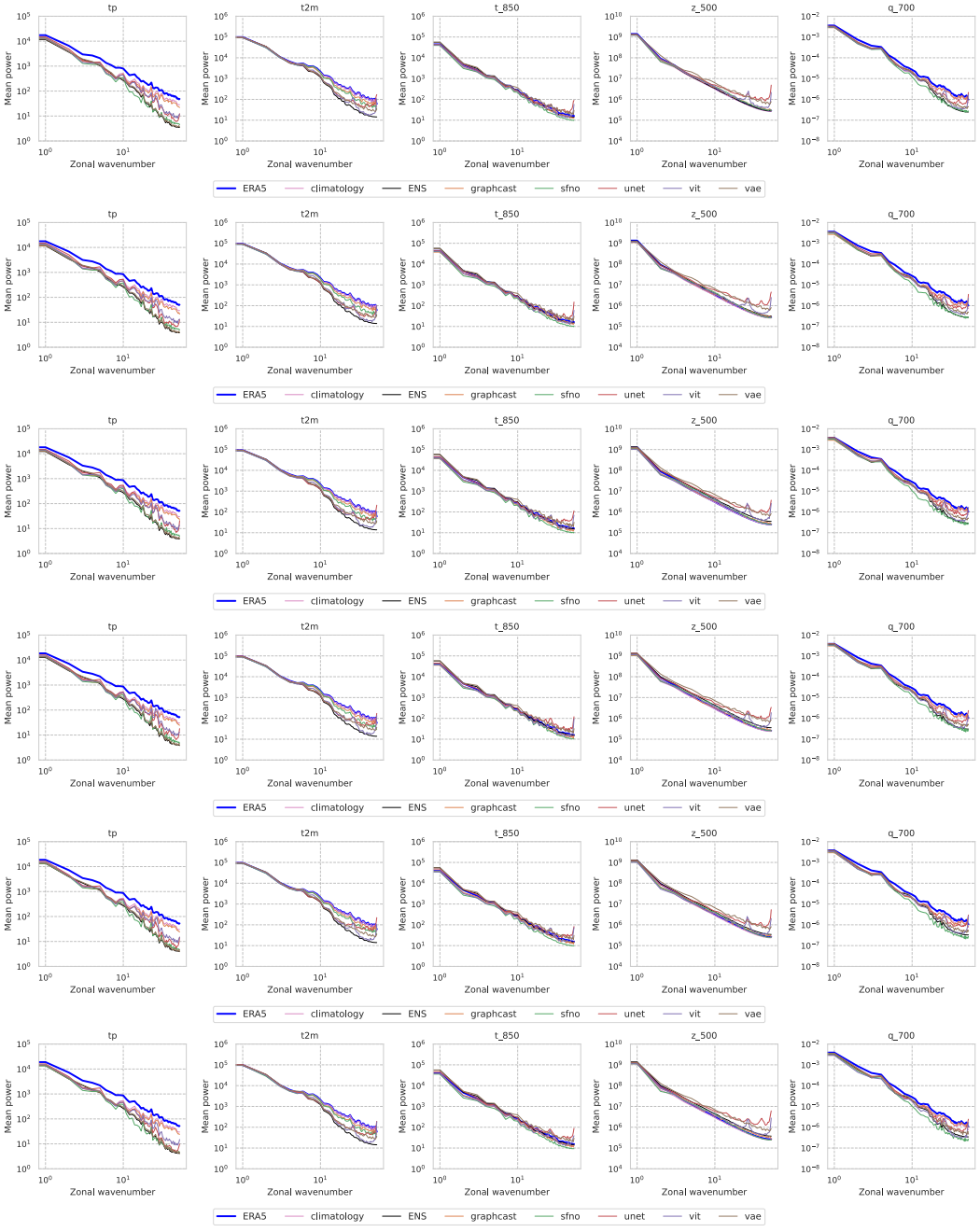


Figure 11: Energy spectra of post-processed results from lead month 1 to 6 (top to bottom) for different models.

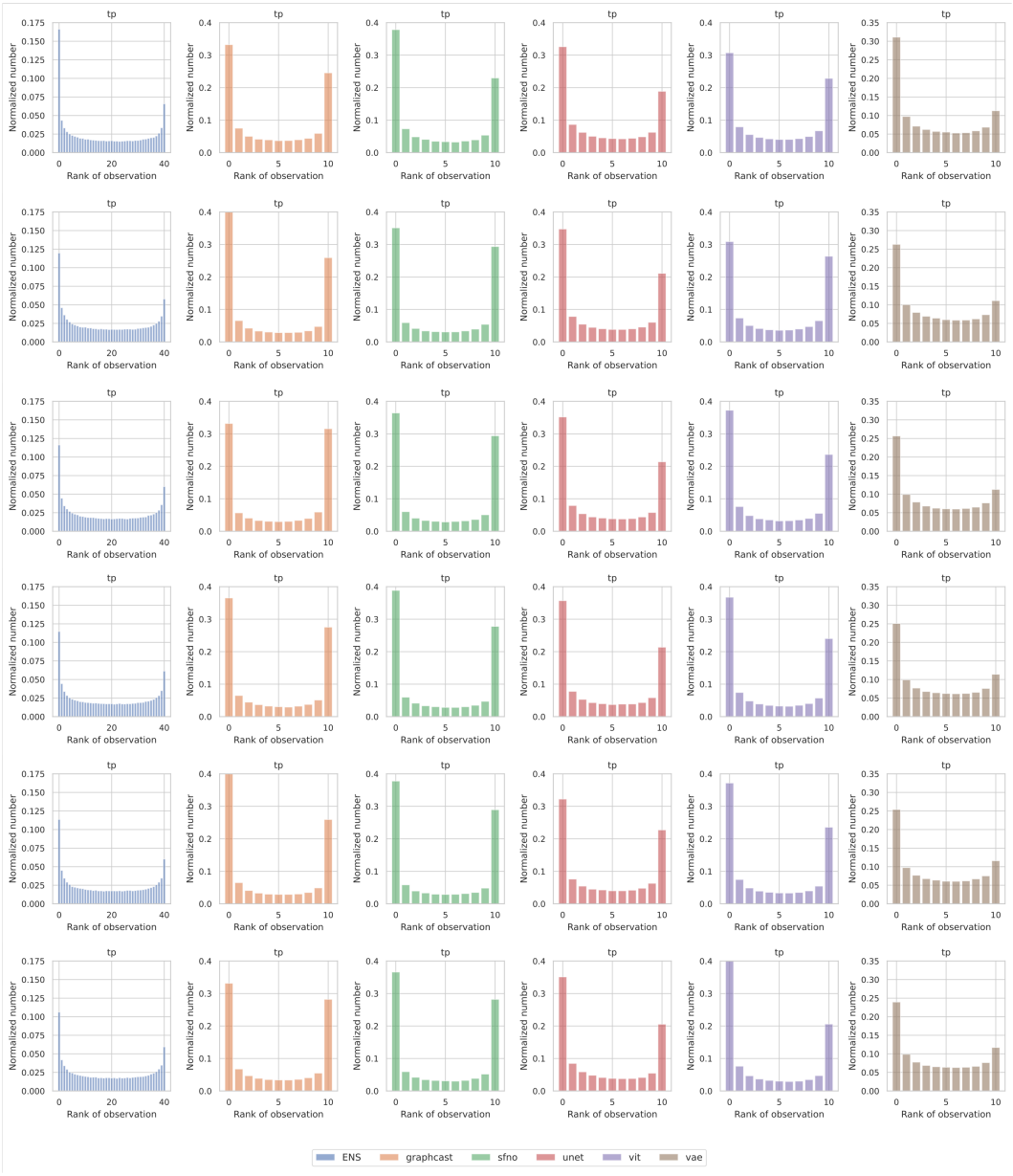


Figure 12: Rank histograms of post-processed total precipitation from lead month 1 to lead month 6 (top to bottom) for different models.

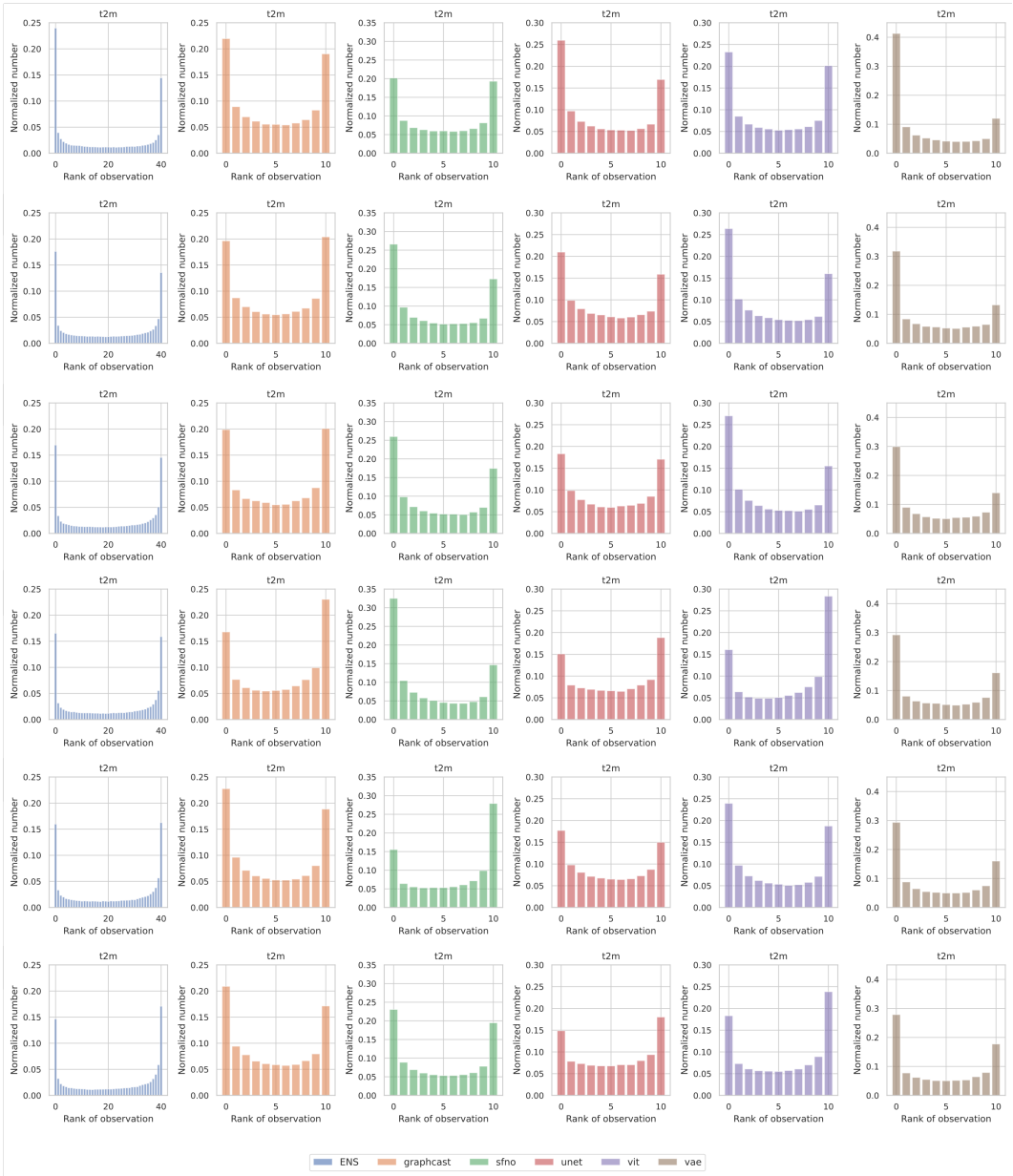


Figure 13: Same as Figure 12, but for 2 m temperature.

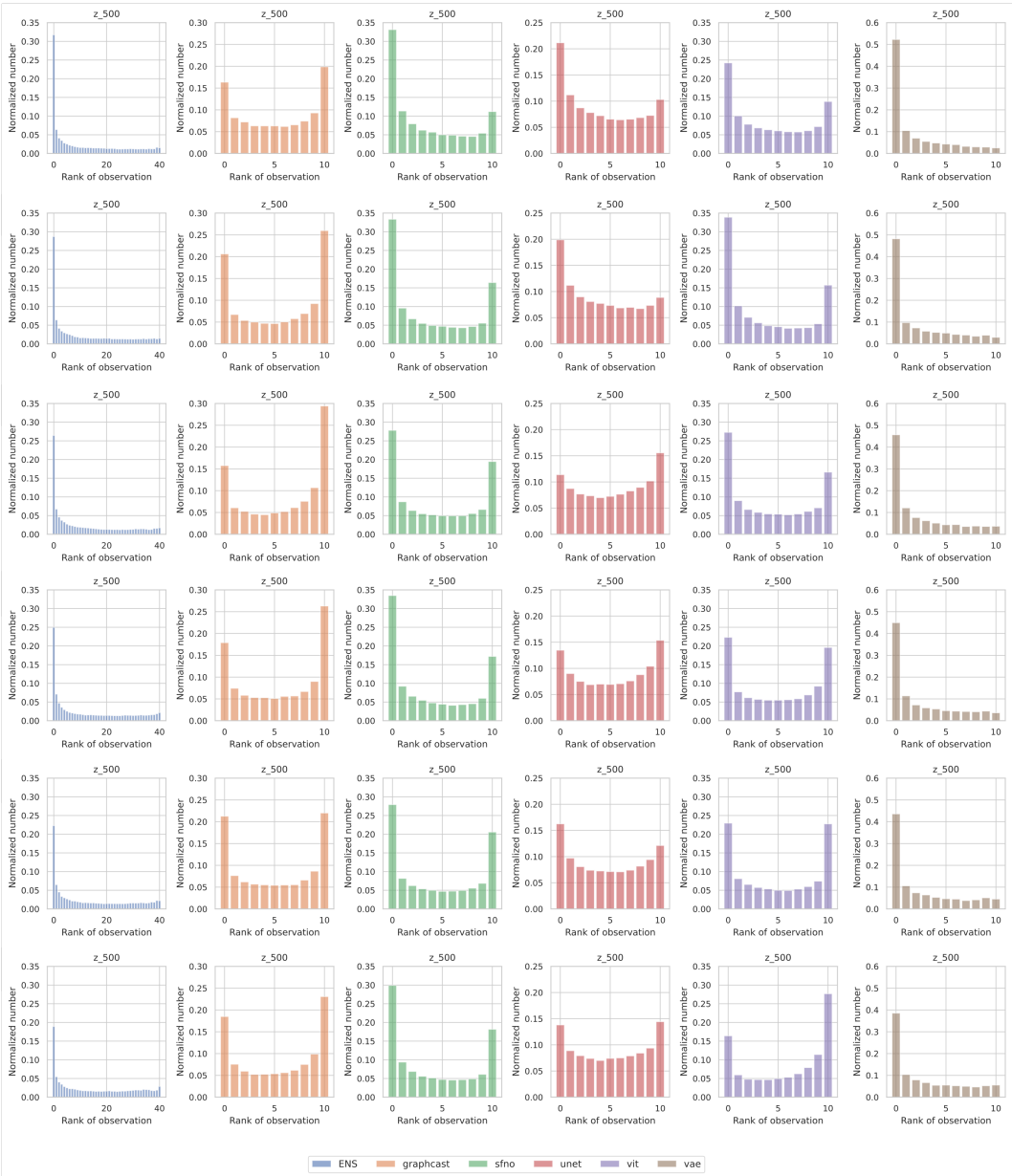
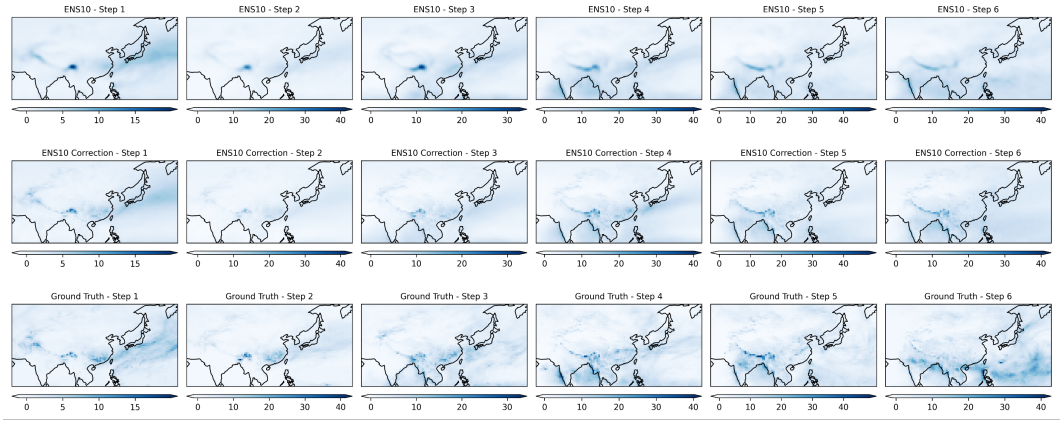


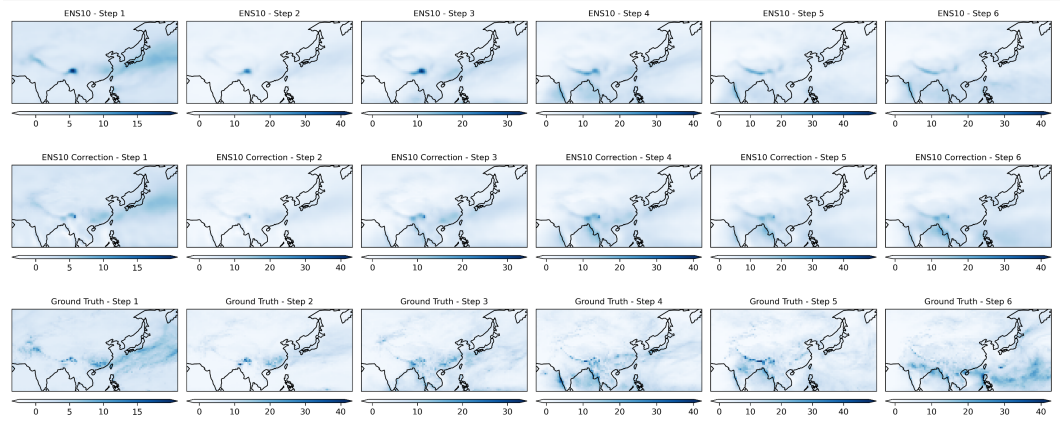
Figure 14: Same as Figure 12, but for 500 hPa geopotential.

E.2 Post-processed Precipitation Results Based on CMCC Forecasts

Figure 15 and 16 present the post-processed precipitation results based on ensemble forecasts initialized in March 2016. Steps 1 to 6 correspond to March to August 2016, covering the East Asian summer monsoon season. For each model, the top row displays the ensemble mean of 10 CMCC members, the middle row shows the precipitation fields post-processed by each model, and the bottom row presents the ERA5 ground truth.

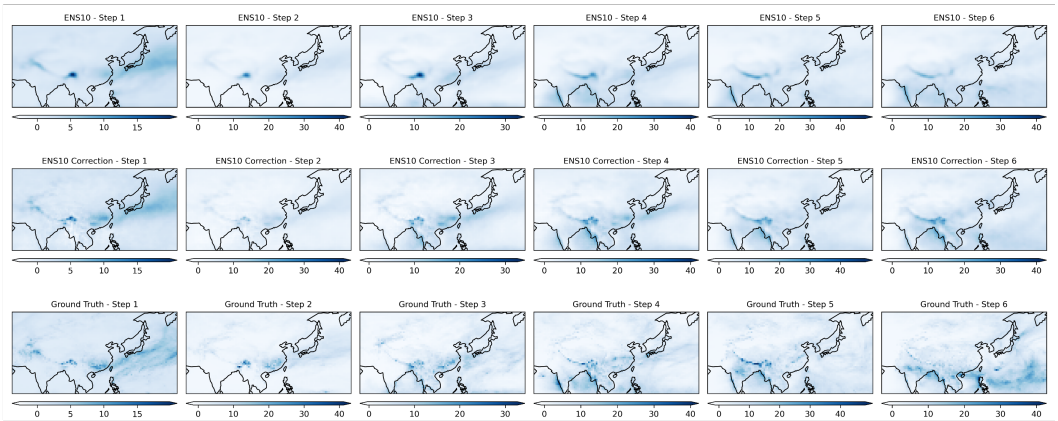


(a) GraphCast

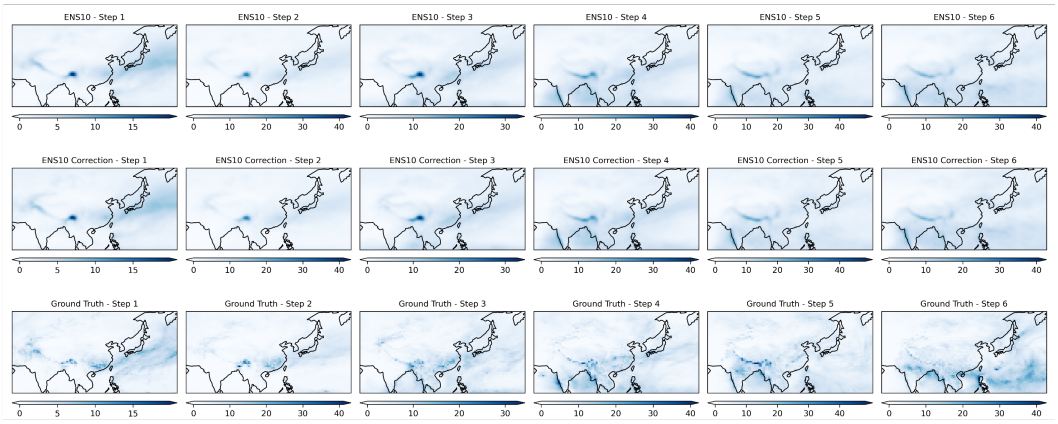


(b) SFNO

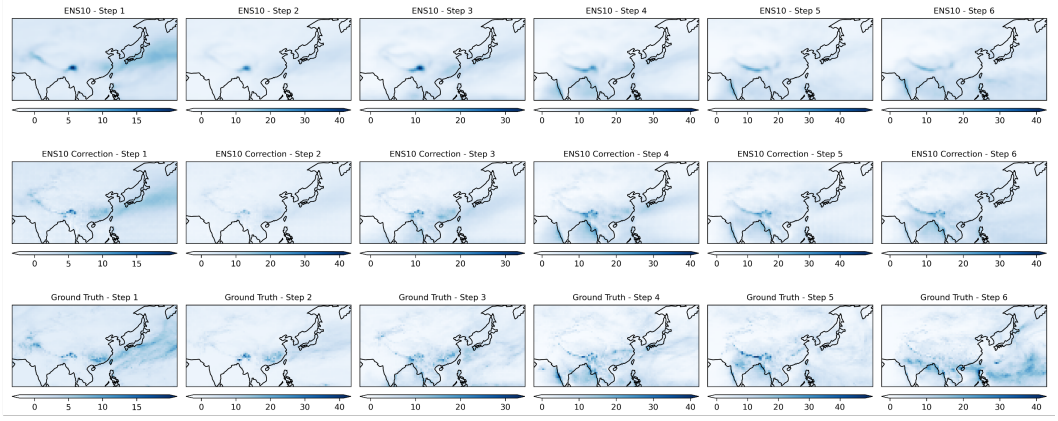
Figure 15: Post-processed precipitation (in mm/day) from GraphCast and SFNO models based on 10-member CMCC ensemble forecasts initialized in March 2016. Lead steps 1 to 6 correspond to March to August 2016.



(c) U-Net



(d) VAE



(e) ViT

Figure 16: Same as Figure 15, but from U-Net, VAE and ViT.

E.3 Training with different seeds

To access robustness of model training, we conduct experiments using different random seeds to quantify the mean and standard deviation across different runs. For the post-processing task, we perform multi-seed evaluation with GraphCast model, training it under three additional random seeds (four runs in total). The results are summarized in Table 4.

Table 4: Performance of GraphCast model trained with different random seeds. Each value represents the *mean* $\pm$ *standard deviation* across runs.

Variable	Metric	Lead month 1	Lead month 2	Lead month 3	Lead month 4	Lead month 5	Lead month 6
Z500	RMSE	256.05 ± 4.54	298.32 ± 3.63	296.98 ± 2.26	300.60 ± 5.88	304.01 ± 3.56	310.78 ± 4.80
	ACC	0.40 ± 0.02	0.13 ± 0.02	0.13 ± 0.01	0.15 ± 0.04	0.13 ± 0.01	0.10 ± 0.02
t2m	RMSE	1.46 ± 0.04	1.62 ± 0.03	1.66 ± 0.03	1.66 ± 0.02	1.69 ± 0.04	1.73 ± 0.02
	ACC	0.33 ± 0.01	0.10 ± 0.03	0.03 ± 0.04	0.04 ± 0.02	0.02 ± 0.03	0.04 ± 0.03
tp	RMSE	1.68 ± 0.01	1.80 ± 0.01	1.85 ± 0.01	1.87 ± 0.01	1.88 ± 0.01	1.91 ± 0.02
	ACC	0.25 ± 0.00	0.12 ± 0.01	0.10 ± 0.00	0.10 ± 0.01	0.10 ± 0.01	0.08 ± 0.02

E.4 GraphCast-Based Post-Processing on CMCC and ECMWF Ensembles

We compare the post-processing performance of GraphCast applied to CMCC and ECMWF seasonal forecasts with selected evaluation metrics, as shown in Figure 17.

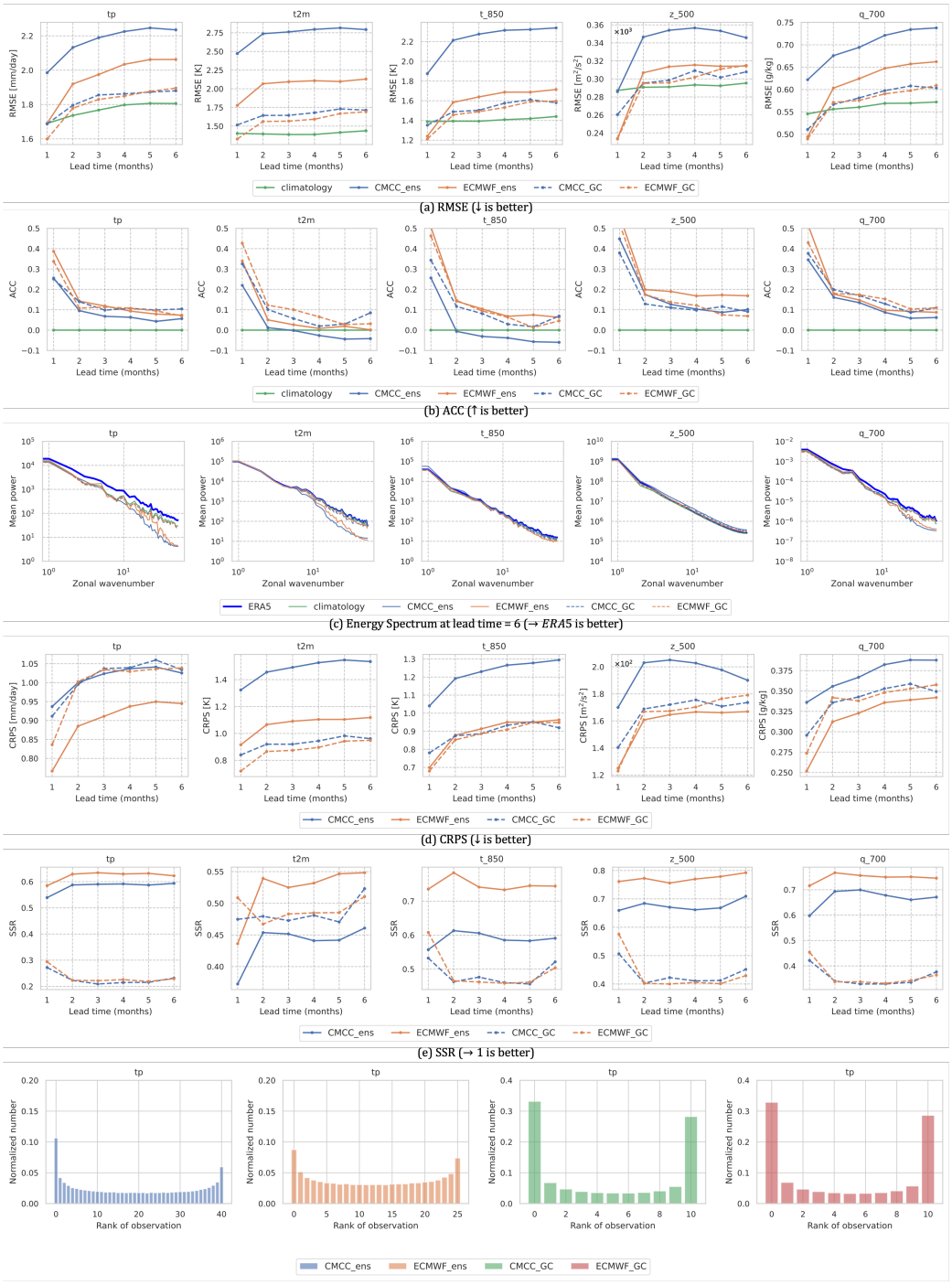


Figure 17: Comparison of post-processing results for seasonal forecasts from CMCC and ECMWF. "CMCC_ens" and "ECMWF_ens" refer to the raw ensemble means from 40-member and 25-member systems, respectively. "CMCC_GC" and "ECMWF_GC" refer to the GraphCast-processed forecasts using the first 10 ensemble members.

Figure 18 presents the post-processed precipitation results from the GraphCast model, applied to 10-member ensemble forecasts from both CMCC and ECMWF initialized in March 2016.

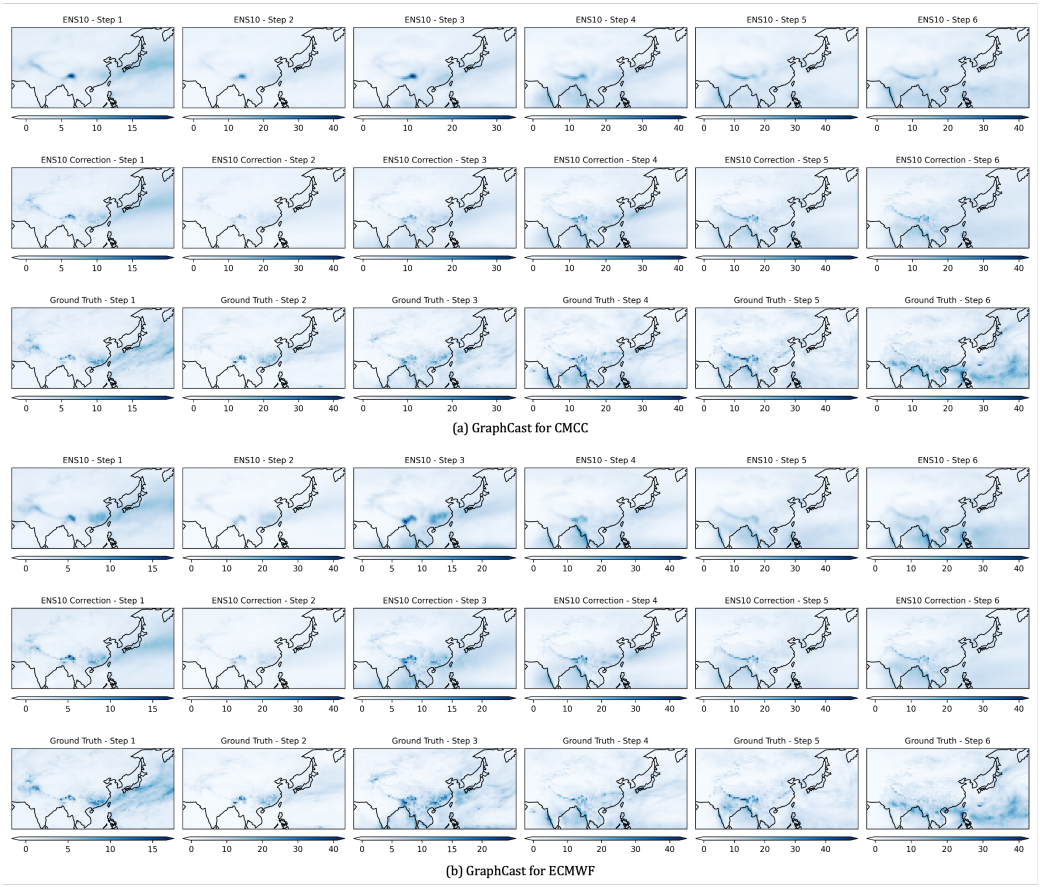


Figure 18: Post-processed precipitation (in mm/day) from GraphCast model based on 10-member CMCC and 10-member ECMWF ensemble forecasts initialized in March 2016. Lead steps 1 to 6 correspond to March to August 2016.

F Model Configurations

The configurations of all baseline models are provided in Table 5 $\sim$ 10. The number of parameters for each model is computed using the function `sum(p.numel() for p in model.parameters())`, and the GPU memory usage is measured with `torch.cuda.memory_allocated(device)`.

Table 5: Model hyperparameters, parameter count, and memory usage for FNO.

Hyperparameters	**Values**
Non-Spectral and Spectral Channel	[64, 128, 256, 512, 1024]
Activation	GELU
Fourier Modes	(16, 16)
Optimizer	AdamW
Learning Rate	CosineAnnealing (1e-3 $\rightarrow$ 1e-4)
Batch Size	16
Crop Size	(200, 400)
Parameters	714.4 M
Memory (batch size = 1)	4559.56 MB

Table 6: Model hyperparameters, parameter count, and memory usage for U-Net. For entries with a slash ("/"), the format is Prediction / Post-Processing.

Hyperparameters	**Values**
Channel	[64, 128, 256, 512]
Activation	LeakyReLU
Optimizer	AdamW
Learning Rate (prediction)	CosineAnnealing (1e-3 $\rightarrow$ 1e-4) / (5e-4 $\rightarrow$ 5e-5)
Batch Size	16 / 24
Crop Size	(200, 400) / (180, 360)
Parameters	31.1 M / 31.1 M
Memory (batch size = 1)	546.08 MB / 455.45 MB

Table 7: Model hyperparameters, parameter count, and memory usage for ViT.

Hyperparameters	**Values**
Patch Size	4
Hidden Dimension	384
Layers	8
Heads	12
MLP Dimension	768
Learning Rate	CosineAnnealing (1e-3 $\rightarrow$ 1e-4)
Batch Size	16 / 24
Crop Size	(200, 400) / (180, 360)
Parameters	12.5 M / 12.1 M
Memory (batch size = 1)	932.60 MB / 774.49 MB

Table 8: Model hyperparameters, parameter count, and memory usage for VAE.

Hyperparameters	**Values**
Encoder Channel	[64, 128, 256, 512]
Decoder Channel	[512, 256, 128, 64]
Latent Dimension	256
Activation	ReLU
Reconstruct Weight	1
KL Weight	0.001
Learning Rate	CosineAnnealing (1e-3 $\rightarrow$ 1e-4)
Batch Size	16 / 24
Crop Size	(200, 400) / (180, 360)
Parameters	497.7 M / 395.2 M
Memory (batch size = 1)	2034.48 MB / 1616.77 MB

Table 9: Model hyperparameters, parameter count, and memory usage for GraphCast.

Hyperparameters	**Values**
Mesh Level	6
Processor Layers	12
Hidden Dimension	512
Activation	SiLU
Optimizer	AdamW
Learning Rate	CosineAnnealing (1e-3 $\rightarrow$ 1e-4)
Batch Size	1
Crop Size	(181, 360)
Parameters	28.0 M
Memory (batch size = 1)	60931.46 MB

Table 10: Model hyperparameters, parameter count, and memory usage for SFNO.

Hyperparameters	**Values**
Scale Factor	4
Embed Dimension	384
Layers	8
Activation	ReLU
Learning Rate	CosineAnnealing (1e-3 $\rightarrow$ 1e-4)
Batch Size	24
Crop Size	(180, 360)
Parameters	82.7 M
Memory (batch size = 1)	1263.83 MB

G Data Preparation and Usage Instructions

For details on dataset organization, preprocessing, model training, and evaluation, please refer to our public repository `https://github.com/SauryChen/SeasonBench-EA`.

H Statement of Importance and Social Impacts

Improving seasonal prediction over complex regions like East Asia is not only a scientific challenge, but also critical for understanding and responding to the impacts of climate variability and change. Enhanced seasonal forecasts in specific regions support more effective water resource planning, disaster preparedness, and agricultural decision-making, which are increasingly vulnerable under a changing climate. However, existing AI-based forecasting methods are often constrained by limited attention to regional and climate dynamics and a lack of standardized benchmarks.

SeasonBench-EA addresses these challenges by providing a multi-resolution, multi-source dataset specifically designed for seasonal prediction in East Asia. It supports both data-driven seasonal forecasting and post-processing of numerical model outputs, a hybrid approach that combines physical insights with data-driven techniques, enabling robust model development and fair comparison across methods. By doing so, SeasonBench-EA aims to accelerate progress in AI-based seasonal prediction, particularly for precipitation, and regional-specific applications.

Beyond the core prediction tasks, the dataset's multi-resolution design also enables downscaling applications and supports the development of nested model architectures, similar to the grid nesting strategies commonly employed in regional numerical weather and climate models. Such architectures enable high-resolution regional forecasts that integrate large-scale boundary information from global contexts, while maintaining reasonable computational costs. This could also address a key challenge in current global-scale forecasting models, where massive input data can impose significant pressure on system I/O and storage resources.

I Statement of Limitations and Future Work

While SeasonBench-EA establishes a foundation for benchmarking AI-based seasonal forecasting, several limitations remain:

1. Computational constraints: Due to the high cost of data processing and training, only a subset of available numerical forecast models is currently included for post-processing evaluation. Future expansions of the evaluation are planned.
2. Boundary condition coverage: Although key boundary variables are included, additional boundary conditions, such as sea surface salinity and subsurface ocean temperatures, are expected to further improve the representation of long-term climate drivers. These variables will be progressively incorporated in future updates of the dataset.
3. Cross-model evaluation consistency: Differences in hindcast periods, ensemble sizes, and initialization strategies across numerical forecast systems introduce challenges in ensuring standardized and fair evaluation of data-driven models. In addition, differences in spatial resolution and test periods between the reanalysis data and ensemble forecasts may introduce slight inconsistencies when directly comparing the performance of direct prediction models and post-processing methods. Nevertheless, these inconsistencies are generally minor and do not affect the overall evaluation trends.

SeasonBench-EA will be continuously updated to incorporate more variables, models, and evaluation strategies.

References

[1] Hai Lin, William J Merryfield, Ryan Muncaster, Gregory C Smith, Marko Markovic, Frédéric Dupont, François Roy, Jean-François Lemieux, Arlan Dirkson, Viatcheslav V Kharin, et al. The canadian seasonal to interannual prediction system version 2 (cansipsv2). *Weather and Forecasting*, 35(4):1317–1343, 2020.

[2] Stephanie J Johnson, Timothy N Stockdale, Laura Ferranti, Magdalena A Balmaseda, Franco Molteni, Linus Magnusson, Steffen Tietsche, Damien Decremer, Antje Weisheimer, Gianpaolo Balsamo, et al. Seas5: the new ecmwf seasonal forecast system. *Geoscientific Model Development*, 12(3):1087–1117, 2019.

[3] Ray Bell, Aaron Spring, Riley Brady, Andrew Huang, Dougie Squire, Zachary Blackwood, Maximilian Cosmo Sitter, and Taher Chegini. xarray-contrib/xskillscore: Metrics for verifying forecasts.