SeasonBench-EA: A Multi-Source Benchmark for Seasonal Prediction and Numerical Model Post-Processing in East Asia

Mengxuan Chen
Tsinghua University

Guowen Li
Sun Yat-sen University

Ziheng Zou
Tsinghua SIGS

Fang Wang
CMA Earth System Modeling and Prediction Centre (CEMC)

Jinxiao Zhang
Tsinghua University

Runmin Dong
Sun Yat-sen University

Juepeng Zheng
Sun Yat-sen University

Haohuan Fu *
Tsinghua University, Tsinghua SIGS

{chenmx21, zouzh24, zhang-jx22}@mails.tsinghua.edu.cn
ligw8@mail2.sysu.edu.cn; fangwang@cma.gov.cn
{dongrm3, zhengjp8}@mail.sysu.edu.cn; haohuan@tsinghua.edu.cn

Abstract

Seasonal-scale climate prediction plays a critical role in supporting agricultural planning, disaster prevention, and long-term decision making. In particular, reliable forecasts issued 1-6 months in advance are essential for early warning of flood and drought risks associated with precipitation during the East Asian summer monsoon season. However, while the use of machine learning techniques has advanced rapidly in weather and subseasonal-to-seasonal forecasting, partly driven by the availability of benchmark datasets, their application to seasonal-scale prediction remains limited. Existing seasonal prediction primarily relies on ensemble forecasts from numerical models, which, while physically grounded, are subject to biases and uncertainties at long lead times. Motivated by these challenges, we propose SeasonBench-EA, a benchmark dataset for seasonal prediction in East Asia region. It features multi-resolution, multi-source data with both regional and global coverage, integrating ERA5 reanalysis data and ensemble forecasts from multiple leading forecast centers. Beyond key atmospheric fields, the dataset also includes boundary-related variables, such as ocean state, soil and solar radiation, that are essential for capturing seasonal-scale atmospheric variability. Two tasks are defined and evaluated: 1) machine learning-based seasonal prediction using ERA5 reanalysis, and 2) post-processing of seasonal forecasts from numerical model ensembles. A suite of deterministic and probabilistic metrics is provided for tasks evaluation, along with a hindcast assessment focused on precipitation during the East Asian summer monsoon, aligned with model evaluation protocols used in operations. By offering a unified data and evaluation framework, SeasonBench-EA aims to promote the development and application of data-driven methods for seasonal prediction, a challenging yet highly impactful task with board implications for society and public well-being. Our benchmark is available at https://github.com/SauryChen/SeasonBench-EA.

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

1 Introduction

Seasonal-scale climate prediction plays a critical role in various socioeconomic activities, including agricultural planning, disaster prevention and reduction, and water resource management. As extreme weather and climate events become more frequent under climate change, there is growing demand for accurate and reliable forecasts across multiple timescales, spanning from weather to subseasonal and seasonal predictions [1, 2, 3, 4]. In the domains of weather and subseasonal-to-seasonal (S2S) forecasting, recent progress has been supported by the availability of public datasets and benchmarks [5, 6, 7], which have accelerated the development of machine-learning models. Some of these models have demonstrated performance exceeding that of long-established earth system models [8, 9, 10]. In contrast, seasonal prediction has received relatively less attention within the field of artificial intelligence. Machine-learning models specifically designed for this task remains limited, partly due to the lack of standardized datasets and benchmarks.

Different from weather (1-15 days) and S2S (15-45 days) forecasting, seasonal targets lead time of 1–6 months, with a focus on accurately capturing monthly-mean climate states and their anomalies relative to climatology. While weather forecasting is highly sensitive to initial conditions [11], seasonal prediction depends predominantly on slowly varying boundary conditions, such as ocean states, sea ice coverage, soil temperature, and solar radiation, which regulate long-term atmospheric dynamics. Moreover, in contrast to S2S forecast that focus on the evolution of short-term disturbances [12], seasonal prediction emphasizes monthly variability and deviations from climatology normals. These fundamental differences pose unique challenges for the development and evaluation of seasonal prediction models.

To bridge this gap, we introduce SeasonBench-EA, a multi-resolution, multi-source benchmark dataset focused on seasonal prediction, with an emphasis on the East Asia region, as described in Figure 1. East Asia presents unique challenges for seasonal forecasting due to its complex monsoon systems, strong ocean-atmosphere interactions, and highly uneven spatiotemporal distribution of precipitation. These characteristics demand models that can capture regional-specific dynamics, which are often underrepresented in global-scale approaches. However, current AI-based forecasting models have been developed at global-scale [8, 13, 14], with limited exploration of regional solutions. Considering the critical role of boundary conditions in seasonal prediction and the need for regional-specific modeling, as different regions are influenced by distinct climate conditions, SeasonBench-EA provides 0.25° resolution data over East Asia (58-163°E, 8-60°N) and 1° resolution data globally. This design captures fine-scale regional features while preserving global boundary information, enabling more accurate regional forecasts under reasonable computational costs. The data configuration is also inspired by the nested-grid approach commonly used in regional numerical models, and may facilitate the development of nested architectures in AI-based seasonal prediction [15, 16].

With the collected data, we benchmark two practical tasks with various representative data-driven models to evaluate seasonal prediction in a systematic manner: 1) machine learning-based prediction

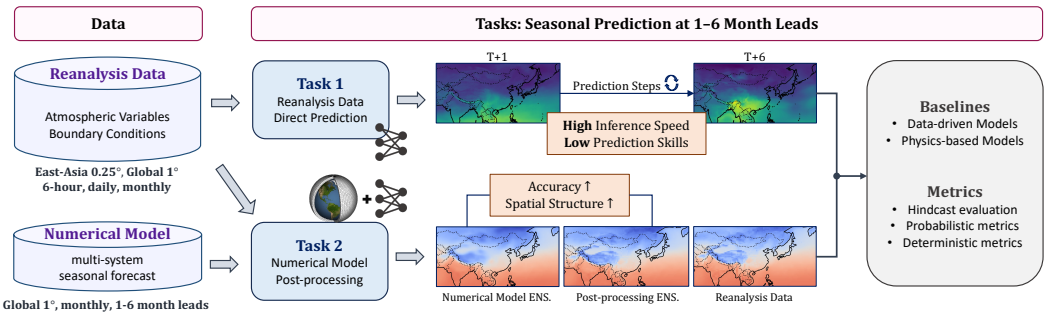


Figure 1: Overview of SeasonBench-EA, a multi-resolution, multi-source benchmark dataset designed for seasonal prediction in East Asia. It integrates ERA5 reanalysis data, including atmospheric variables and key boundary conditions, as well as ensemble seasonal forecast results from leading operational centers. SeasonBench-EA supports two tasks: 1) machine learning–based prediction from reanalysis, and 2) post-processing of the numerical model ensemble outcomes. In addition to standard deterministic and probabilistic metrics, it also provides a hindcast evaluation for assessing model's long-term predictive skill and robustness.

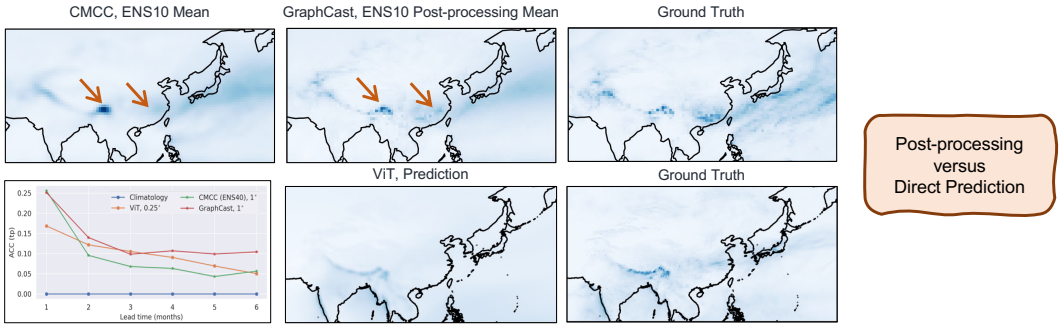


Figure 2: Overview of the performance for the two benchmark tasks at a two-month lead time, using ViT for direct prediction and GraphCast for ensemble post-processing as examples.

using ERA5 reanalysis, and 2) post-processing of numerical model ensemble forecasts. While the former directly leverages historical data, it often suffers from the blurring effect and substantial uncertainty at extended lead times [7], limiting its ability to capture sharp spatial structures and extreme events. In contrast, post-processing offers a hybrid strategy that combines physically informed guidance with data-driven corrections, which can mitigate some of these limitations (orange arrows in Figure 2). However, its effectiveness remains highly dependent on the quality of the ensemble forecasts. Moreover, the anomaly correlation coefficient (ACC) scores across both tasks underscore the current limitations and highlight the challenges faced in the seasonal prediction task.

2 Related Work

In recent years, a growing number of benchmark datasets have been introduced to support the development of data-driven models across various forecasting timescales, including weather [5, 17, 6], S2S [18, 19, 7], and long-term climate projection [20, 21]. Besides, datasets have also been proposed for related applications such as statistical downscaling [22] and extreme weather detection [23, 24]. A summary of representative datasets focused on prediction is provided in Table 1.

Gap at the seasonal timescale. While weather and climate benchmarks have advanced machine learning for medium-range weather forecasting and long-term climate projections[25, 26, 13, 8, 27, 9], a clear gap remains at the seasonal timescale. At this timescale, machine learning models exhibit blurring effects and struggle with boundary condition signals, while numerical models suffer from systematic drift and model bias.

Limited variable diversity for physical consistency. Most existing datasets beyond the weather scale focus on surface variables like temperature and precipitation [19, 18, 21, 20]. However, variables such as geopotential and specific humidity are essential for capturing circulation patterns. For instance,

Table 1: Comparison of SeasonBench-EA with existing datasets for AI-based prediction. SeasonBench-EA fills the gap between medium-range weather forecasting and long-term climate projection by supporting both prediction and post-processing with multiple variables. Note: H = Hour, D = Day, W = Week, M = Month, Y = Year. ✗ in Multi-Var means that only temperature and precipitation are included.

Benchmark	Time Res./ Lead Time	Spatial Res./ Region	Reanalysis / NWP Data	Prediction/ Post-processing	Multi-Var
ENS-10 [17]	24H / 48H	0.5° global	✓ / ✓	✗ / ✓	✓
WeatherBench2 [6]	6H / 15D	0.25° global	✓ / ✓	✓ / ✗	✓
ChaosBench [7]	1D / 44D	1.5° global	✓ / ✓	✓ / ✗	✓
SubseasonalClimateUSA [19]	2W / 6W	1° contiguous U.S	✓ / ✓	✓ / ✓	✗
SubseasonalRodeo [18]	2W / 6W	1° western U.S	✓ / ✓	✓ / ✗	✗
SeasonBench-EA	1M / 6M	0.25° EA & 1° global	✓ / ✓	✓ / ✓	✓
ClimateSet [21]	1M / 251Y	250km global	✗ / ✓	✓ / ✗	✗
ClimateBench [20]	1Y / 500Y	250km global	✗ / ✓	✓ / ✗	✗

Table 2: Reanalysis variables included in SeasonBench-EA. Variables highlighted in brown are available in the reanalysis dataset, while the others are included in both the reanalysis and the numerical model ensembles.

Type	Variables
surface	2m temperature, mean sea level pressure, total precipitation
pressure @ 1000, 850, 700, 500 200 hPa	temperature, u/v component of wind, geopotential, specific humidity
boundary	boundary layer height, surface solar radiation downwards, soil temperature, volumetric soil water layers, snow albedo, snow depth, sea surface temperature, sea ice cover
constant	geopotential at surface, land sea mask, soil type

experts often interpret geopotential anomalies to infer likely precipitation distributions, helping ensure the physical consistency of seasonal forecasts.

Seasonal prediction in East Asia. Machine learning has shown promise in seasonal prediction and model correction tasks over East Asia, especially for precipitation prediction [28, 29, 30, 31, 32]. However, existing studies are often developed for specific regions and variable sets, and adopt inconsistent evaluation protocols. Currently, no publicly available dataset supports seasonal prediction over East Asia with standardized evaluation settings, making it difficult to systematically compare methods and limiting broader participation.

3 SeasonBench-EA

SeasonBench-EA integrates ERA5 reanalysis data [33, 34, 35, 36] and ensemble forecasts [37, 38] from leading operational centers. Given that seasonal prediction is sensitive to boundary conditions, dataset covers essential atmospheric variables and critical boundary layer variables such as ocean, soil, and solar radiation, supporting robust modeling of long-term climate conditions. While evaluation is conducted at the monthly scale, aligning with the typical temporal resolution of seasonal prediction, the reanalysis data is available at hourly, daily, and monthly resolutions, and the ensemble forecasts are provided at monthly resolution, facilitating flexible training and assessment across temporal scales.

3.1 Reanalysis Data

Reanalysis data serve as input variables for the seasonal prediction task and the ground truth for numerical model post-processing. The selected variables are summarized in Table 2. Variable selection follows two criteria: the ability to characterize the large-scale atmospheric circulation, and well-established relevance to seasonal precipitation prediction [28, 31, 32, 39, 40, 41, 42], which is a key focus of seasonal prediction in the East Asian monsoon region. For example, sea surface temperature and sea ice cover influence teleconnections, while snow depth and albedo over the Tibetan Plateau affect moisture transport. The reanalysis data spans from 1940 to 2024 at monthly resolution, and from 1991 to 2024 at 6-hourly and daily resolutions. The total volume of the reanalysis data is approximately 715 GB.

3.2 Seasonal Forecasts from Numerical Model Ensembles

The variables used in numerical model ensemble are a subset of those in the reanalysis dataset, as summarized in Table 2, selected to support the ensemble post-processing task. The selection is also based on the variables' availability and temporal coverage across different forecasting systems. For instance, certain systems lack key variables: the JMA model does not provide variables at 1000 hPa, while the UK Met Office, NCEP and BOM models do not include variables such as snow depth, therefore omitted in the dataset. Additionally, we prioritize system versions of each numerical model that cover all calendar months and span longer temporal coverage. When multiple system versions are available for a numerical model, we choose the latest version available at the time of data download.

Table 3: Summary of numerical model ensemble systems included in SeasonBench-EA.

Center	CMCC	DWD	ECCC	ECMWF	Meteo-France
System	SPS 3.5	GCFS2.1	GEM5-NEMO	SEAS5	System 8
Ensemble members	40	30	10	25	25

Seasonal prediction data from five operational centers are included, as summarized in Table 3, with additional details for each system provided in Supplementary Section A.

All ensemble forecasts provide global coverage at a spatial resolution of 1°, with a monthly temporal resolution, consistent with the global reanalysis data. The dataset spans the period from 1993 to 2024, with a total data volume of approximately 1.3 TB for the multi-model ensemble component. The data processing focused on cropping the data to the East Asia region, aligning the spatial and temporal grids between the numerical model ensembles and reanalysis data, as well as performing unit conversion and normalization.

3.3 Baselines

SeasonBench-EA supports two tasks: 1) seasonal prediction based on reanalysis data, and 2) post-processing of seasonal forecasts from numerical model ensembles. While our baseline focus on a commonly used set of target variables (*t2m, tp, t_850, z_500, q_700*), the benchmark remains flexible, allowing users to define custom variable combinations for specific research needs.

We construct separate baselines for the two tasks. For the seasonal prediction task, models are trained on reanalysis data within the East Asia region, aligning with the goal of regional forecasting. For the post-processing task, models are built on global scale, but evaluated over the East Asia domain to assess improvements in regional skill.

SeasonBench-EA includes a variety of representative data-driven architectures for both tasks: U-Net [43], ViT [44], FNO [45], and VAE [46]. For the post-processing, we additionally include architectures designed for global-scale modeling, including the SFNO [14] and GraphCast [8]. Besides, the monthly climatology and persistence predictions are used as two physics baselines, following [5, 22, 7]. These baselines are distinguished from data-driven models because they rely solely on climatological statistics. A description of these model configuration is provided in Supplementary Section F.

3.4 Metrics

SeasonBench-EA provides both deterministic and probabilistic metrics that are commonly used in seasonal prediction and ensemble forecast. For deterministic metrics, we include root mean square error (RMSE), bias, Willmott's index of agreement (WI), anomaly correlation coefficient (ACC), energy spectrum, and critical success index (CSI). For probabilistic metrics, we adopt rank histogram, continuous ranked probability score (CRPS), and spread–skill ratio (SSR). A detailed description of these metrics is provided in Supplementary Section B.

3.5 Hindcast Evaluation

Hindcast evaluation provides a retrospective framework for validating predictive models by comparing their forecasts anomalies with historical observation anomalies. This approach is widely adopted in operations to assess model performance across multiple years and different initial time. In SeasonBench-EA, we employ two evaluation metrics:

Anomaly Correlation Coefficient ($\mathbf{ACC_{hindcast}}$) Unlike the standard ACC metric, $\mathrm{ACC}_{\mathrm{hindcast}}$ (Eq. 1) further removes the climatological mean specific to each data source, *i.e.* forecast climatology for predictions and observation climatology for ground truth. This adjustment helps correct for the model's systematic biases and enables a more robust and accurate evaluation of a model's ability to

capture interseasonal variability.

$$ACC = \frac{\sum_{i=1}^{H\times W}\left(\Delta f - \overline{\Delta f}\right)\left(\Delta O - \overline{\Delta O}\right)}{\sqrt{\sum_{i=1}^{H\times W}\left(\Delta f - \overline{\Delta f}\right)^2 \sum_{i=1}^{H\times W}\left(\Delta O - \overline{\Delta O}\right)^2}}, \Delta f = f - \overline{f}, \Delta O = O - \overline{O}, \quad (1)$$

where f and O represent the forecast and observe values at each grid point, $\overline{f}$ and $\overline{O}$ represent their climatological means over the evaluation years. The terms Δf and ΔO are the corresponding anomalies. H and W indicate the number of latitude and longitude grid points, respectively.

Temporal Correlation Coefficient (TCC) TCC (Eq. 2) is calculated at each spatial grid to evaluate the temporal consistency between forecasts and observations anomalies over multiple years. It reflects a model's capacity to reproduce interannual variability for the target month at local scales, which is essential for skillful seasonal prediction.

$$\text{TCC}i,j = \frac{\sum_{t=1}^{T}(f_t^{(i,j)} - \bar{f}^{(i,j)})(O_t^{(i,j)} - \bar{O}^{(i,j)})}{\sqrt{\sum_{t=1}^{T}(f_t^{(i,j)} - \bar{f}^{(i,j)})^2}\sqrt{\sum_{t=1}^{T}(O_t^{(i,j)} - \bar{O}^{(i,j)})^2}}, \quad (2)$$

where $f_t^{(i,j)}$ and $O_t^{(i,j)}$ are the forecast and observe values at grid point (i, j) for year t, while $\bar{f}^{(i,j)}$ and $\bar{O}^{(i,j)}$ are their corresponding temporal means at that grid point over the evaluation period of T years.

4 Results and Analysis

In this section, we present the results for the following variables: total precipitation (tp), 2-meter temperature (t2m), temperature at 850 hPa (t_850), geopotential at 500 hPa (z_500), and specific humidity at 700 hPa (q_700). These variables are critical for describing large-scale atmospheric circulation, with total precipitation serving as a core predictand in seasonal forecasting applications.

4.1 Prediction

For the seasonal prediction task, we adopt an auto-regressive forecasting strategy. Models are trained using reanalysis data from 1940 to 2015, validated on 2016 to 2019, and evaluated over 2020 to 2024. Monthly climatology for anomaly computation is derived from the 1991-2020 reference period. The years 2020 to 2024 are selected for testing due to their diverse seasonal precipitation patterns over East Asia, allowing for evaluation under a range of climate conditions (see Supplementary Section C). Additional experiments, including a simple linear regression model, rolling-window evaluations to assess temporal robustness, multi-seed training to evaluate model stability, and detailed results for the seasonal prediction task, are provided in Supplementary Section D.

Loss of predictive skills relative to climatology. As shown in Figure 3 (a) and (b), all models exhibit higher RMSE compared to the monthly climatology baseline, with ACC values even dropping below zero at several lead months. This consistent performance degradation across different architectures suggests that the limitation is not specific to any particular model design, but rather stems from fundamental challenges in seasonal prediction. In particular, the baselines lack sensitivity to boundary-driven signals such as solar radiation and sea surface temperature, which are critical at seasonal timescales. Moreover, they fail to incorporate the broader environmental context, where global or surrounding regional conditions serve as essential boundary constraints for local climate evolution. These findings highlight the need to develop models that are physically informed, aware of boundary conditions, and capable of capturing long-range spatiotemporal dependencies.

Lack of small-scale variability in model predictions. The energy spectrum plots depicted in Figure 3 (d) show that models exhibit a significant reduction in spectral amplitude at high wavenumbers compared to the reanalysis, particularly for tp. This indicates a substantial loss of small-scale variability in the predictions. At a six-month lead time, the predicted tp fails to reproduce detailed spatial structures and deviates from the climatology patterns (Figure 3 (c)). In addition, all models except FNO show spurious peaks at certain small-scale wavenumbers, which may reflect the instability in predicting fine-scale processes. Potential contributing factors include the lack of physical constraints, error accumulation in auto-regressive inference, and the limited capacity of pixel-level loss functions to penalize spatial discontinuities.

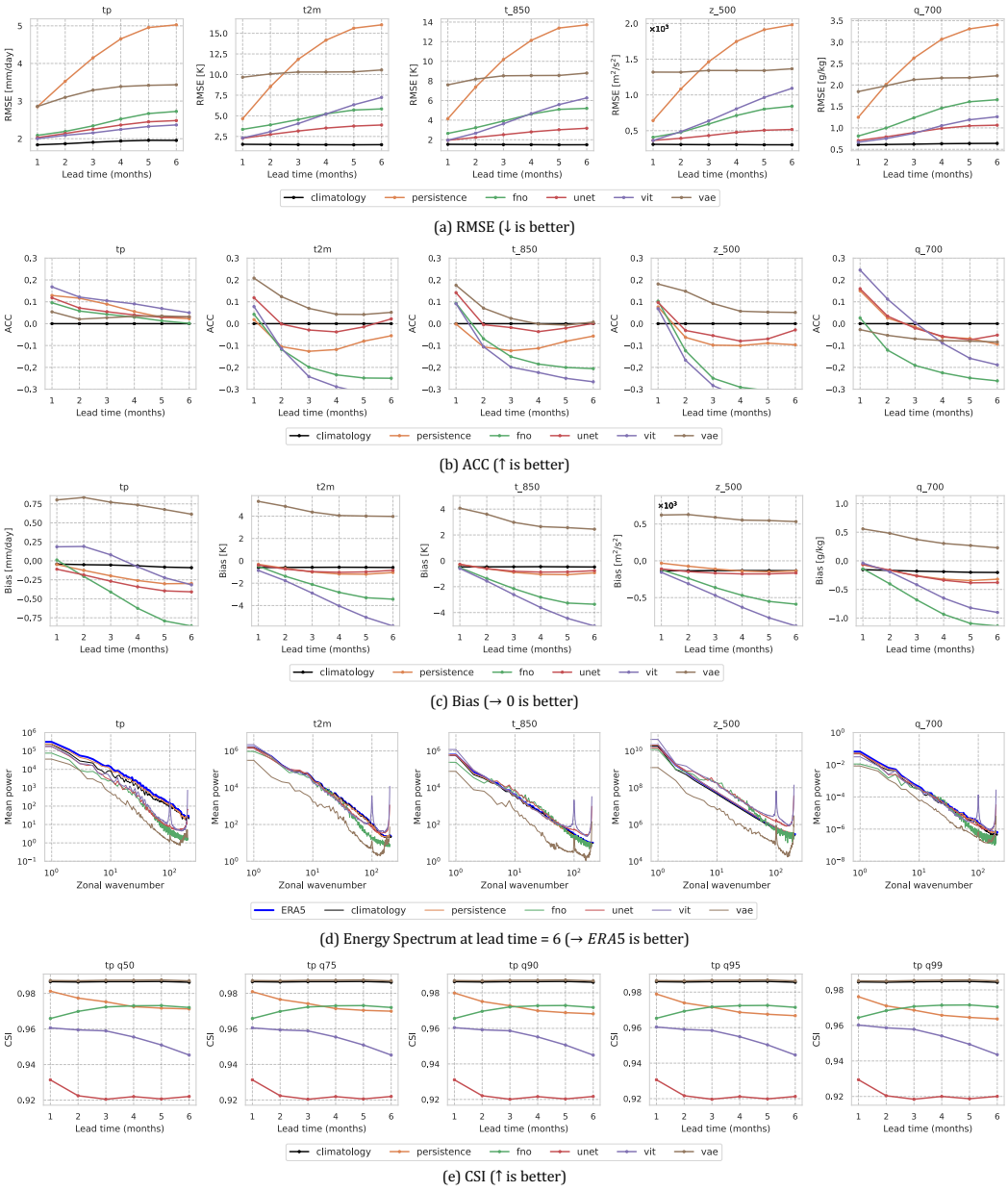


Figure 3: Evaluation of seasonal prediction performance using deterministic metrics.

Limited benefits from longer autoregressive steps. In medium-range weather forecasting and S2S prediction, training models with longer autoregressive steps has been shown to improve stability and performance by better capturing temporal dependencies [7]. However, extended steps fails to enhance model performance in this task. As shown in Figure 4, model performance does not improve monotonically with longer training steps. The growing uncertainty and weakened signal-to-noise ratio at long lead times could limit the effectiveness of autoregressive learning strategies. Further details are presented in Supplementary Section D.

4.2 Post-Processing

For the post-processing of seasonal forecasts from numerical model ensembles, models are trained to directly output corrected variable fields at all lead times. Training is conducted on global-scale data to incorporate large-scale boundary information, while evaluation is performed over East Asia

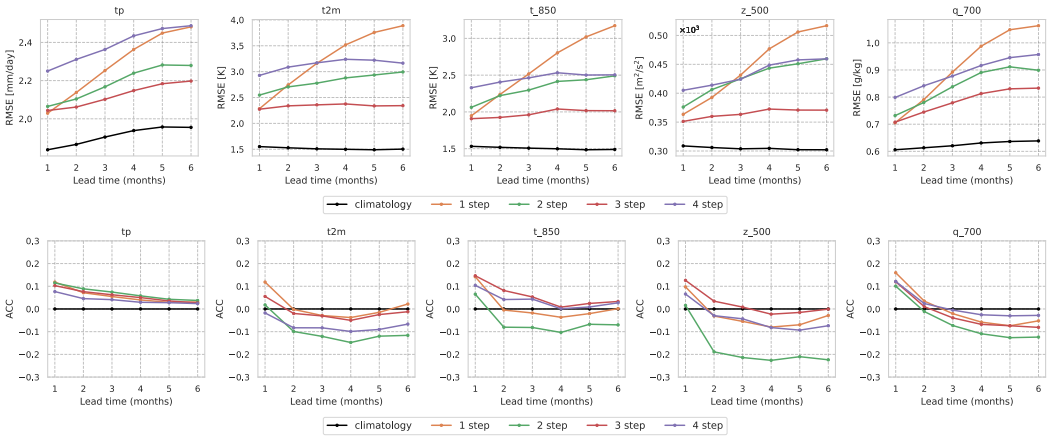


Figure 4: Performance comparison of U-Net models trained with different autoregressive steps.

to ensure consistency with the prediction task setup. Specifically, models are trained on data from 1993 to 2024, excluding the validation (2009–2011) and test (2013–2016) periods. The test period is chosen to accommodate inconsistencies in data coverage, as several numerical models have missing years in their forecast records. The monthly climatology is computed using the 1991–2020 period from reanalysis data with a global resolution of 1°. Figure 5 shows the post-processing results using ensemble forecasts from CMCC as an example. Following [17], SeasonBench-EA uses the first 10 ensemble members. Additional evaluation results for post-processing, multi-seed training to assess model stability, and GraphCast-based results for ECMWF are listed in Supplementary Section E.

Numerical model post-processing improves forecasting skills. Post-processed results based on 10 ensemble members outperform the original 40-member ensemble in both RMSE and ACC, demonstrating that data-driven correction can effectively improve forecast accuracy. Notably, the post-processed forecasts also surpass the direct prediction models, highlighting the advantage of incorporating numerical model guidance and ensemble diversity to enhance seasonal prediction. From the perspective of energy spectrum, spurious peaks at small scales are significantly reduced, suggesting improved physical consistency and spatial coherence. Despite the improvements, ACC values remain below the threshold of skillful prediction, indicating that the post-processed forecasts still fall short in capturing accurate anomaly signals, particularly at longer lead times.

Precipitation correction remains challenging. Among all target variables, the improvement from post-processing is limited for total precipitation. As shown in Figure 5 (c), the corrected forecasts exhibit a significant drop in amplitude at high wavenumbers, indicating a failure to recover fine-scale spatial variability. Additionally, Figure 5 (e) shows that the rank histogram for total precipitation retains a U-shaped pattern after correction, suggesting that the ensemble remains underdispersive. Notably, the climatology baseline continues to outperform the corrected forecasts, especially for precipitation. During training, models optimized with RMSE-based objectives are likely to converge to average, leading to smooth predictions and the loss of spatial details. These results highlight the need for improved model architectures, such as GraphCast, and loss functions that better preserve spatial details and capture precipitation variability.

4.3 Hindcast

We further perform a hindcast evaluation on three representative cases: 1) direct prediction using the ViT model, 2) the ensemble mean of the first 10 members from CMCC, and 3) the ensemble forecasts post-processed by GraphCast. The evaluation targets total precipitation during the summer season (June–August), with all forecasts initialized in March. To assess both spatial and temporal forecast skills, we report $\text{ACC}_{\text{hindcast}}$ and TCC. The hindcast period spans 2006-2020, while 2021–2024 is used for validation. All remaining years are included in the training set. Both the GraphCast and ViT models are trained with four random seeds to compute the mean and standard deviation of their hindcast performance. The hindcast results for total precipitation during the summer season are demonstrated in Figure 6.

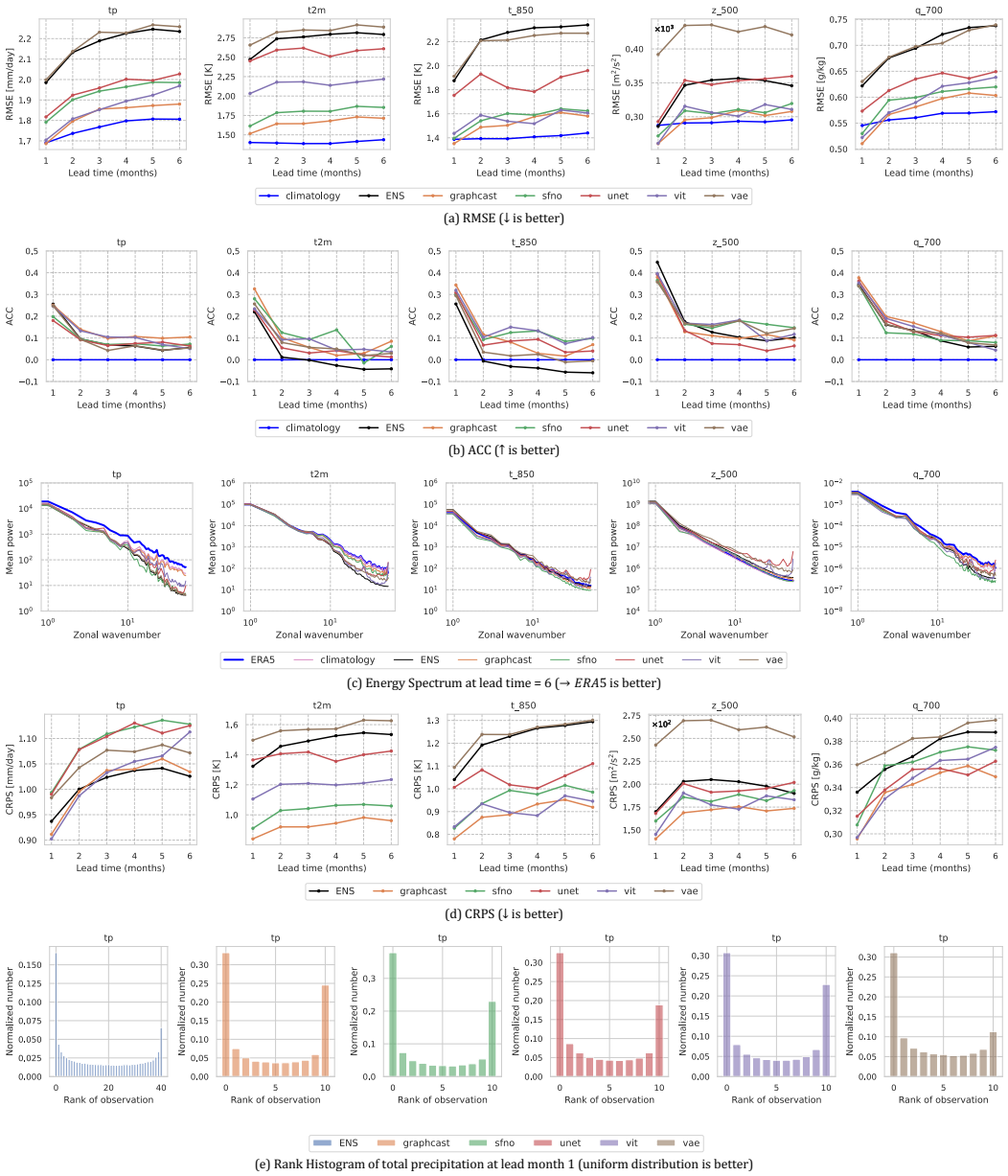
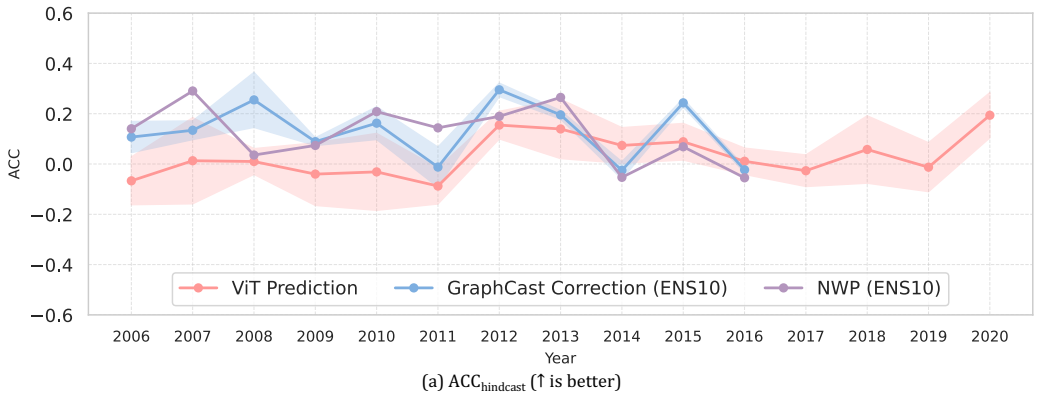
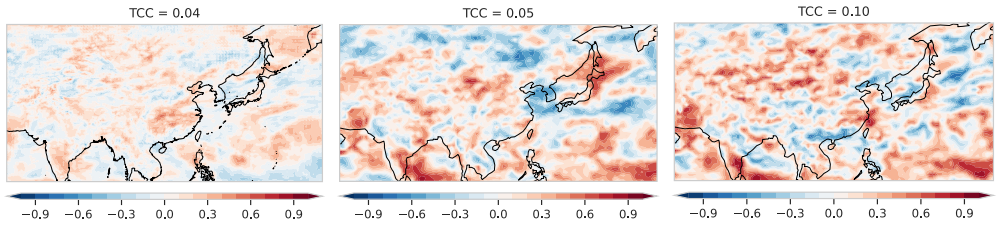


Figure 5: Post-processing results for the ensemble forecasts. "ENS" refers to the original 40-member ensemble forecasts from CMCC, while the other models apply correction using 10 ensemble members.

Physically informed models outperform data-driven ones, but challenges persist. Although all ACC values fluctuate around zero, with the value of 0.032 ± 0.131 (ViT), 0.129 ± 0.122 (GraphCast), and 0.119 (numerical ensemble), physically informed methods, such as GraphCast-based post-processing and numerical ensemble forecasts, outperform the purely data-driven ViT model. This highlights the benefit of incorporating physical priors or leveraging ensemble diversity to improve predictive skill at seasonal timescales. However, these gains in spatial accuracy do not necessarily translate into better temporal consistency. Despite achieving higher ACC values, post-processed forecasts exhibit a decline in TCC scores compared to raw ensembles, suggesting limited preservation of interannual coherence. This may originate from the strong dependence of post-processing models on numerical inputs, and the fact that commonly used loss functions primarily emphasize spatial accuracy while overlooking year-to-year variability.



(a) $ACC_{hindcast}$ (↑ is better)



(b) TCC (↑ is better), from left to right: ViT prediction, GraphCast correction and NWP with 10 ensemble members.

Figure 6: Hindcast results for total precipitation during the summer season from 2006 to 2020, with March as the initialization month. Note that CMCC SPS3.5 lacks data from 2017 to 2020, leading to missing values in those years.

5 Conclusion

In this work, we present SeasonBench-EA, a multi-resolution, multi-source benchmark dataset designed to advance data-driven research in season prediction. By integrating ERA5 reanalysis and numerical ensemble forecasts from leading centers, SeasonBench-EA provides a unified framework to evaluate both direct prediction and post-processing tasks. Our benchmark across a range of representative models reveal that the performance remains limited, especially for key variables like precipitation. Directed prediction models can reproduce large-scale variability but lose skill rapidly with increasing lead time, showing oversmoothed and low-variance forecasts due to the lack of explicit physical constraints. Post-processing models, leverage ensemble diversity to enhance spatial accuracy and reduce small-scale noise, yet remain limited by biases in numerical forecasts and weakened temporal coherence. Overall, physically informed methods outperform purely data-driven ones at the current stage, underscoring the importance of boundary information and physical priors. It demands further methodological innovations that can effectively incorporate physical constraints, capture boundary-driven variability, and model long-range spatiotemporal dependencies.

Several limitations remain in SeasonBench-EA. Due to computational constraints, not all available numerical forecast models are evaluated (additional results for ECMWF are provided in Section E). Also, while current dataset includes key atmospheric and boundary variables, incorporating additional boundary conditions, such as sea surface salinity and subsurface ocean temperatures, could further improve the representation of long-term drivers of seasonal variability. We provide data download script so that users can add variables to the dataset for their specific research need. SeasonBench-EA will be continuously updated to include more data sources, variables, and evaluation protocols.

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. Although nested data-driven models have not yet been benchmarked in this work, the dataset provides a solid foundation for the future exploration. Such architectures allow for high-resolution regional predictions that incorporate broader-scale boundary conditions from global contexts, while maintaining reasonable computational costs. This is particularly valuable in regions like East Asia, where complex land–ocean–atmosphere interactions demand both local precision and global awareness.

Acknowledgments and Disclosure of Funding

The work is supported by the National Key Research and Development Plan of China (Grant 2023YFB3002400) and National Natural Science Foundation of China (Grant T2125006).

References

[1] Thomas Stocker. *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge university press, 2014.

[2] Ying Sun, Xuebin Zhang, Francis W Zwiers, Lianchun Song, Hui Wan, Ting Hu, Hong Yin, and Guoyu Ren. Rapid increase in the risk of extreme summer heat in eastern china. *Nature Climate Change*, 4(12):1082–1085, 2014.

[3] Peter Stott. How climate change affects extreme weather events. *Science*, 352(6293):1517–1518, 2016.

[4] NN Ridder, AM Ukkola, AJ Pitman, and SE Perkins-Kirkpatrick. Increased occurrence of high impact compound events under climate change. *Npj Climate and Atmospheric Science*, 5(1):3, 2022.

[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 Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. 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 A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 43715–43729. Curran Associates, Inc., 2024.

[8] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

[9] 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(8027):1060–1066, 2024.

[10] Lei Chen, Xiaohui Zhong, Hao Li, Jie Wu, Bo Lu, Deliang Chen, Shang-Ping Xie, Libo Wu, Qingchen Chao, Chensen Lin, et al. A machine learning model that outperforms conventional global subseasonal forecast models. *Nature Communications*, 15(1):6425, 2024.

[11] Tilmann Gneiting and Adrian E Raftery. Weather forecasting with ensemble methods. *Science*, 310(5746):248–249, 2005.

[12] Frédéric Vitart and Andrew W Robertson. The sub-seasonal to seasonal prediction project (s2s) and the prediction of extreme events. *npj climate and atmospheric science*, 1(1):3, 2018.

[13] Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

[14] Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere. In *International conference on machine learning*, pages 2806–2823. PMLR, 2023.

[15] Joel Oskarsson, Tomas Landelius, Marc Deisenroth, and Fredrik Lindsten. Probabilistic weather forecasting with hierarchical graph neural networks. *Advances in Neural Information Processing Systems*, 37:41577–41648, 2024.

[16] Yuan Gao, Hao Wu, Ruiqi Shu, Huanshuo Dong, Fan Xu, Rui Ray Chen, Yibo Yan, Qingsong Wen, Xuming Hu, Kun Wang, et al. Oneforecast: a universal framework for global and regional weather forecasting. *arXiv preprint arXiv:2502.00338*, 2025.

[17] Saleh Ashkboos, Langwen Huang, Nikoli Dryden, Tal Ben-Nun, Peter Dueben, Lukas Gianinazzi, Luca Kummer, and Torsten Hoefler. Ens-10: A dataset for post-processing ensemble weather forecasts. *Advances in Neural Information Processing Systems*, 35:21974–21987, 2022.

[18] Jessica Hwang, Paulo Orenstein, Judah Cohen, Karl Pfeiffer, and Lester Mackey. Improving subseasonal forecasting in the western u.s. with machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2325–2335, New York, NY, USA, 2019. Association for Computing Machinery.

[19] Soukayna Mouatadid, Paulo Orenstein, Genevieve Flaspohler, Miruna Oprescu, Judah Cohen, Franklyn Wang, Sean Knight, Maria Geogdzhayeva, Sam Levang, Ernest Fraenkel, and Lester Mackey. Subseasonalclimateusa: A dataset for subseasonal forecasting and benchmarking. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 7960–7992. Curran Associates, Inc., 2023.

[20] Duncan Watson-Parris, Yuhan Rao, Dirk Olivié, Øyvind Seland, Peer Nowack, Gustau Camps-Valls, Philip Stier, Shahine Bouabid, Maura Dewey, Emilie Fons, et al. Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems*, 14(10):e2021MS002954, 2022.

[21] Julia Kaltenborn, Charlotte Lange, Venkatesh Ramesh, Philippe Brouillard, Yaniv Gurwicz, Chandni Nagda, Jakob Runge, Peer Nowack, and David Rolnick. Climateset: A large-scale climate model dataset for machine learning. *Advances in Neural Information Processing Systems*, 36:21757–21792, 2023.

[22] Tung Nguyen, Jason Jewik, Hritik Bansal, Prakhar Sharma, and Aditya Grover. Climatelearn: Benchmarking machine learning for weather and climate modeling. *Advances in Neural Information Processing Systems*, 36:75009–75025, 2023.

[23] Evan Racah, Christopher Beckham, Tegan Maharaj, Samira Ebrahimi Kahou, Mr Prabhat, and Chris Pal. Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. *Advances in neural information processing systems*, 30, 2017.

[24] Prabhat, Karthik Kashinath, Mayur Mudigonda, Sol Kim, Lukas Kapp-Schwoerer, Andre Graubner, Ege Karaismailoglu, Leo von Kleist, Thorsten Kurth, Annette Greiner, et al. Climatenet: An expert-labelled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development Discussions*, 2020:1–28, 2020.

[25] Qing Yi Feng, Ruggero Vasile, Marc Segond, Avi Gozolchiani, Yang Wang, Markus Abel, Shilomo Havlin, Armin Bunde, and Henk A Dijkstra. Climatelearn: A machine-learning approach for climate prediction using network measures. *Geoscientific Model Development Discussions*, 2016:1–18, 2016.

[26] Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, et al. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

[27] Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

[28] Jialin Wang, Jing Yang, Hong-Li Ren, Jinxiao Li, Qing Bao, and Miaoni Gao. Dynamical and machine learning hybrid seasonal prediction of summer rainfall in china. *Journal of Meteorological Research*, 35(4):583–593, 2021.

[29] Qimin Deng, Peirong Lu, Shuyun Zhao, and Naiming Yuan. U-net: A deep-learning method for improving summer precipitation forecasts in china. *Atmospheric and Oceanic Science Letters*, 16(4):100322, 2023.

[30] Peter B Gibson, William E Chapman, Alphan Altinok, Luca Delle Monache, Michael J DeFlorio, and Duane E Waliser. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, 2(1):159, 2021.

[31] Weixin Jin, Yong Luo, Tongwen Wu, Xiaomeng Huang, Wei Xue, and Chaoqing Yu. Deep learning for seasonal precipitation prediction over china. *Journal of Meteorological Research*, 36(2):271–281, 2022.

[32] Peirong Lu, Qimin Deng, Shuyun Zhao, Yongguang Wang, and Wuke Wang. Deep learning for seasonal prediction of summer precipitation levels in eastern china. *Earth and Space Science*, 10(11):e2023EA003129, 2023.

[33] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on single levels from 1940 to present, 2023.

[34] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on pressure levels from 1940 to present, 2023.

[35] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 monthly averaged data on single levels from 1940 to present, 2023.

[36] H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 monthly averaged data on pressure levels from 1940 to present, 2023.

[37] Copernicus Climate Change Service. Seasonal forecast monthly statistics on single levels, 2018.

[38] Copernicus Climate Change Service. Seasonal forecast monthly statistics on pressure levels, 2018.

[39] Yanbo Nie, Jianqi Sun, and Jiehua Ma. Seasonal prediction of summer extreme precipitation frequencies over southwest china based on machine learning. *Atmospheric Research*, 294:106947, 2023.

[40] Nick Dunstone, Doug M Smith, Steven C Hardiman, Paul Davies, Sarah Ineson, Shipra Jain, Chris Kent, Gill Martin, and Adam A Scaife. Windows of opportunity for predicting seasonal climate extremes highlighted by the pakistan floods of 2022. *Nature Communications*, 14(1):6544, 2023.

[41] Xuan Tong and Wen Zhou. Assessing predictive attribution in nmme forecasts of summer precipitation in eastern china using deep learning. *npj Climate and Atmospheric Science*, 7(1):304, 2024.

[42] Jieru Ma, Hong-Li Ren, Ming Cai, Yi Deng, Chenguang Zhou, Jian Li, Huizheng Che, and Lin Wang. Skillful seasonal predictions of continental east-asian summer rainfall by integrating its spatio-temporal evolution. *Nature Communications*, 16(1):273, 2025.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[44] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[45] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[46] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: While recent research has primarily focused on weather forecasting and subseasonal-to-seasonalprediction, less attention has been given to the equally challenging task of seasonal prediction. This work addresses that gap by collecting two datasets, reanalysis data and numerical model ensemble forecasts data, and evaluate a range of baseline models on two core tasks central to advancing seasonal prediction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: answerYes

   Justification: The limitations are discussed in Conclusion.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our benchmark do not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The dataset and code are publicly released, and the training and evaluation details are provided in the supplemental material.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer:[Yes]

   Justification: SeasonBench-EA dataset is publicly available at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EPEUGO`, and the link to the code is provided in the Abstract.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The data splits and how they were chosen are described in Section 4, the hyperparameters, type of optimizer, etc. are provided in the supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Several experiments are repeated with four random seeds to ensure statistical robustness. We also examine the impact of long-term data shifts in prediction tasks.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Information about compute resources is included in the supplemental material.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have carefully reviewed the NeurIPS Code of Ethics. All data used to construct the benchmark are publicly available and properly cited.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Societal impacts are discussed in Section 5.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: While the paper itself does not pose high risk, we note that individuals are not authorized to issue official weather or climate forecasts.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the model and data used in SeasonBench-EA are properly cited in the paper.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: SeasonBench-EA dataset is publicly available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EPEUGO, code is submitted to GitHub. Both the dataset and the code are accompanied by documentation that describes the structure and usage of the assets.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: We do not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: LLM is used only for writing, editing, and formatting purposes.