AGU ADVANCING EARTH AND SPACE SCIENCES

JGR Machine Learning and Computation

RESEARCH ARTICLE
10.1029/2025JH000993

Distribution-Guided Ensemble Postprocessing for S2S Precipitation Forecasts: A Seamless Pathway Using Deep Generative Models

Wen Shi[1], Baoxiang Pan[2], Jianbin Huang[3,4], Tingfeng Dou[4,5], Jie Feng[6], Huihui Yuan[1], Anran Wang[1], Xin Xia[1], Lei Huang[7,8,9], and Yong Luo[1]

[1]Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, China, [2]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China, [3]Beijing Yanshan Earth Critical Zone National Research Station, University of Chinese Academy of Sciences, Beijing, China, [4]College of Resources and Environment, University of Chinese Academy of Sciences, Beijing, China, [5]Key Laboratory of Earth System Numerical Modeling and Application, Chinese Academy of Sciences, Beijing, China, [6]Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China, [7]National Satellite Meteorological Center (National Center for Space Weather), Beijing, China, [8]Innovation Center for FengYun Meteorological Satellite (FYSIC), Beijing, China, [9]Key Laboratory of Radiometric Calibration and Validation for Environmental Satellites China Meteorological Administration, Beijing, China

Key Points:

- Generative post-processing improves probabilistic and extreme event skill, meanwhile alleviates blurry effect and physical inconsistency
- Seamless weather-to-climate training is the key to making full use of data across all S2S forecast horizons for robust post-processing
- Proposed framework also disentangles uncertainty sources, providing an effective advance estimate of local forecast error

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Y. Luo,
yongluo@tsinghua.edu.cn

Citation:

Shi, W., Pan, B., Huang, J., Dou, T., Feng, J., Yuan, H., et al. (2026). Distribution-guided ensemble postprocessing for S2S precipitation forecasts: A seamless pathway using deep generative models. *Journal of Geophysical Research: Machine Learning and Computation*, *3*, e2025JH000993. https://doi.org/10.1029/2025JH000993

Received 1 SEP 2025
Accepted 27 FEB 2026

Author Contributions:

Funding acquisition: Yong Luo
Project administration: Yong Luo
Supervision: Yong Luo
Writing – review & editing: Yong Luo

Abstract

Atmosphere-ocean-land coupled forecasting systems, despite their comprehensiveness, face substantial challenges in the "predictability desert" at subseasonal to seasonal (S2S) timescales, particularly for precipitation—a variable crucial for socioeconomic activities yet of stunning spatiotemporal variance. Post-processing methods developed for numerical weather prediction and climate projections are not directly applicable to S2S forecasting, as they cannot distinctively address initialization errors, chaos-induced state uncertainty growth, and model systematic biases. Additionally, regression-based deep learning corrections introduce smoothing artifacts and ensemble under-dispersion, limiting their ability to capture key processes and extreme events. We propose an integrated framework using Generative Adversarial Networks (GANs) for ensemble post-processing. The approach exploits the ability of deep generative models to represent high-dimensional distributions, combining trajectory constraints from short-term forecasts with distributional constraints from long-term climatology. In a case study using ECMWF hindcasts over Southern China, our model calibrates ensemble forecasts while enhancing both ensemble size and spatial resolution. The post-processed forecasts maintain deterministic skill (anomaly correlation coefficient) while showing improved probabilistic forecast metrics, such as Continuous Ranked Probability Score (CRPS) and Brier score, extending the skillful probabilistic forecast horizon to week 3. The predicted fields demonstrate improved spatial distribution matching and maintain linear covariance across variables. The framework demonstrates strong spread-error correlations for effective advance error estimation, and helps disentangle forecast uncertainties into propagated dynamical and post-processing components, each with distinct lead-time dependencies. This unified framework demonstrates the potential to advance seamless forecast capabilities while addressing the growing demand for high-resolution, physically consistent S2S products.

Plain Language Summary

Forecasting weather conditions from a few days to several weeks ahead is a significant challenge. The physics-based models used for this task have complex errors, making additional statistical correction (i.e., post-processing) essential. The atmosphere's chaotic nature creates a key challenge: the post-processing methods for the first few days of a forecast and for predictions many weeks ahead are treated as separate problems. This creates a dilemma for developing a unified method capable of providing consistent forecasts across this critical intermediate timescale. We have developed a new framework using deep generative models—advanced statistical tools that learn the full distribution of observation data—to resolve this dilemma. For precipitation forecasting, our method generates a complete set of forecasts, each with realistic spatial detail, to better represent the full range of possibilities for potential weather conditions. This not only improves forecast skill from multiple perspectives but also provides a more informative view of the uncertainty in the forecast, detailing its different sources. Such products can provide decision-makers with more usable information about the likelihood of extreme events in the coming weeks.

1. Introduction

Subseasonal to seasonal (S2S) prediction bridges the gap between medium-range weather forecasts and seasonal climate predictions, covering lead times from approximately 2 weeks to 2 months. Accurate S2S forecasts, particularly for precipitation, provide essential information for decision makers about water resource management, agricultural planning, and disaster early warnings, yielding substantial socioeconomic value (White et al., 2017). Yet, this timescale has long been considered a "predictability desert" as predictability signal from initial atmospheric state is eroded by chaos, while boundary forcings from ocean and other slowly evolving climate components have not fully developed (Vitart et al., 2012).

Efforts in S2S prediction are increasingly converging on seamless forecasting frameworks, pursued both through unified dynamical models (Brown et al., 2012) and rapidly evolving data-driven approaches (Ling et al., 2024). By explicitly coupling processes across time-scales, such seamless frameworks aim to enhance forecast skill by capturing more comprehensive physics and multiple sources of predictability. They promise to deliver temporally consistent forecasts applicable across diverse downstream application domains (Brunet et al., 2010). Meanwhile, there is growing demand to move beyond skillful mean-state predictions toward diverse, probabilistic products at higher spatial-temporal resolutions, solidifying ensemble forecasting as a standard practice and fostering the development of more high-resolution models (Molod et al., 2020; J. Li et al., 2021; Li et al., 2022). These advances are essential for informing decision-makers more detailed about high-impact events (Robertson et al., 2020).

Despite these promises, current prediction systems show limited forecast skill and substantial uncertainty at S2S timescales, particularly for precipitation. The skills of precipitation forecast rapidly deteriorate within the second week across most global and regional domains (de Andrade et al., 2019; Liang & Lin, 2018). Dynamical forecasting system products suffer from systematic biases and insufficient spatial resolution to capture local or regional variabilities. While developing new dynamical or data-driven models remains prohibitively expensive for most institutions, post-processing existing forecast products presents a more accessible approach for improving forecast accuracy and usability.

Post-processing across S2S timescales is necessary for the reasons above, yet presents several significant challenges. S2S model errors stem from multiple sources, including assimilation errors, initial shocks, climate drift, and parameterization errors (with the latter being particularly problematic for precipitation) (Merryfield et al., 2020). Each error source contributes distinct uncertainties that not only interact with each other but also with the inherent chaotic uncertainty of the dynamical system. Furthermore, extracting flow-dependent systematic biases while properly representing and disentangling forecast uncertainties requires advanced approaches beyond conventional methods.

Given these challenges, existing post-processing methodologies remain limited, with most attempts emerging only recently. One class of approaches extends regression-based techniques from the post-processing of numerical weather prediction (NWP), employing statistical models (Vigaud et al., 2020), conventional machine learning methods (Fan et al., 2023; Yin et al., 2023), and deep learning methods (Dong et al., 2024; Horat & Lerch, 2024; Lyu et al., 2023; Nie & Sun, 2024) to establish paired relationships between dynamical model outputs and corresponding observations. Another class of approaches adapts distribution-based methods from the post-processing of climate projections, employing parametric adjustments of the overall data distribution such as quantile mapping (Kolachian & Saghafian, 2019; Li et al., 2019; Monhart et al., 2018), Bayesian methods (M. Li et al., 2021; Y. Li et al., 2021; Schepen et al., 2018; Specq & Batté, 2020; Zhang et al., 2025) or more complex parametric forms (Huang et al., 2022).

Both classes of approaches face significant limitations when applied to S2S timescale. Most post-processed targets are aggregated to relatively coarse temporal resolutions (pentad to biweekly), focusing on mean-state predictions that inadequately capture extreme events and weather processes. While some studies take extreme event occurrence or probability directly as targets of optimization (Horat & Lerch, 2024; Liu et al., 2023; Lyu et al., 2024; Zhang et al., 2023), these single-purpose products sacrifice consistency with skillful fundamental precipitation forecasts. Spatially, most methods, particularly distribution-based approaches, are limited to grid-point or station-specific modeling, and fail to preserve the spatial coherence necessary for high-resolution gridded products. Furthermore, few methods adequately address forecast uncertainty, let alone disentangle contributions from different sources of uncertainty.

10.1029/2025JH000993

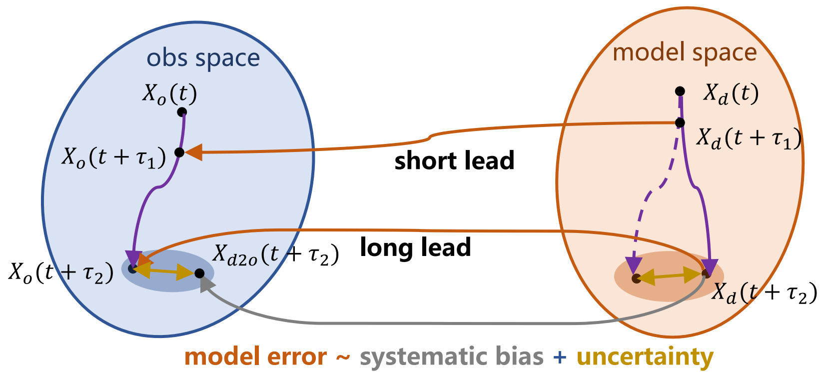

Figure 1. Conceptual diagram of the key challenge posed by chaotic divergence in S2S post-processing. The solid purple line represents the evolution of observed or modeled state (X_o, X_d), while the dashed purple line is a hypothetical perfect forecast trajectory containing only systematic bias. At short leads (τ_1), the dynamical forecast X_d remains paired with the observation. At long leads (τ_2), chaotic divergence breaks this pairing. Directly modeling the total model error becomes ineffective as it is confounded by chaotic uncertainty (represented by shaded areas).

Such limitations suggest that post-processing methods from either NWP or climate projection domain cannot be directly applied to S2S timescales. S2S hindcast data sets typically span only about few years, raising questions about the effectiveness of distribution-based methods under insufficient sample sizes. A parallel and more fundamental challenge is that S2S forecast trajectories cross typical atmospheric predictability limits due to chaotic divergence, making conventional regression-based post-processing methods ineffective, as conceptualized in Figure 1. For short-lead samples (τ_1), a regression-based approach to model the error and achieve correction is feasible. The key challenge lies with long-lead samples (τ_2). As predictability decays, the model state and the observed state become fundamentally unpaired. The error of the dynamical model is a composite of the systematic bias (the target "signal" for correction) and nonnegligible chaotic uncertainty (which acts as "noise" for a regression-based method). This poses a dilemma for any regression-based method: it must either discard the long-lead samples, which constitute the majority of the available data, or attempt to derive a correction from samples that inherently possess a low signal-to-noise ratio.

A post-processing methodology that remains valid across both short- and long-lead regimes is essential for seamless S2S forecast to deliver temporally consistent forecast information. Barthel Sorensen et al. (2024) addressed this through nudging techniques to create paired samples, but this approach demands substantial computational resources and expertise in model development. As a simpler alternative, Yin et al. (2023) restricted training to the first 7 days of trajectories to mitigate this issue. While extending short-lead corrections to longer leads is conceptually reasonable given the shared underlying physics (Rodwell & Palmer, 2007), its quantitative application requires careful consideration, as post-processing relationships learned on short-lead samples may overrely on assimilation errors and initial shocks.

In recent years, deep generative models have demonstrated remarkable capabilities in distribution modeling. By learning high-dimensional distributions, they can produce spatiotemporally coherent and high-resolution fields, proving effective for applications such as nowcasting (Nai et al., 2024; Ravuri et al., 2021), downscaling (Hess et al., 2025), and data assimilation (Chao et al., 2025; Sun et al., 2025; Yang, Nai, et al., 2025). Furthermore, their capacity to generate large ensembles offers a powerful approach to better represent forecast uncertainty (Nai et al., 2025; Pan et al., 2022; Price et al., 2025). In this study, we propose GAN-W2C (weather-to-climate), a generative framework that seamlessly integrates regression-based learning and distribution matching methods, providing a pathway to address the multiple challenges faced in S2S post-processing. The proposed method aims to (a) provide high-resolution, realistic forecasts that preserve extreme event information, (b) enhance probabilistic and extreme event forecast skill, (c) deliver extra uncertainty estimates that effectively predict product errors and explain uncertainty sources.

The rest of this paper is organized as follows. Section 2 introduces our methodology, outlining the design principles of the proposed GAN-W2C framework, the data sets, and the experimental setup. Section 3 presents a comprehensive evaluation of the results, assessing model performance in terms of probabilistic and deterministic

10.1029/2025JH000993

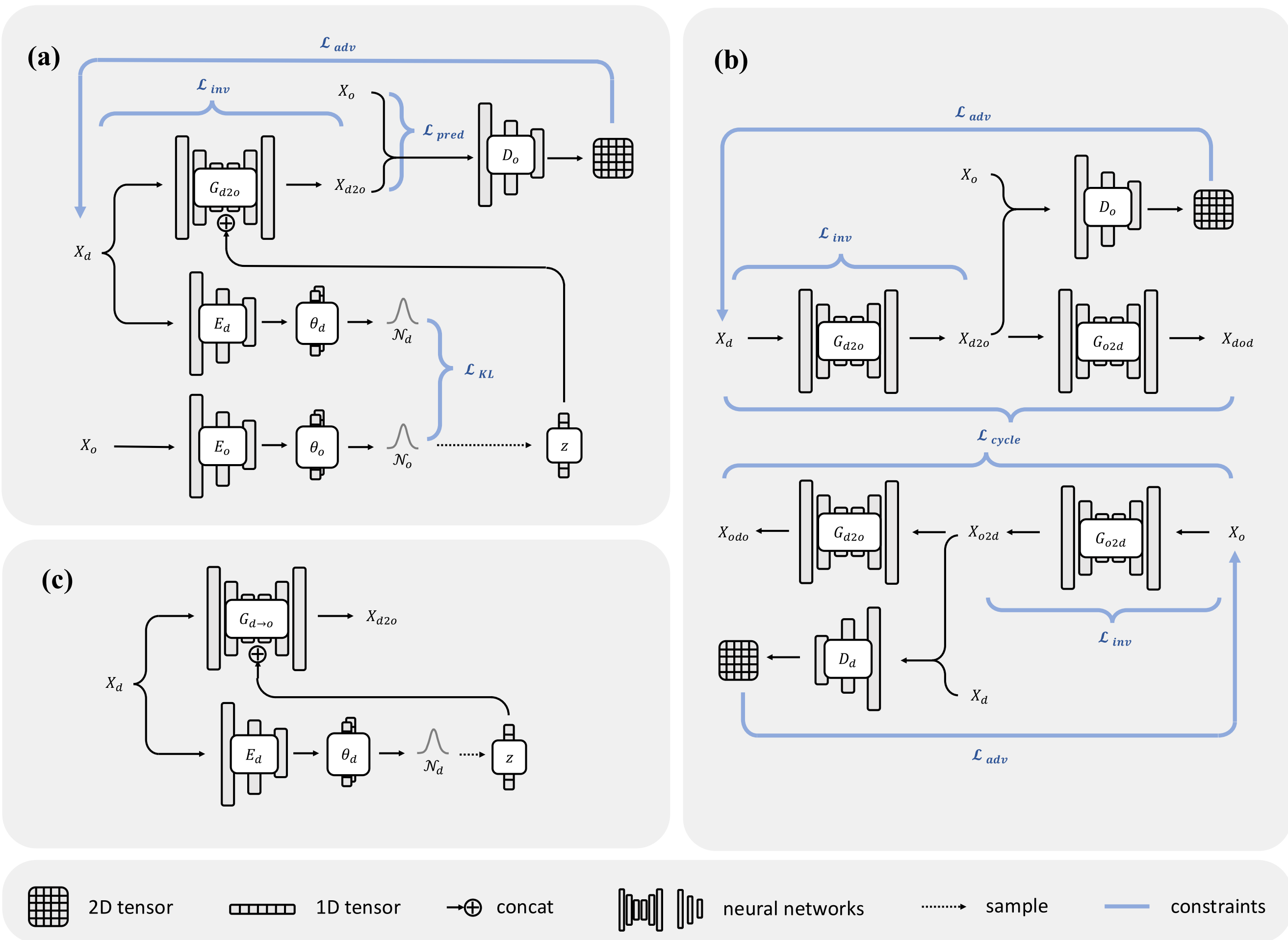

Figure 2. Schematic of the unified architecture and workflows of the proposed GAN-W2C model. The figure conceptually decomposes the framework's constraints based on their physical motivation. (a) The weather-scale inspired components, including a direct prediction loss ($\mathcal{L}_{\mathrm{pred}}$) and a KL-divergence loss ($\mathcal{L}_{\mathrm{KL}}$) for aligning latent spaces, primarily address paired samples. (b) The climate-scale inspired components, featuring a CycleGAN structure with a cycle consistency loss ($\mathcal{L}_{\mathrm{cycle}}$), address the unpaired nature of long-lead forecasts. All samples are trained under a unified objective combining these constraints. (c) The inference process, where a mini-ensemble is generated by sampling latent vectors (z) from the aligned encoder (E_d).

skill, spatial distribution matching, multivariate consistency, the effectiveness of its uncertainty decomposition, and an ablation study. Section 4 provides conclusions and Section 5 discusses the implications of our findings as well as the future research directions.

2. Methods

2.1. Proposed Model

We propose addressing flow-dependent systematic biases while mitigating the interference of assimilation errors through a distribution-guided approach, which helps overcome the limitations inherent in conventional post-processing approaches. Deep generative models currently represent the most powerful tools for distribution modeling, and among them, Generative Adversarial Networks (GANs) effortlessly integrate into existing regression-based frameworks. Through the adversarial interplay of two neural networks, GANs learn an implicit loss function that penalizes potential structural discrepancies between outputs and targets. We integrate variants of GAN models, each dedicated to post-processing either short-lead or long-lead forecasts, into a unified framework (Figure 2).

The proposed framework operates on daily joint states containing both circulation and precipitation fields, and treats each ensemble member independently as an individual sample in the model space.

For weather-scale predictions (short lead times), forecasts and observations can be treated as paired data. A method adapted from the image-to-image translation approach of Henry et al. (2021) is introduced to extend traditional Model Output Statistics (MOS) by incorporating adversarial constraints. A prediction loss ($\mathcal{L}_{\text{pred}}$) supervises the network on paired samples, and an adversarial loss ($\mathcal{L}_{\text{adv}}$) enhances detail refinement and guides the search for the nearest observational state space samples given approximate, blurred post-processed targets (Figure 2a).

To address the inherently one-to-many nature of coarse-to-fine resolution mapping and parameterization uncertainties, we inject perturbations into our workflow to explore the neighborhood around the post-processed results. We use a probabilistic method inspired by variational autoencoders (VAEs, Kingma & Welling, 2014), using a Kullback-Leibler divergence loss ($\mathcal{L}_{\text{KL}}$) to ensure consistency between the perturbations and the outputs (Figures 2a and 2c). As a result of such perturbation, each member of the dynamical model is processed into a compact mini-ensemble.

For predictions beyond the weather scale, forecasts and observations are unpaired because of the chaotic growth of initial state errors. We adopt CycleGAN (Zhu et al., 2017), a method that performs domain adaptation by matching the underlying data distributions. It constructs both forward and inverse transformation functions, encouraging bijective mappings between the two domains through a cycle consistency loss ($\mathcal{L}_{\text{cycle}}$) (Figure 2b). This constraint, combined with the adversarial loss, effectively bridges the full state spaces of dynamical models and observations.

To unify these regimes, the deep learning framework enables seamless integration by tapering the weight of the prediction loss ($\mathcal{L}_{\text{pred}}$), ensuring that the regression constraint receives larger weight at short leads and decays to zero at long leads. In practice, we adopt a quadratic tapering function (Text S1 in Supporting Information S1, Equation 6), though the specific functional form serves as a hyperparameter and should be insensitive to reasonable choices. This simple set ensures that long-lead samples are governed by unpaired constraints while short-lead samples are influenced by the complete set of available constraints.

Furthermore, although post-processing precipitation is the primary goal, the framework still transforms the full state space, including both circulation and precipitation fields. This approach is motivated by established practices in forecasting where skillful large-scale circulation forecasts serve as a reliable "carrier" for weather systems, from which precipitation is diagnosed (Cai et al., 2022; Chen et al., 2012; Pathak et al., 2022; Wilby et al., 1998). Instead of constructing a direct circulation-to-precipitation emulator, our framework leverages the power of generative modeling to implicitly learn the physical relationships between variables.

We introduce an invariance loss ($\mathcal{L}_{\text{inv}}$) to serve as a crucial anchor and encourage the large-scale circulation fields to remain close to the raw dynamical forecast during transformation. This is complemented by the adversarial loss ($\mathcal{L}_{\text{adv}}$), which enforces output products to follow a realistic joint distribution, thereby enhancing physical consistency by promoting reasonable inter-variable correlations. A combination of anchoring and physical consistency enforces that post-processed precipitation is generated conditional on the synoptic situation predicted by the dynamical model, helps to reduce the risk of mode collapse or producing precipitation fields that look realistic but irrelevant to current weather patterns, and further stabilizes adversarial training.

Text S1 in Supporting Information S1 provides the complete mathematical formulation of our framework, detailing the explicit definition and physical interpretation of each loss term, as well as the per-sample dataflow.

2.2. Baseline Models

We establish a series of baseline models and products for comparison. The first baseline is the raw dynamical model output, where its 1° output is bilinearly interpolated onto the 0.1° grid of the reference precipitation product.

For the Model Output Statistics (MOS) baseline, we implement a conventional post processing paradigm by training a correction function on paired samples with MSE loss (corresponding to $\mathcal{L}_{\text{pred}}$ in our formulation). Sensitivity experiments indicate that the first three lead days yield reliable paired samples. While sharing the same G_{d2o} architecture as our proposed model, the MOS baseline model excludes circulation outputs since they impose no effective constraints on precipitation under this training scheme.

We further include Ensemble Model Output Statistics (EMOS) as a statistical ensemble post-processing baseline, implemented using the recent framework of Trotta et al. (2024). The parametric distribution assumption employed in this method is well-suited for the intermittency (zero inflation) and strong skewness of daily precipitation. Transformation parameters are estimated by maximizing the likelihood. We train separate models for each lead time and apply them grid-wise. To ensure a fair and scalable comparison, we draw samples from the fitted analytical distribution at each grid point, matching the total ensemble size of GAN-W2C method and using the same verification pipeline as for all other methods. Since EMOS is not designed to preserve spatial coherence, we do not evaluate the spatial structure of its sampled fields.

The probabilistic climatological forecast is implemented following configurations similar to Yuval et al. (2024). For each target timestamp, we select all historical samples from other years within ±15 days of the target's day-of-year, then randomly draw 30 members from this climatological pool to form the baseline ensemble.

2.3. Data

This study employs reforecasts of dynamical models, reanalysis, and gridded precipitation products. The study domain covers most of Southern China (21°N−33°N, 106°E−122°E), a region predominantly influenced by the East Asian monsoon. Among various models in the S2S project (Vitart et al., 2017), the ECMWF IFS CY47R3 model is selected for experimental post-processing. This model provides extended-range forecasts with lead times of 1 day up to 46 days. On-the-fly reforecast data of its 2022 operational version are retrieved at 1° horizontal resolution from its publicly available data set, covering the period from 2002 to 2021.

ERA5, the fifth-generation global climate and weather reanalysis (Hersbach et al., 2020) at 0.25° horizontal resolution, is used as the observational circulation reference. The China Meteorological Forcing Dataset (CMFD, He et al., 2020) is selected as the reference gridded precipitation data, which provides 0.1° horizontal resolution coverage of China's land region from 1979 to 2018. Static topographic information consists of bedrock elevation from ETOPO 2022 at 0.1° horizontal resolution and sub-grid scale orography characteristics at 0.25° horizontal resolution retrieved from ERA5, including angle, anisotropy, slope, and standard deviation.

Based on prior knowledge of synoptic systems and data availability, the selected circulation variables are: 2 m temperature, mean sea level pressure, and circulation fields at multiple pressure levels (zonal/meridional winds, temperature, specific humidity, geopotential height at 500/700/850 hPa, and 200 hPa zonal winds and geopotential height). These variables represent key atmospheric features including synoptic pressure systems, moisture transport, the South Asian High, and jet streams.

The overlapping period of multi-source data is partitioned into training (year 2002–2013), validation (year 2014–2016), and testing (year 2017–2018) sets, with all scores, distributions, and case evaluations conducted on the test set. During the training phase, loss terms are calculated on the naive resolutions of respective data sets. Precipitation is primarily assessed on the 0.1° CMFD grid.

2.4. Preprocessing and Training Configurations

All non-static variables are averaged into daily statistics. A minimax scaling to [0,1] is applied to circulation variables, static terrain parameters, and coordinates. For precipitation, a transformation of $x' = 5\ln(x+1)$ is adopted to alleviate the long-tailed characteristics, preserving extreme values while normalizing magnitudes to $\sim\mathcal{O}(1)$.

All neural networks in our framework, including generators G, discriminators D, and encoders E, share a convolution-based backbone and condition on auxiliary inputs such as month, geographical coordinates, and topographic features. Notably, as evident from our above formalization, all networks are designed to be independent of the dynamical model's lead-time. The network architectures are fundamentally composed of UNet-like convolutional downsampling and upsampling modules together with ResNet blocks, following configurations like Hess et al. (2022), which also represents a standard practice in deep learning applications for meteorological data.

A PatchGAN strategy is adopted, whereby the discriminator D outputs a two-dimensional map of realism scores, ensuring that each local region receives an explicit adversarial constraint for enhanced spatial detail. To stabilize the notoriously challenging adversarial training process, we utilize the Wasserstein GAN with gradient penalty

10.1029/2025JH000993

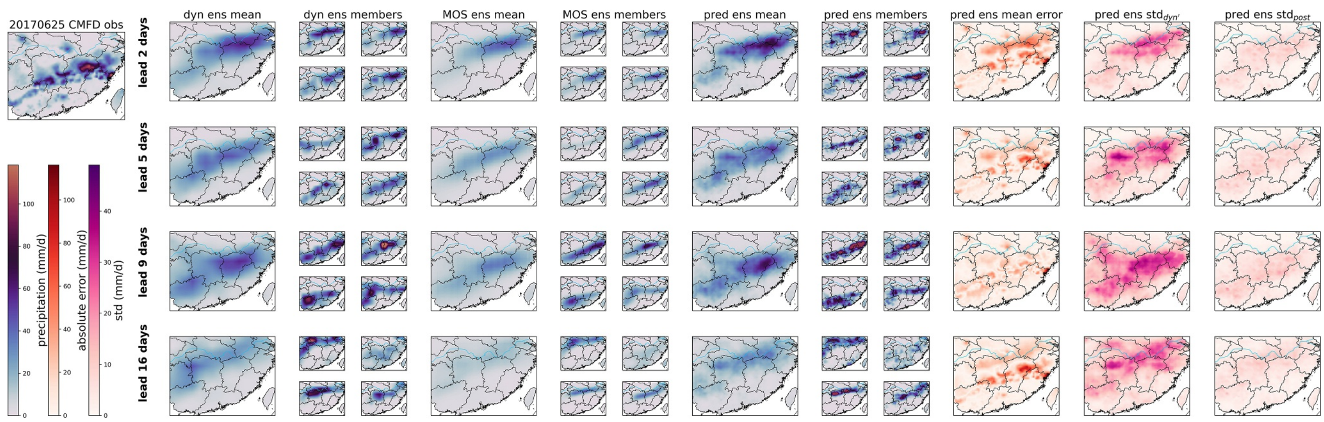

Figure 3. Comparison of precipitation forecasts for a representative case across multiple lead times (rows). The figure contrasts the CMFD observation with forecasts from the raw dynamical model, the MOS baseline, and the proposed GAN-W2C framework. Each forecast product includes the ensemble mean and four randomly sampled members. For the GAN-W2C model, the corresponding absolute prediction error and its decomposed uncertainty components (std$_{dyn'}$ and std$_{post}$) are also shown.

(WGAN-GP) variant this approach replaces the outputs of the discriminator D from binary classification to a continuous distance metric, reducing training oscillations. Additional constraints on gradient magnitudes of D mitigates the risk of gradient explosion or vanishing effectively removing the need for frequent manual intervention on hyperparameters during GAN training, as detailed in Harris et al. (2022). All networks are trained using the Adam optimizer with optimal network parameters selected based on both quantitative performance on the validation set and qualitative visual assessment of the inferenced sample distributions.

For the training of our proposed model, we increase the sampling frequency of samples with lead time $\leq$3 days by a factor of three to balance the contribution between short-lead and long-lead samples. The size of raw dynamical forecasts M and post-processed mini ensemble N is 11 and 10, respectively.

Detailed configurations, hyperparameters, and computational resource requirements are provided in Table S1 and Text S5 of Supporting Information S1.

3. Results

3.1. Representative Forecast Examples

To illustrate the qualitative differences in spatial representation that are not fully captured by spatial aggregated metrics, we present a representative forecast case featuring a heavy precipitation event, typical for Southern China during summer (Figure 3). In general, both methods correct systematic biases without drastically altering the position of precipitation bands from the dynamical forecasts. The MOS post-processing exhibits a notable blurring effect, tending to predict moderate precipitation across the domain. This approach underestimates spatial heterogeneity and extremes while overestimating the spatial extent of light precipitation. Consequently, for a lead time extending to approximately 2 weeks, the MOS ensemble mean has shown implausibly weakened precipitation. In contrast, our GAN-W2C framework preserves reasonable spatial details and extreme precipitation magnitudes. Some ensemble members successfully provide early warning signals for heavy precipitation events even at extended lead times.

An additional benefit of GAN-W2C approach is the ability to quantify and decompose the uncertainty of forecasts. std$_{dyn'}$ denotes the dispersion between post-processed products from different dynamical ensemble members, representing uncertainty inherent to the dynamical integration. Meanwhile, std$_{post}$ captures the average internal dispersion within the small ensemble generated from each single dynamical member, quantifying uncertainty introduced by the post-processing model itself. Detailed calculations are in Text S4 of Supporting Information S1. These uncertainties exhibit significant distinctions in their spatial patterns, while collectively estimating prediction errors in advance by highlighting spatial locations likely to have larger errors. Further quantitative analysis of these uncertainty components will be presented in Section 3.5.

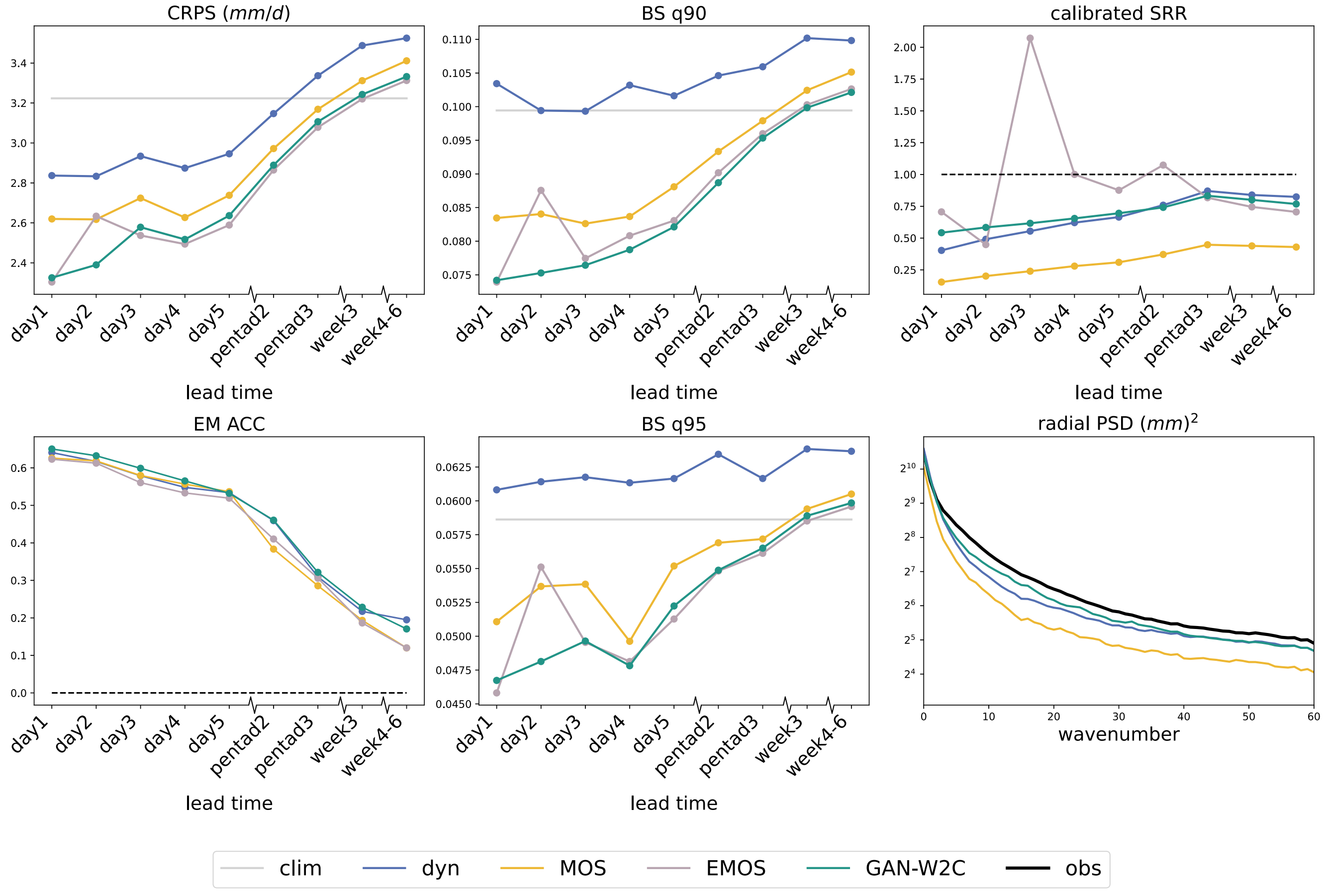

Figure 4. Comparison of verification scores for the proposed GAN-W2C model against baseline methods: raw dynamical forecast (dyn), MOS and EMOS post-processing, and climatology (clim). Panels show key metrics for probabilistic skill (CRPS), extreme event prediction (BS q90, BS q95), deterministic skill (ensemble mean ACC), reliability (calibrated SRR), and spatial structure (radial PSD). To clarify the visualization of skill evolution, scores are presented for daily lead times (day 1–5) and as averages for subsequent pentad and weekly intervals. EMOS is excluded from the PSD panel considering its designing purpose.

3.2. Forecasting Skill Assessment

We employ a key set of verification metrics to evaluate the ensemble forecasts, with all scores evaluated at the same grid as observational data (Figure 4). The probabilistic forecast skill is assessed using the CRPS and Brier Score (BS) for extreme precipitation events, while the deterministic forecast skill is evaluated using the Anomaly Correlation Coefficient (ACC) of the ensemble mean. The overall ensemble reliability is quantified through the Spread-RMSE Ratio (SRR, RMSE denotes Root Mean Square Error). Spatial detail is evaluated using radial spectra. The detailed calculation of these scores is provided in Text S2 of Supporting Information S1. Specific procedures, including the variable-window evaluation for ACC (Pan et al., 2019; Zhu et al., 2014), ensemble size correction for SRR (Leutbecher & Palmer, 2008), and a smoothing procedure to obtain a more stable annual cycle of the reference statistics (Rasp et al., 2024), are also detailed therein.

As expected, forecast skill for both the dynamical model and post-processed products decays with increasing lead time, eventually converging toward or even becoming worse than climatological performance. For CRPS skill, the GAN-W2C products outperform both the dynamical products and MOS baselines across all lead times, with the mean score reduced to 92.9% relative to the dynamical model. According to this metric, the GAN-W2C model also extends the forecast timescale beyond climatology to approximately week 3, compared to around pentad 3 for MOS and only pentad 2-3 for the dynamical model. Similar improvements are observed for extreme event prediction using the Brier score; for 90th and 95th percentile thresholds, the GAN-W2C model reduces the BS to 89.4% and 91.0% relative to the dynamical model, respectively. While the raw dynamical model often underperforms climatology for extremes across virtually all lead times due to systematic biases, all post-processing methods provide skillful extreme forecasts relative to climatology throughout the first 2 weeks, with GAN-W2C showing consistently strong performance. The EMOS baseline achieves marginal advantages in CRPS and, for limited lead times, in Brier scores. In terms of variable-window ACC scores, the GAN-W2C products maintain performance comparable to or slightly better than the dynamical model across most lead times, though performance remains limited at extended forecasting horizons (weeks 4–6). In contrast, both MOS and EMOS baselines show degradation from dynamical forecasts at most lead-times. Regarding ensemble reliability, SRR results indicate that dynamical products are under-dispersive, especially at short lead times, and MOS post-processing largely exacerbates this under-dispersion across all lead times (SRR further from 1). Meanwhile, the GAN-W2C model effectively calibrates the ensemble in the first 7 days (improving SRR by 22.5%). Although it increases under dispersion at longer lead times (reducing SRR to 93.8% of the dynamical product's score), its performance remains substantially better than the MOS calibration. The EMOS baseline exhibits unstable calibration behavior with short-lead overspread. This is likely due to its grid-wise nature, which can generate extreme values at individual grids without spatial constraints. These improvements are robust across seasons, with the largest magnitude in spring and summer when precipitation errors are generally larger (Figure S2 in Supporting Information S1). Spatially, the probabilistic skill gains are widespread across the domain, whereas Brier Score reductions are strongest over inland western South China, consistent with the larger baseline errors in that region (Figure S3 in Supporting Information S1). Finally, assessing the representation of spatial details via power spectra shows that the dynamical products, constrained by coarser spatial resolution, lose power at wavenumbers higher than 3 (scales finer than ~500 km). MOS post-processing intensifies this limitation with spectral power underestimation across all wavenumbers. The GAN-W2C products, however, have a power spectrum more closely aligned with observations than these two baselines, particularly enhancing power at wavenumbers lower than 40 (scales coarser than *sim* 35 km).

We further extended the evaluation to confirm the robustness of these findings. On the 1° grid (with EMOS refit at 1°), GAN-W2C leads across nearly all metrics, notably including ACC scores of ensemble mean that consistently exceed the raw dynamical model (Figure S4 in Supporting Information S1). For 3-day accumulated precipitation, GAN-W2C maintains its advantage in probabilistic scores, with only ACC lower than the raw dynamical model but still outperforming other post-processing methods (Figure S5 in Supporting Information S1). These results indicates that the model remains effective at coarser spatial scales and demonstrates potential in capturing temporal correlations. Sensitivity experiments on ensemble size (Figure S6 in Supporting Information S1) indicate that flow-dependent uncertainty from the dynamical ensemble primarily span the informative spread of the probabilistic post-processed products, and the performance gains in our default configuration mainly stem from distribution-guided post-processing rather than simple ensemble inflation, although additional members can still provide marginal benefits when computational resources allow.

3.3. Ensemble Calibration

To diagnose ensemble calibration more clearly, we further evaluate rank histograms for precipitation intensity and reliability diagrams for the binary occurrence of precipitation, calculation procedures also detailed in Text S2 of Supporting Information S1. Figure 5 presents these results for daily fields on the 0.1° grid.

The Rank histograms reveal that the raw dynamical forecasts and the MOS baseline display the U-shape indicative of ensemble under dispersion, and this behavior intensifies at longer lead times. Both GAN-W2C and EMOS flatten these histograms, indicating improved calibration of precipitation intensity. Quantitatively, the rank distributions from these experiments have means closer to 0.5 and smaller variances, reflecting reduced dispersion from the U-shaped pattern. At extended leads (week 3–6), a slight positive skew emerges in the calibrated experiments, suggesting over forecasting bias under limited predictability and is consistent with the SRR decline observed at longer leads in Section 3.2.

The reliability diagrams show that the raw dynamical forecasts are overconfident (lower slope) in predicting precipitation occurrence, particularly at short leads where they predict too many wet days, partly attributable to the coarse spatial resolution of the raw data. The MOS baseline primarily reduces mean bias but does not improve the slope of the reliability curve. EMOS is closest to the perfect calibration, as expected from its explicit training procedure. GAN-W2C also demonstrates competitive calibration performance by correcting the slope bias, particularly showing more reliable predictions at higher forecast probabilities corresponding to more significant precipitation events.

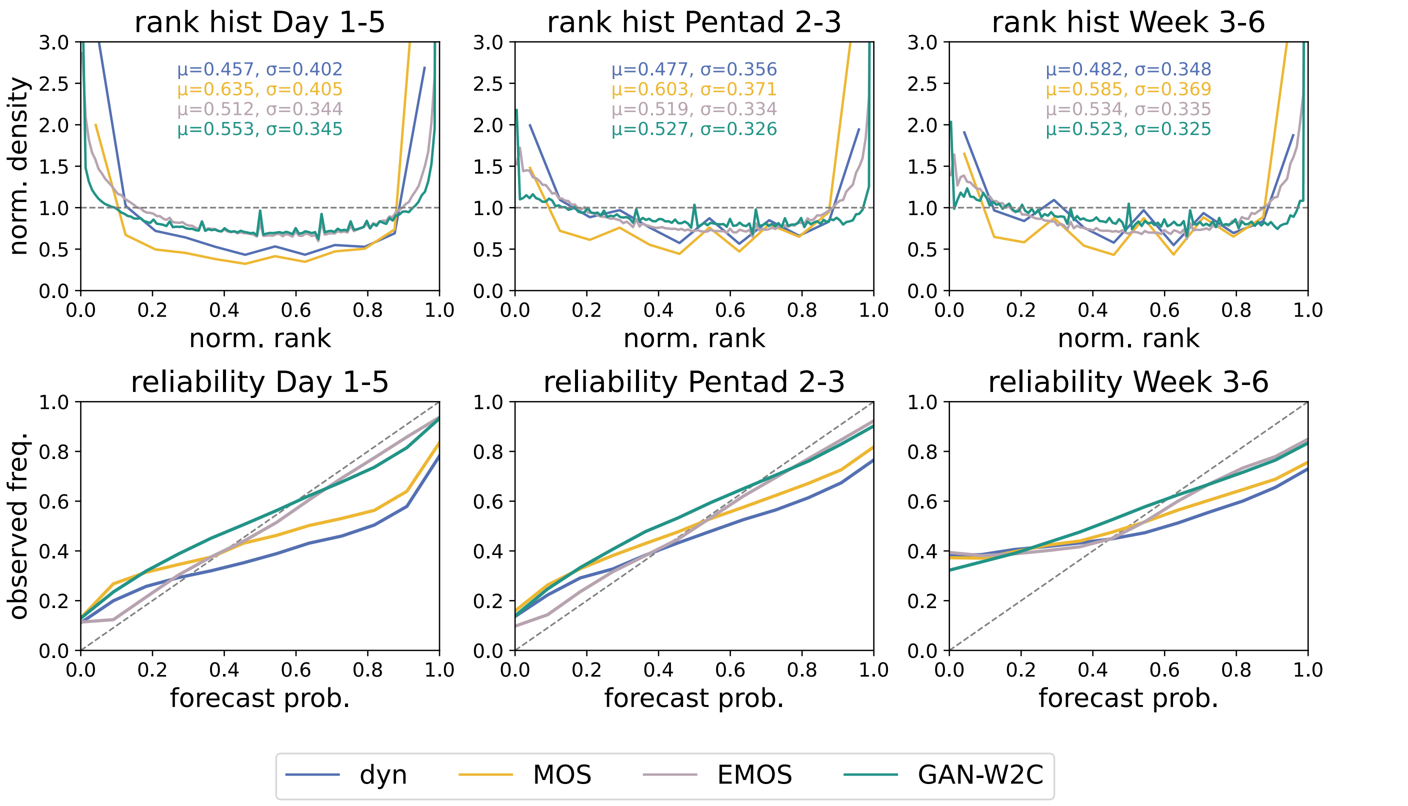

Figure 5. Ensemble calibration diagnostics for daily precipitation forecasts on the 0.1° grid. Top row: rank histograms conditioned on observed wet days (colored lines connect histogram bin tops; annotations show rank mean and variance; text annotations indicate the mean and variance of the rank distribution; dashed line indicates the uniform reference distribution). Bottom row: reliability diagrams for precipitation occurrence (curves show observed frequency vs. forecast probability; dashed line indicates perfect reliability) Columns summarize performance at day 1–5, pentads 2–3, and weeks 3–6.

These results confirm that the probabilistic skill gains of GAN-W2C are not only due to mean adjustment but also reflect recalibration of the intensity and occurrence distribution. Evaluations on the 1° daily grid (Figure S7 in Supporting Information S1) and for 3-day accumulated precipitation (Figure S8 in Supporting Information S1) yield consistent conclusions, with more distinct advantage over baselines at long lead times in rank histogram performance.

3.4. Climatology Distribution Matching

To evaluate the model's ability to reproduce the overall precipitation distribution, we calculate grid-wise statistics: the proportion of PD (precipitation days, defined as precipitation >0.25 mm/d, following Rasp et al. (2024)), as well as the mean, standard deviation, and the 33rd, 66th, 90th, and 99th percentiles of precipitation on PD (Figure 6). The analysis encompasses first- and second-order moments, as higher-order moments exhibit artifacts in the observational data sets, and three parameters sufficiently characterize precipitation probability forecasts (Vandal et al., 2018). Spatial correlations (r) and root mean square errors (RMSE) compared to observational statistics are also evaluated.

As is shown in Figure 6, the raw dynamical model outputs exhibit considerable and grid-dependent systematic biases compared to observations. These include overestimation of precipitation day proportion, mean values, and most percentiles (33rd, 66th, 90th), coupled with the underestimation of precipitation variability and upper extremes (standard deviation and 99th percentile), particularly over regions such as the lower Yangtze River basin and the Pearl River Delta. Importantly, the dynamical products fail to capture the spatial heterogeneity across all evaluated statistics. While the MOS post-processing reduces the bias in precipitation day proportion, it introduces a substantial underestimation across all other statistics, often exceeding the biases of the raw dynamical output. Moreover, MOS further flattens spatial contrasts across all statistics, which is also a manifestation of the characteristic blurring effect. In contrast, the GAN-W2C approach, despite still underestimating precipitation day

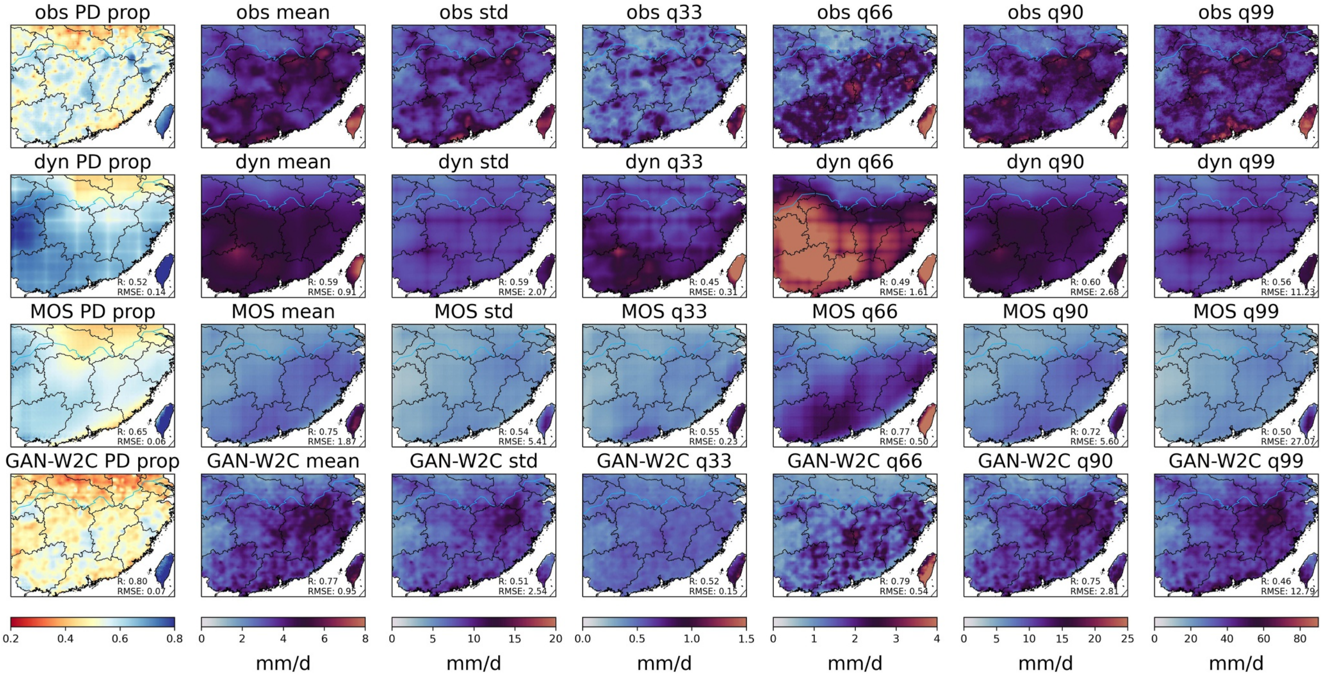

Figure 6. Grid-wise comparison of precipitation statistics from observations (obs) and different forecast products. The panels show the proportion of precipitation days (>0.25 mm/d, PD prop), along with the mean, standard deviation (std), and various percentiles (q33–q99) calculated on precipitation days, annotated with the spatial correlation coefficient (r) and root mean square error (RMSE) against the corresponding observational statistic.

proportion and some statistics (especially extremes) in certain areas, successfully mitigates the systematic underestimation characteristic of MOS post-processing. A key improvement is its ability to preserve spatial heterogeneity in these statistics, effectively capturing features like the extreme precipitation characteristics observed in the middle and lower reaches of the Yangtze River.

3.5. Multivariate Consistency

Multivariate consistency is examined through empirical orthogonal function (EOF) analysis applied to anomaly fields aggregated to a 1° resolution, considering both the focus on large-scale spatial patterns and computational costs. The leading EOF modes capture the dominant patterns of multivariate covariation. Physical coherence within the leading modes serves as a necessary condition for physically consistent results. Both the dynamical model forecasts and GAN-W2C post-processed fields consist of multivariate fields containing circulation and precipitation. In particular, for the MOS products, which only provide precipitation output, the consistency analysis is performed on a joint state that concatenates the post-processed precipitation with the original dynamical model circulation fields, expressed as $({}^{c}X_{d}^{\tau}, {}^{p}X_{d2o}^{\tau})$. Further details of the EOF analysis, including the standardization procedure and mode alignment, are provided in Text S3 of Supporting Information S1.

Here, we specifically examine the relationships between precipitation anomalies and low-level (850 hPa) moisture transport conditions (Figure 7). While the derived modes of all products exhibit varying degrees of deviations from the observed patterns due to residual biases and limited sample size of the test set, the critical focus is on identifying any physically inconsistent covariant features between variables. In the dynamical model, mode 4 shows the largest discrepancy, with positive precipitation anomalies displaced toward the southwest of the domain. This shift is physically consistent with a weakened westerly anomaly in the same region that channels additional oceanic moisture, producing a co-located positive moisture anomaly. Similarly, the leading modes of GAN-W2C that differ from observations still demonstrate internal consistency. In its mode 2, intensified positive precipitation anomalies in the southwest of the domain coincide with enhanced low-level convergence. Likewise, mode 3 exhibits a negative precipitation anomaly center concentrated in the northeastern region, corresponding to enhanced local divergence. Mode 4 displays negative precipitation anomalies extending into the central domain, positioned further north than observed, which aligns with the northward-shifted negative moisture anomalies. In

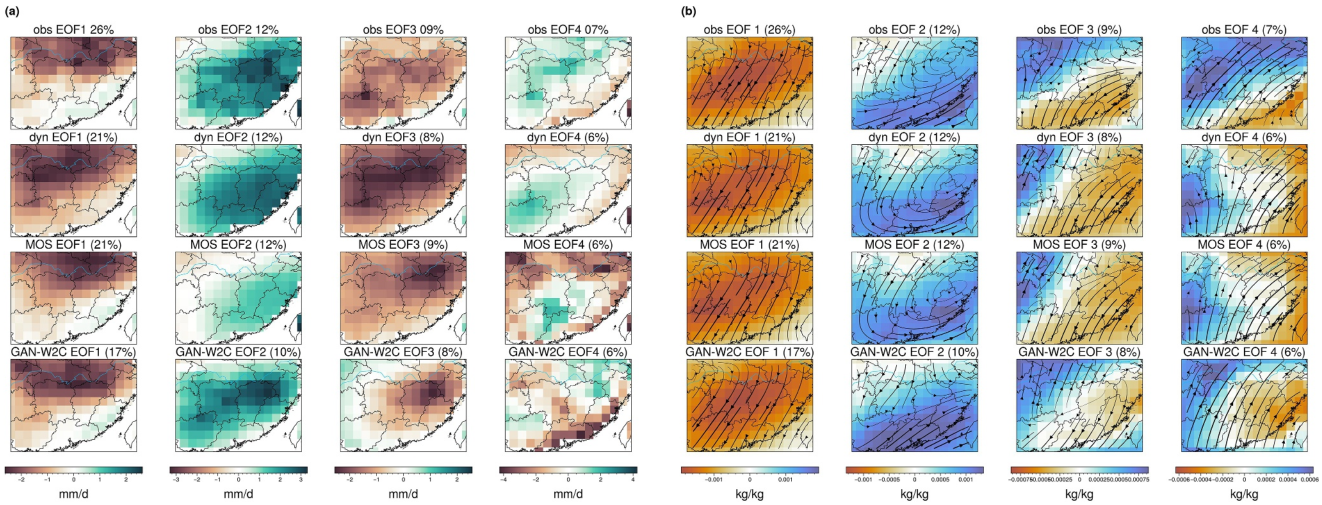

Figure 7. Leading modes of a joint EOF analysis comparing the multivariate structure of observations and different forecast products. This figure shows the dominant patterns of co-variation between (a) daily precipitation anomalies and (b) 850 hPa anomalies of specific humidity (shading) and wind (streamlines). The percentage of explained variance is labeled for each product and mode.

contrast, the MOS modes demonstrate apparent inconsistencies. While modes 3 and 4 retain circulation conditions similar to the dynamical modes, their corresponding precipitation anomaly patterns differ substantially. This is most evident in mode 4, where the moisture transport conditions favor precipitation in the southwest, yet the positive precipitation anomaly is incorrectly positioned in the central domain and precipitation is suppressed across most of the region.

While the EOF analysis demonstrates that GAN-W2C maintains internal physical consistency between precipitation and circulation within its generated fields, this does not necessarily imply that the precipitation improvements are built upon improved circulation forecasts. By design, the generated precipitation fields are anchored to the large-scale circulation patterns predicted by the dynamical model, and we do not expect the framework to correct fundamental circulation errors. Figures S9 and S10 in Supporting Information S1 provides representative cases illustrating this behavior. When the large-scale circulation forecast is relatively accurate, GAN-W2C preserves this background and applies minor circulation adjustments to refine the precipitation. When the raw forecast is misleading, GAN-W2C preserves the deviated large-scale pattern rather than attempting to guess an alternative circulation. This case-dependent performance confirms that the model follows physical constraints rather than making arbitrary changes, indicating that its post-processing skill is bounded by the predictability of the large-scale circulation provided by the dynamical model.

3.6. Uncertainty and Its Decomposition

The decomposed uncertainty components introduced in Figure 3 are analyzed across the entire test set to reveal characteristics in their lead-time dependence and error prediction capabilities. As shown in Figure 8a, the propagated dynamical uncertainty, $\mathrm{std}_{\mathrm{dyn}'}$, increases with lead time and saturates around pentad 3, while $\mathrm{std}_{\mathrm{post}}$ remains largely constant across lead times, showing only slightly higher values in the first 2 days (3.6% higher than subsequent days). This behavior is intuitively reasonable, as $\mathrm{std}_{\mathrm{dyn}'}$ inherits the dynamical model's estimate of chaotic divergence, while $\mathrm{std}_{\mathrm{post}}$ originates from sources such as downscaling and precipitation parameterization, which are inherently lead-time independent. Beyond day 3, the total forecast uncertainty $\mathrm{std}_{\mathrm{all}}$ becomes dominated by $\mathrm{std}_{\mathrm{dyn}'}$.

We further assess whether the estimated uncertainties can flag potential "hot spots" of forecast failure, as shown in Figure 8b. Since aggregated metrics for probabilistic skill, such as CRPS, are insensitive to case-dependent variations in ensemble distribution (Feng et al., 2024), a direct evaluation of the pointwise relationship between the ensemble spread and forecast error (RMSE) is needed. This evaluation employs max pooling over 10 grid points, examining maximum uncertainty and maximum error within 1° ranges. Both quantities have been

10.1029/2025JH000993

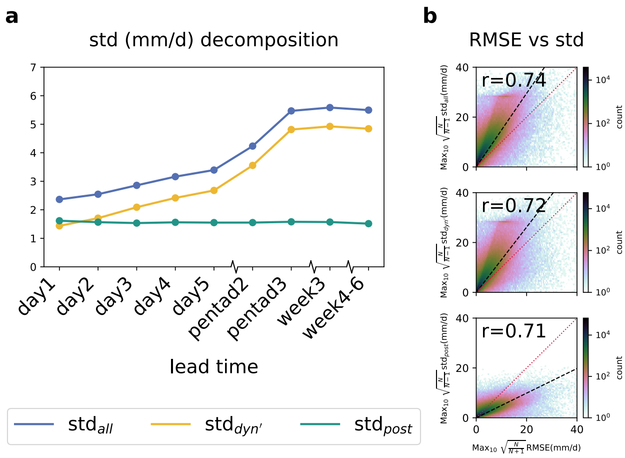

Figure 8. Uncertainty decomposition and spread-error correlation for the GAN-W2C model. (a) Domain-averaged magnitude of the total ($\mathrm{std}_{\mathrm{all}}$), propagated dynamical ($\mathrm{std}_{\mathrm{dyn}'}$), and post-processing ($\mathrm{std}_{\mathrm{post}}$) uncertainty components as a function of lead time. (b) Scatter density plots of local maximum forecast error (RMSE) versus local maximum uncertainty (std) for the three uncertainty components, with the Pearson correlation coefficient (r) labeled in each panel. The black dashed line indicates the 1:1 correspondence for perfect calibration, while the red dotted line shows the intercept-free least squares regression fit.

calibrated to be consistent with the SRR treatment detailed in Text S2 of Supporting Information S1. Although systematic over- or under-estimation exists, the results reveal a strong pointwise correlation ($r > 0.70$) for both total and decomposed uncertainties. While this correlation is, at least partially, attributed to the shared dependence of both spread and error on the magnitude of the forecast anomaly (Feng et al., 2024), it nonetheless confirms that the estimated uncertainty provides the practically valuable, flow-dependent indicator of local forecast error.

Notably, the error prediction capability of propagated dynamical uncertainty $\mathrm{std}_{\mathrm{dyn}'}$ closely matches that of the raw dynamical model's intrinsic uncertainty $\mathrm{std}_{\mathrm{dyn}}$ (both $r \approx 0.72$, see also Figure S12 in Supporting Information S1), indicating that our post-processing framework preserves the uncertainty estimation information inherent in the dynamical model. The total uncertainty $\mathrm{std}_{\mathrm{all}}$ achieves an even stronger correlation with the error ($r \approx 0.74$), indicating that the error prediction capability of our framework builds upon and exceeds that of the dynamical model by effectively combining both sources of uncertainty. In contrast, uncertainties estimated from the MOS post-processed ensemble show much weaker correlation with RMSE ($r \approx 0.57$, see Figure S12 in Supporting Information S1), and the corresponding scatter plot also reveals a distinct branch of systematic underestimation.

3.7. Ablation Analysis

Ablation analyses are conducted to assess the necessity of the weather-scale and climate-scale constraints by evaluating their independent performance. The weather-scale model (GAN-W), which conceptually isolates the workflow shown in Figure 2a, involves $\mathcal{L}_{\mathrm{pred}}$, $\mathcal{L}_{\mathrm{inv}}$, $\mathcal{L}_{\mathrm{adv}}$ and $\mathcal{L}_{\mathrm{KL}}$ as its loss functions, and excludes the inverse transformation G_{o2d}. The climate-scale model (GAN-C), which similarly isolates the workflow of Figure 2b, employs only $\mathcal{L}_{\mathrm{inv}}$, $\mathcal{L}_{\mathrm{cycle}}$, and $\mathcal{L}_{\mathrm{adv}}$, with G_{d2o} configured deterministically. This design choice stems from our practical finding that introducing uncertainty in pure CycleGAN frameworks brings little score improvement but markedly increases the training cost. Detailed training configurations for both ablation models are provided in Table S1 of Supporting Information S1.

Quantitative evaluation Figure 9 reveals that both the GAN-W and GAN-C models underperform the full GAN-W2C framework across all lead times in probabilistic (CRPS), deterministic (ACC), and extreme event (Brier score) metrics. For the overall distribution (Figure S11 in Supporting Information S1), GAN-W post-processing products effectively capture spatial heterogeneity in grid-wise statistics but exhibit a pronounced underestimation of precipitation intensity. This bias is also manifested through reduced power spectral density across nearly all wavenumbers (Figure 9) and consistently smaller ensemble spread (Figure S13 in Supporting Information S1) compared to GAN-W2C. The spread-error correlation is weakened for both total and decomposed uncertainty terms. The GAN-C model achieves comparable performance to GAN-W2C in matching grid-wise statistics and

10.1029/2025JH000993

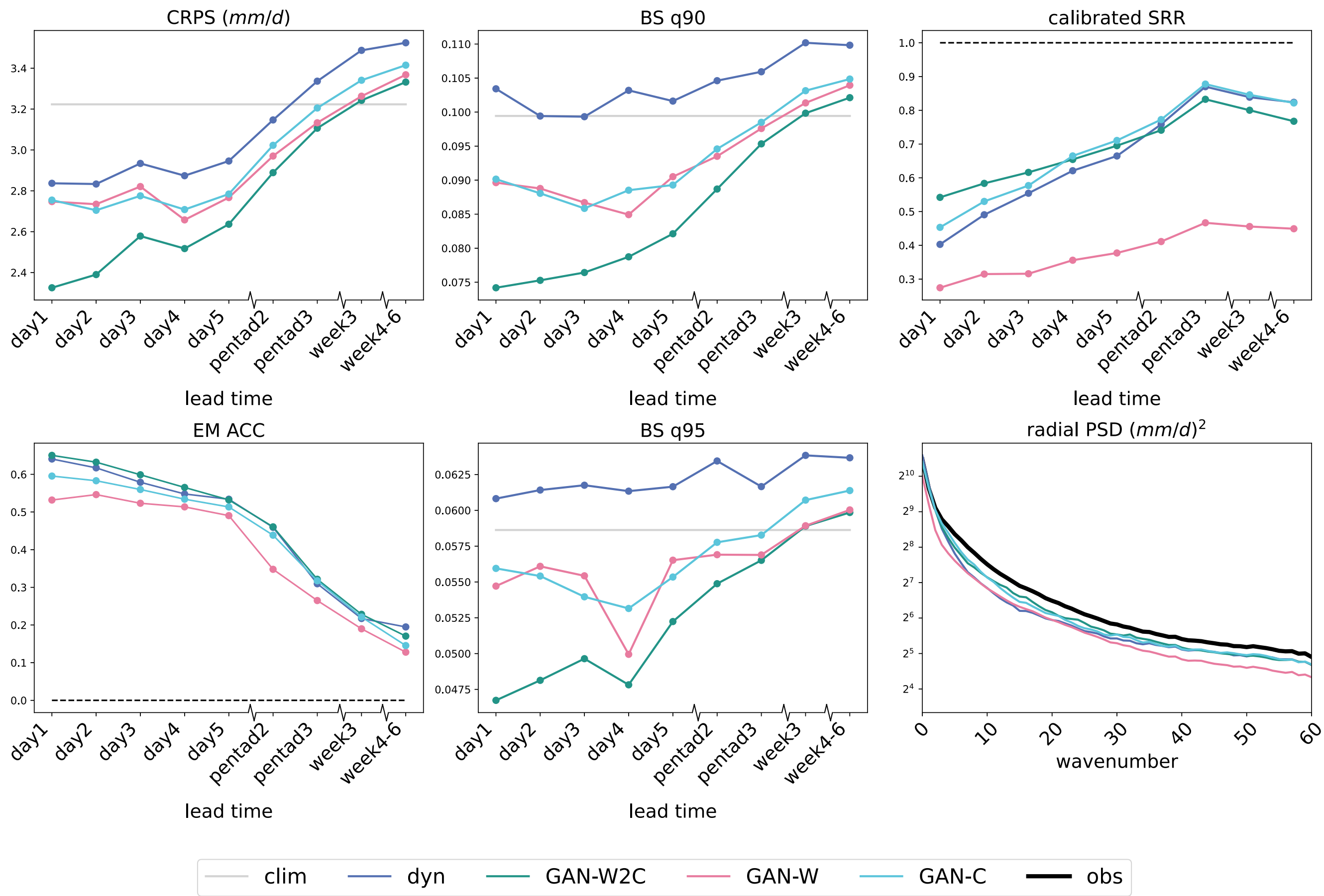

Figure 9. Same as Figure 4, but for the ablation study. The comparison is among models trained with only weather-scale constraints (GAN-W), only climate-scale constraints (GAN-C), and the raw dynamical forecast (dyn).

power spectra with observational distribution. and it calibrates the ensemble spread by mitigating the under-dispersion of the raw dynamical forecasts for at least the first 3 weeks. However, its spread-error correlation remains only comparable to the raw dynamical forecasts (Figure S12 in Supporting Information S1). These findings indicate that removing either the weather-scale or climate-scale constraint yields a model that, while capable of producing well-aligned marginal distributions, fails to deliver the improved ensemble reliability or provide the meaningful uncertainty quantification required for practical forecasts.

4. Conclusions

Dynamical S2S precipitation forecast products require correction of their flow-dependent biases. There is growing demand for products with high spatiotemporal resolution that maintain physical consistency, represent extreme events effectively, and provide informative uncertainty quantification. A fundamental and particular difficulty for developing post-processing models at the S2S timescale is the growing mismatching between forecasts and observations as lead time increases due to the chaos effect. This challenge has been partially masked in existing post-processing applications that target coarser temporal resolutions (pentad to biweekly) or focus on limited stations. However, it becomes unavoidable when developing daily gridded products.

In this study, we propose GAN-W2C, a seamless generative post-processing framework that unifies weather-scale (regression-based) and climate-scale (distribution-matching) constraints, enabling modeling at a daily, 10 km spatiotemporal resolution.

Our key findings reveal the limitations of simpler approaches (the MOS baseline) and the advantages of our proposed framework. While an improved version of the conventional regression-based MOS model, trained on short-lead samples, does improve some metrics across all leads - particularly achieving optimal RMSE and MAE

10.1029/2025JH000993

(Figure S1 in Supporting Information S1)—it suffers from characteristic blurring effects that further result in misrepresentation of extreme events and ensemble under-dispersion. This approach also disrupts the physical consistency and ensemble information inherent in the raw dynamical forecasts, impairing the predictive value of deterministic and probabilistic products.

In contrast, our GAN-W2C model enhances probabilistic (CRPS) and extreme event (Brier score) forecast skill beyond all baselines without degrading the deterministic skill of the raw dynamical forecast, effectively extending the skillful forecast horizon from a probabilistic perspective. Preliminary evaluations on spatiotemporal subsets suggest larger improvements in spatiotemporal subsets when and where dynamical errors are more pronounced. The method ensures more realistic spatial detail (power spectra) and better-aligned overall climatological distributions (grid-wise statistics), alleviating the mean-biased smoothing typical of regression approaches. By fitting the multivariate joint distribution, the framework constrains physical consistency between precipitation and circulation fields, as evidenced by the coherent moisture transport configurations in the leading EOF modes. Importantly, our model produces uncertainty-aware products that properly propagate uncertainty from the dynamical ensemble and decompose it from the uncertainty inherent to the post-processing model, thus providing a valuable advance estimate of forecast error.

5. Discussions

The seamless integration of constraints proves crucial to our post-processing approach. Our framework builds upon existing generative methods that have individually proven effective for their respective domains. For instance, methods combining direct regression with adversarial constraints to learning mappings between paired fields have been successfully employed in short-term weather forecast post-processing (Harris et al., 2022), while CycleGAN has demonstrated success in post-processing decadal climate projections or seasonal forecasts (Hess et al., 2023; Pan et al., 2021; Yang, Ling, et al., 2025). However, as our ablation experiments confirm, models based purely on weather-scale (GAN-W) or climate-scale (GAN-C) constraints fail to achieve the comprehensive performance of the full GAN-W2C framework when applied to the unique challenges of post-processing S2S forecasts. We adopt a unified optimization objective across both weather and climate regimes by tapering the regression constraint as lead time grows. This design explicitly acknowledges and mitigates the issue of forecast-observation mismatch arising from chaotic divergence and enables training on data from all lead times. Treating each ensemble member as an independent sample further expands the training data set, making it feasible to leverage powerful deep generative models. In practical terms, this unified framework is designed to be directly applicable to seamless forecast systems and meets the growing operational demand for consistent S2S products across all forecast horizons.

The generative nature of our approach is fundamental to its success. Most existing data-driven methods are regression-based. They learn to predict the conditional expectation of a future state and therefore suffer the well-known double-penalty effect when predictability is limited, inevitably leading to blurred outputs. Deep generative models, by contrast, learn to represent the complex, high-dimensional probability distribution and make predictions by sampling from a conditional distribution, so individual samples remain sharp, thereby alleviating this issue. Interestingly, our post-processing framework is "ensemble-agnostic"—it operates member-by-member both during training and inference without explicitly taking the entire ensemble as input to calibrate spread. Without using a probabilistic loss function (e.g., CRPS) for optimization, it instead optimizes deterministic scores for individual members, yet implicitly achieves ensemble calibration. This ensemble-agnostic framework brings additional benefit of computational flexibility by reducing the memory requirements because the full ensemble does not need to be processed simultaneously. This suggests that distribution-guided generative approaches are naturally suited for ensemble forecasting problems.

As a representative of traditional ensemble post-processing, EMOS achieves competitive CRPS and Brier score performance. However, its parametric assumptions and grid-wise design have inherent limitations, leading to unstable distribution fitting and a lack of spatial constraints, which restrict its applicability beyond threshold-based probability products. Our framework offers more informative products with broader applicability compared to methods optimized for a single class of metrics.

Temporal generalization from hindcast to operational forecasts is supported by our experimental design. The independent performance on test sets confirms that the model can effectively process unseen data from future dates. Once trained on hindcasts, the weights are directly applicable for near-real-time forecasting, with inference being computationally lightweight (processing a complete forecast initialization in approximately 10 s on a consumer-grade GPU). However, it should be noted that major updates to the dynamical model version would necessitate retraining the post-processing model to capture the altered error characteristics.

Several limitations remain. First, from the perspective of inference, the post-processing skill for precipitation is fundamentally bounded by the skill of the large-scale circulation provided by the dynamical model. From the perspective of training, a key implicit assumption of our model is that the raw dynamical model provides skillful circulation forecasts at short lead times to effectively guide the weather-scale component. For regions with poor forecast skill even at very short leads, the model may struggle to benefit from weather-scale training. Second, GAN training remains notoriously unstable. Extensive experimentation on hyper-parameters and evaluation across multiple training iterations are required to determine optimal weights of neural networks. Such computational and storage costs make exhaustive hyper-parameter grid searches, ablation studies of input variables, testing larger mini-ensemble sizes (N), or scaling the model to a larger domain difficult. This complexity also means that spatial generalization presents a significant challenge. Since the framework learns region-specific bias patterns and precipitation climatology, applying it to new regions requires complete retraining with locally available data and empirical adjustment of hyperparameters.

It is noteworthy that while our framework outputs multivariate fields, only precipitation products are evaluated quantitatively. The post-processed circulation fields exist in a state of practical compromise, simultaneously constrained by $\mathcal{L}_{\text{inv}}$ to remain close to the original dynamical forecast and by $\mathcal{L}_{\text{pred}}$ and $\mathcal{L}_{\text{adv}}$ to resemble observed fields. Consequently, these generated circulation fields may have lower scores than the smoother dynamical model outputs under standard quantitative grid-point metrics. The inclusion of circulation variables is designed to provide physical anchors for precipitation and to stabilize adversarial training, rather than to improve circulation forecasts themselves.

Several avenues for future research may address these limitations and extend the framework's capabilities. A self-supervised strategy that adaptively adjusts the transition threshold (τ_0) based on the flow-dependent predictability of each forecast during training would allow smarter adjustments of constraint strength (λ_{pred}). To achieve joint correction of both large-scale circulation and precipitation, future work could explore more robust invariance constraints, potentially formulated on anomaly and spectrally truncated fields, or even incorporating data assimilation concepts, to nudge the large-scale states without over-constraining the fine-scale details, thereby enabling a truly multivariate post-processing framework.

Adopting more powerful backbone architectures (such as Transformers) could further enhance performance. Diffusion-based generative models are an increasingly popular alternative that, in principle, can be adapted to similar setting with our proposed model. By design, diffusion models offer comparatively robust training by avoiding unstable adversarial optimization, It should be noted, however, that their multi-step denoising inference brings orders of magnitude higher computational costs than the single-step generation of GANs. Advances in fast sampling and distillation may narrow this gap, but at present our GAN-based design offers an efficiency advantage for operational scenarios. Overall, these advances may enable the extension of the framework to larger domains or even quasi-global scales, thereby capturing more universal physical relationships and improving applicability to other regions.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

The data used in this study are publicly available from the following sources: The S2S reforecast data can be accessed through the ECMWF data portal (ECMWF, 2022). The ERA5 reanalysis data set is available from the Copernicus Climate Data Store (Hersbach et al., 2023). The ETOPO topography data can be obtained from NOAA National Centers for Environmental Information (2022). The China Meteorological Forcing Dataset (CMFD) is accessible through the Third Pole Environment Database (Yang et al., 2025). The code for the proposed GAN-W2C framework developed in this study is publicly available (Shi, 2025). The EMOS baseline was implemented using the code provided by Trotta (2024).

Acknowledgments

This research is supported by the National Natural Science Foundation of China (U2242207) and the National Key Research and Development Program of China (2022YFE0106700, 2024YFF0809004). We thank Dr. Ya Wang from the Institute of Atmospheric Physics, Chinese Academy of Sciences, and Sencan Sun from Department of Earth System Science, Tsinghua University, for their helpful discussions.

References

Barthel Sorensen, B., Charalampopoulos, A., Zhang, S., Harrop, B. E., Leung, L. R., & Sapsis, T. P. (2024). A non-intrusive machine learning framework for debiasing long-time coarse resolution climate simulations and quantifying rare events statistics. *Journal of Advances in Modeling Earth Systems*, *16*(3), e2023MS004122. https://doi.org/10.1029/2023MS004122

Brown, A., Milton, S., Cullen, M., Golding, B., Mitchell, J., & Shelly, A. (2012). Unified modeling and prediction of weather and climate: A 25-year journey. *Bulletin of the American Meteorological Society*, *93*(12), 1865–1877. https://doi.org/10.1175/BAMS-D-12-00018.1

Brunet, G., Shapiro, M., Hoskins, B., Moncrieff, M., Dole, R., Kiladis, G. N., et al. (2010). Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bulletin of the American Meteorological Society*, *91*(10), 1397–1406. https://doi.org/10.1175/2010BAMS3013.1

Cai, H., Zhao, Z., Zheng, J., Luo, W., & Li, H. (2022). Evaluation of the dynamical–statistical downscaling model for extended range precipitation forecasts in China. *Atmosphere*, *13*(10), 1663. https://doi.org/10.3390/atmos13101663

Chao, J., Pan, B., Chen, Q., Yang, S., Wang, J., Nai, C., et al. (2025). Learning to infer weather states using partial observations. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(1), e2024JH000260. https://doi.org/10.1029/2024jh000260

Chen, H., Sun, J., & Wang, H. (2012). A statistical downscaling model for forecasting summer rainfall in China from Demeter hindcast datasets. *Weather and Forecasting*, *27*(3), 608–628. https://doi.org/10.1175/waf-d-11-00079.1

de Andrade, F. M., Coelho, C. A. S., & Cavalcanti, I. F. A. (2019). Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dynamics*, *52*(9), 5451–5475. https://doi.org/10.1007/s00382-018-4457-z

Dong, N., Hao, H., Yang, M., Wei, J., Xu, S., & Kunstmann, H. (2024). Deep learning based sub-seasonal precipitation and streamflow forecasting over the source region of the Yangtze River. *Hydrology and Earth System Sciences Discussions*, 1–26. https://doi.org/10.5194/hess-2024-212

ECMWF. (2022). S2s reforecasts daily averaged [Dataset]. *Author*. Retrieved from https://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/(S2SPredictionProject (WWRP/WCRP))

Fan, Y., Krasnopolsky, V., Dool, H. v. d., Wu, C.-Y., & Gottschalck, J. (2023). Using artificial neural networks to improve CFS Week-3–4 precipitation and 2-m air temperature forecasts. *Weather and Forecasting*, *38*(5), 637–654. https://doi.org/10.1175/WAF-D-20-0014.1

Feng, J., Toth, Z., Zhang, J., & Peña, M. (2024). Ensemble forecasting: A foray of dynamics into the realm of statistics. *Quarterly Journal of the Royal Meteorological Society*, *150*(762), 2537–2560. https://doi.org/10.1002/qj.4745

Harris, L., McRae, A. T. T., Chantry, M., Dueben, P. D., & Palmer, T. N. (2022). A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS003120. https://doi.org/10.1029/2022MS003120

He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing dataset for land process studies over China. *Scientific Data*, *7*(1), 25. https://doi.org/10.1038/s41597-020-0369-y

Henry, J., Natalie, T., & Madsen, D. (2021). *Pix2pix gan for image-to-image translation* (pp. 1–5). Research Gate Publication.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hersbach, H., et al. (2023). Era5 hourly data on pressure levels from 1940 to present [Dataset]. *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. https://doi.org/10.24381/cds.bd0915c6

Hess, P., Aich, M., Pan, B., & Boers, N. (2025). Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence*, *7*(3), 1–11. https://doi.org/10.1038/s42256-025-00980-5

Hess, P., Drüke, M., Petri, S., Strnad, F. M., & Boers, N. (2022). Physically constrained generative adversarial networks for improving precipitation fields from earth system models. *Nature Machine Intelligence*, *4*(10), 828–839. https://doi.org/10.1038/s42256-022-00540-1

Hess, P., Lange, S., Schötz, C., & Boers, N. (2023). Deep learning for bias-correcting CMIP6-class Earth system models. *Earth's Future*, *11*(10), e2023EF004002. https://doi.org/10.1029/2023EF004002

Horat, N., & Lerch, S. (2024). Deep learning for postprocessing global probabilistic forecasts on subseasonal time scales. *Monthly Weather Review*, *152*(3), 667–687. https://doi.org/10.1175/MWR-D-23-0150.1

Huang, Z., Zhao, T., Xu, W., Cai, H., Wang, J., Zhang, Y., et al. (2022). A seven-parameter Bernoulli-Gamma-Gaussian model to calibrate subseasonal to seasonal precipitation forecasts. *Journal of Hydrology*, *610*, 127896. https://doi.org/10.1016/j.jhydrol.2022.127896

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In *International conference on learning representations (ICLR)*.

Kolachian, R., & Saghafian, B. (2019). Deterministic and probabilistic evaluation of raw and post processed sub-seasonal to seasonal precipitation forecasts in different precipitation regimes. *Theoretical and Applied Climatology*, *137*(1), 1479–1493. https://doi.org/10.1007/s00704-018-2680-5

Leutbecher, M., & Palmer, T. N. (2008). Ensemble forecasting. *Journal of Computational Physics*, *227*(7), 3515–3539. https://doi.org/10.1016/j.jcp.2007.02.014

Li, J., Bao, Q., Liu, Y., Wang, L., Yang, J., Wu, G., et al. (2021). Effect of horizontal resolution on the simulation of tropical cyclones in the Chinese academy of sciences fgoals-f3 climate system model. *Geoscientific Model Development*, *14*(10), 6113–6133. https://doi.org/10.5194/gmd-14-6113-2021

Li, M., Jin, H., & Shao, Q. (2021). Improvements in subseasonal forecasts of rainfall extremes by statistical postprocessing methods. *Weather and Climate Extremes*, *34*, 100384. https://doi.org/10.1016/j.wace.2021.100384

Li, W., Chen, J., Li, L., Chen, H., Liu, B., Xu, C.-Y., & Li, X. (2019). Evaluation and bias correction of S2S precipitation for hydrological extremes. *Journal of Hydrometeorology*, *20*(9), 1887–1906. https://doi.org/10.1175/JHM-D-19-0042.1

Li, W., Song, J., Hsu, P.-c., & Wang, Y. (2022). Evaluation of the forecast performance for week-2 winter surface air temperature from the model for prediction across scales–atmosphere (mpas-a). *Weather and Forecasting*, *37*(11), 2035–2047. https://doi.org/10.1175/waf-d-22-0054.1

Li, Y., Wu, Z., He, H., Wang, Q. J., Xu, H., & Lu, G. (2021). Post-processing sub-seasonal precipitation forecasts at various spatiotemporal scales across China during boreal summer monsoon. *Journal of Hydrology*, *598*, 125742. https://doi.org/10.1016/j.jhydrol.2020.125742

Liang, P., & Lin, H. (2018). Sub-seasonal prediction over East Asia during boreal summer using the ECCC monthly forecasting system. *Climate Dynamics*, *50*(3), 1007–1022. https://doi.org/10.1007/s00382-017-3658-1

Ling, F., Chen, K., Wu, J., Han, T., Luo, J.-J., Ouyang, W., & Bai, L. (2024). FengWu-W2S: A deep learning model for seamless weather-to-subseasonal forecast of global atmosphere. *arXiv*. https://doi.org/10.48550/arXiv.2411.10191

Liu, L., Bai, H., Feng, G., Gong, Z., & Zhou, B. (2023). Evaluation and correction of sub-seasonal dynamic model forecast of precipitation in eastern China. *Climate Dynamics*, *61*(9), 4643–4659. https://doi.org/10.1007/s00382-023-06788-6

Lyu, Y., Zhu, S., Zhi, X., Ji, Y., Fan, Y., & Dong, F. (2023). Improving subseasonal-to-seasonal prediction of summer extreme precipitation over Southern China based on a deep learning method. *Geophysical Research Letters*, *50*(24), e2023GL106245. https://doi.org/10.1029/2023GL106245

Lyu, Y., Zhu, S., Zhi, X., Wang, J., Ji, Y., Fan, Y., & Dong, F. (2024). Significant advancement in subseasonal-to-seasonal summer precipitation ensemble forecast skills in China mainland through an innovative hybrid CSG-UNET method. *Environmental Research Letters*, *19*(7), 074055. https://doi.org/10.1088/1748-9326/ad5577

Merryfield, W. J., Baehr, J., Batté, L., Becker, E. J., Butler, A. H., Coelho, C. A. S., et al. (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896. https://doi.org/10.1175/bams-d-19-0037.1

Molod, A., Hackert, E., Vikhliaev, Y., Zhao, B., Barahona, D., Vernieres, G., et al. (2020). Geos-s2s version 2: The gmao high-resolution coupled model and assimilation system for seasonal prediction. *Journal of Geophysical Research: Atmospheres*, *125*(5), e2019JD031767. https://doi.org/10.1029/2019jd031767

Monhart, S., Spirig, C., Bhend, J., Bogner, K., Schär, C., & Liniger, M. A. (2018). Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *Journal of Geophysical Research: Atmospheres*, *123*(15), 7999–8016. https://doi.org/10.1029/2017JD027923

Nai, C., Chen, X., Yang, S., Xiao, Z., & Pan, B. (2025). Boosting weather forecast via generative superensemble. *Npj Climate and Atmospheric Science*, *8*(1). https://doi.org/10.1038/s41612-025-01255-x

Nai, C., Pan, B., Chen, X., Tang, Q., Ni, G., Duan, Q., et al. (2024). Reliable precipitation nowcasting using probabilistic diffusion models. *Environmental Research Letters*, *19*(3), 034039. https://doi.org/10.1088/1748-9326/ad2891

Nie, Y., & Sun, J. (2024). Improving dynamical-statistical subseasonal precipitation forecasts using deep learning: A case study in Southwest China. *Environmental Research Letters*, *19*(7), 074013. https://doi.org/10.1088/1748-9326/ad5370

NOAA National Centers for Environmental Information. (2022). Etopo 2022 global relief model [Dataset]. *NOAA NCEI*. https://doi.org/10.25921/fd45-gt74

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., & Lee, J. (2022). Improving seasonal forecast using probabilistic deep learning. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002766. https://doi.org/10.1029/2021MS002766

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., et al. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509. https://doi.org/10.1029/2021MS002509

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., & Higgins, W. (2019). Precipitation prediction skill for the west coast United States: From short to extended range. *Journal of Climate*, *32*(1), 161–182. https://doi.org/10.1175/JCLI-D-18-0355.1

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., & Mardani, M. (2022). *Fourcastnet: A global data-driven high-resolution weather model using adaptive Fourier neural operators*. arXiv preprint arXiv:2202.11214. https://doi.org/10.48550/arXiv.2202.11214

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2024). WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, *16*(6), e2023MS004019. https://doi.org/10.1029/2023MS004019

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., et al. (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z

Robertson, A. W., Vitart, F., & Camargo, S. J. (2020). Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *Journal of Geophysical Research: Atmospheres*, *125*(6), e2018JD029375. https://doi.org/10.1029/2018jd029375

Rodwell, M. J., & Palmer, T. N. (2007). Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society*, *133*(622), 129–146. https://doi.org/10.1002/qj.23

Schepen, A., Zhao, T., Wang, Q. J., & Robertson, D. E. (2018). A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrology and Earth System Sciences*, *22*(2), 1615–1628. https://doi.org/10.5194/hess-22-1615-2018

Shi, W. (2025). sw-meteo/gan_weather2climate: Code_for_gan-w2c [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.17786420

Specq, D., & Batté, L. (2020). Improving subseasonal precipitation forecasts through a statistical–dynamical approach: Application to the southwest tropical Pacific. *Climate Dynamics*, *55*(7), 1913–1927. https://doi.org/10.1007/s00382-020-05355-7

Sun, S., Nai, C., Pan, B., Li, W., Li, L., Li, X., et al. (2025). Fusion of multi-source precipitation records via coordinate-based generative models. *Nature Communications*, *17*(1). https://doi.org/10.1038/s41467-025-67987-9

Trotta, B. (2024). rainforests-paper-code [Software]. *GitHub*. Retrieved from https://github.com/btrotta-bom/rainforests-paper-code

Trotta, B., Owen, B., Liu, J., Weymouth, G., Gale, T., Hume, T., et al. (2024). Rainforests: A machine learning approach to calibrating NWP precipitation forecasts. *Weather and Forecasting*, *39*(11), 1715–1732. https://doi.org/10.1175/WAF-D-23-0211.1

Vandal, T., Kodra, E., Dy, J., Ganguly, S., Nemani, R., & Ganguly, A. R. (2018). *Quantifying uncertainty in discrete-continuous and skewed data with Bayesian deep learning*. (pp. 2377–2386). Association for Computing Machinery. https://doi.org/10.1145/3219819.3219996

Vigaud, N., Tippett, M. K., Yuan, J., Robertson, A. W., & Acharya, N. (2020). Spatial correction of Multimodel ensemble subseasonal precipitation forecasts over North America using local Laplacian eigenfunctions. *Monthly Weather Review*, *148*(2), 523–539. https://doi.org/10.1175/MWR-D-19-0134.1

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., et al. (2017). The subseasonal to seasonal (s2s) prediction project database. *Bulletin of the American Meteorological Society*, *98*(1), 163–173. https://doi.org/10.1175/bams-d-16-0017.1

Vitart, F., Robertson, A. W., & Anderson, D. L. T. (2012). Subseasonal to Seasonal Prediction Project: Bridging the gap between weather and climate. *Bulletin of the World Meteorological Organization*, *61*(2), 23–28.

White, C. J., Carlsen, H., Robertson, A. W., Klein, R. J., Lazo, J. K., Kumar, A., et al. (2017). Potential applications of subseasonal-to-seasonal (S2S) predictions: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorological Applications*, *24*(3), 315–325. https://doi.org/10.1002/met.1654

Wilby, R. L., Hassan, H., & Hanaki, K. (1998). Statistical downscaling of hydrometeorological variables using general circulation model output. *Journal of Hydrology*, *205*(1–2), 1–19. https://doi.org/10.1016/s0022-1694(97)00130-3

Yang, K., He, J., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2025). China meteorological forcing dataset v1.6 (1979-2018) [Dataset]. *National Tibetan Plateau Data Center*. https://doi.org/10.11888/AtmosphericPhysics.tpe.249369.file

Yang, S., Ling, F., Luo, J.-J., & Bai, L. (2025). Improving the seasonal forecast of summer precipitation in southeastern China using a cyclegan-based deep learning bias correction method. *Advances in Atmospheric Sciences*, *42*(1), 26–35. https://doi.org/10.1007/s00376-024-4003-3

Yang, S., Nai, C., Liu, X., Li, W., Chao, J., & Wang, J. (2025). Generative assimilation and prediction for weather and climate. arXiv preprint arXiv:2503.03038. https://doi.org/10.48550/arXiv.2503.03038

10.1029/2025JH000993

Yin, G., Yoshikane, T., Kaneko, R., & Yoshimura, K. (2023). Improving global subseasonal to seasonal precipitation forecasts using a support vector machine-based method. *Journal of Geophysical Research: Atmospheres*, *128*(17), e2023JD038929. https://doi.org/10.1029/2023JD038929

Yuval, J., Langmore, I., Kochkov, D., & Hoyer, S. (2024). Neural general circulation models optimized to predict satellite-based precipitation observations. Retrieved from https://arxiv.org/abs/2412.11973

Zhang, L., Yang, T., Gao, S., Fan, M., Lu, D., Xu, H., & Xiao, C. (2025). An alternative Ensemble Streamflow Prediction approach using improved subseasonal precipitation forecasts from the North America Multi-Model Ensemble Phase II. *Journal of Hydrometeorology*, *1*(aop), 309–326. https://doi.org/10.1175/JHM-D-24-0048.1

Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X., & Cheng, C. (2023). Improving subseasonal-to-seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous U.S. using machine learning models. *Atmospheric Research*, *281*, 106502. https://doi.org/10.1016/j.atmosres.2022.106502

Zhu, H., Wheeler, M. C., Sobel, A. H., & Hudson, D. (2014). Seamless precipitation prediction skill in the tropics and extratropics from a global model. *Monthly Weather Review*, *142*(4), 1556–1569. https://doi.org/10.1175/MWR-D-13-00222.1

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the ieee international conference on computer vision* (pp. 2223–2232).