Fusion of multi-source precipitation records via coordinate-based generative models

Sencan Sun[1], Congyi Nai[2], Baoxiang Pan[2*], Wentao Li[3], Lu Li[4], Xin Li[5], Efi Foufoula-Georgiou[6], Yanluan Lin[1*]

[1*]Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, 100084, China.
[2*]Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China.
[3]State Key Laboratory of Hydrology–Water Resources and Hydraulic Engineering and College of Hydrology & Water Resources, Hohai University, Nanjing, China.
[4]School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China.
[5]Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China.
[6]Department of Earth System Science, University of California Irvine, Irvine, CA, USA.

*Corresponding author(s). E-mail(s): panbaoxiang@lasg.iap.ac.cn; yanluan@tsinghua.edu.cn;

Abstract

Precipitation remains one of the most challenging climate variables to observe and predict accurately. Existing datasets face intricate trade-offs: gauge observations are relatively trustworthy but sparse, satellites provide global coverage with retrieval uncertainties, and numerical models offer physical consistency but are biased and computationally intensive. Here we introduce PRIMER (Precipitation Record Infinite MERging), a deep generative framework that fuses these complementary sources to produce accurate, high-resolution, full-coverage precipitation estimates. PRIMER employs a coordinate-based diffusion model that learns from arbitrary spatial locations and associated precipitation values, enabling seamless integration of gridded data and irregular gauge observations. Through two-stage training—first learning large-scale patterns, then refining with accurate gauge measurements—PRIMER captures both large-scale climatology and local precision. Once trained, it can downscale forecasts, interpolate sparse observations, and correct systematic biases within a principled Bayesian framework. Using gauge observations as ground truth, PRIMER effectively corrects biases in existing datasets, yielding statistically significant error reductions at most stations and furthermore enhancing the spatial coherence of precipitation fields. Crucially, it generalizes without retraining, correcting biases in operational forecasts it has never seen. This demonstrates how generative AI can transform Earth system science by combining imperfect data, providing a scalable solution for global precipitation monitoring and prediction.

1 Introduction

Precipitation–when, where, and how much water falls from the sky to the earth surface–governs freshwater availability, agricultural productivity, flood hazards, and ecosystem health across the globe [1]. Despite its significance, precipitation remains one of the most challenging climate variables to observe and predict accurately. This challenge stems from precipitation's fundamental nature: unlike most climate variables that vary smoothly across space and time, precipitation manifests as discrete, intermittent pulses with striking discontinuities [2, 3]. These processes depend crucially on small-scale cloud microphysics processes [4] that remain poorly understood or simulated. Besides, these processes are highly sensitive to environmental conditions: small perturbations in temperature, humidity, or aerosol concentrations can determine whether clouds produce no rain, light drizzle, or torrential downpours [5, 6]. Furthermore, the triggering and organization of convection–the primary mechanism for intense precipitation–depends on complex interactions between boundary layer turbulence [7], atmospheric stability [8], and mesoscale circulations [9, 10] that remain computationally prohibitive to simulate explicitly. These complexities create fundamental observational and predictive challenges.

Currently, we rely upon three sources to derive precipitation information, namely in-situ gauge observations, remote sensing, and numerical simulation that potentially assimilate in-situ and remote-sensed data [11]. Each of these three sources comes with inherent limitations regarding their accuracy, coverage, and resolution. Ground-based observations from rain gauges provide the most direct and accurate measurements at point locations. However, gauge networks exhibit severe spatial limitations: even $2.5° \times 2.5°$ grid cells contain less than two gauges on avearge [12], let alone the oceanic and remote regions. Satellite remote sensing offers near-global coverage, but measures precipitation indirectly. Passive microwave sensors on polar-orbiting satellites detect emission and scattering signatures from hydrometeors, providing relatively direct estimates but with limited temporal sampling (2-4 observations per day) [13]. Infrared sensors on geostationary satellites offer frequent observations (every 10-30 minutes) but only measure cloud-top temperatures, requiring empirical relationships to infer surface precipitation–a particularly poor assumption for shallow, warm clouds that produce significant precipitation in tropical and maritime regions [14]. Numerical weather prediction and reanalysis products provide physically consistent, complete spatiotemporal coverage by assimilating available observations into dynamical models [15]. However, precipitation in these systems emerges as the end result of a complex chain of parameterized processes—radiation, convection, cloud microphysics, and boundary layer turbulence—each contributing its own errors [16], with their errors compounding multiplicatively. The consequence of these observational and simulational limitations is profound: current precipitation datasets often disagree by as much as the signal itself [11, 16]. In tropical regions, the spread among different products can exceed 300 mm/hr of the mean precipitation [2], fundamentally limiting our ability to close the global water budget, validate climate models, or provide reliable information for water resource management.

A promising solution to these challenges lies in data fusion–leveraging the complementary strengths of multiple data sources to produce precipitation estimates that surpass any individual source in accuracy, resolution, and coverage [17–29]. Among data fusion approaches, Bayesian methods provide the most principled solution. The key insight is elegant: by deriving an informative prior distribution from all available sources, we can encode existing knowledge in a statistically coherent form. Once established, this prior can be updated via Bayes' theorem with any new observation–accounting for each source's unique error characteristics through tailored likelihood functions [30, 31]. The framework naturally weights observations by their reliability and propagates uncertainties to yield full posterior distributions [32], essential for risk assessment.

Recent advances in deep generative models [33–35], particularly probabilistic diffusion models [36, 37], offer a transformative opportunity for implementing the above Bayesian framework. Diffusion models have demonstrated remarkable ability to learn complex, high-dimensional distributions–from natural images [38] to protein structures [39]–making them ideal candidates for capturing the intricate patterns of precipitation. Crucially, these models can serve as learned priors for Bayesian inference, where their probabilistic foundations enable principled uncertainty quantification. Once trained, they function as "plug-and-play" priors [40–42]: the same learned distribution can be applied to diverse inference tasks–bias correction, downscaling, or gap-filling–by simply changing the likelihood function without retraining. Despite the promises, implementing this framework for precipitation faces three fundamental challenges. First, precipitation's extreme spatiotemporal variability–from localized convective cells to continental-scale fronts–makes it extraordinarily difficult to be captured in a single prior distribution. Second, constructing an informative prior becomes paradoxical when no individual data source is trustworthy or comprehensive. Each source captures different aspects of precipitation across mismatched scales, creating a circular dependency where we need accurate data to build a prior, yet need a prior to evaluate data accuracy. Third, even with a reasonable prior, posterior sampling is challenging due to the high dimensionality of precipitation fields and the complexity of observation likelihoods. These barriers define the frontier for deploying generative AI in Earth system science, demanding innovations that transcend conventional generative modelling approaches.

To address these challenges, we introduce PRIMER (Precipitation Record Infinite MERging), a novel framework that reconceptualizes how diffusion models can learn from imperfect, heterogeneous precipitation records here after for relevant probabilistic inference tasks. Our key insight is that probabilistic diffusion models need not be trained on perfect samples – instead, they can be viewed as spectral regression models that progressively learn from low-frequency structures to high-frequency details as we gradually corrupt the target distribution using Gaussain noise [43]. This property enables us to construct an informative prior by learning conditional distributions of precipitation patterns for each data source, where the conditioning explicitly captures each dataset's characteristic biases.

Implementing this multi-source learning faces a fundamental architectural barrier. Conventionally, diffusion models work on samples residing on fixed-resolution grids [44], forcing us to interpolate heterogeneous observations to common resolutions. This interpolation is particularly destructive for precipitation: it smooths sharp gradients at convective boundaries, introduces artificial correlations between sparse gauge points, and—most critically—destroys the very precision that makes gauges valuable. For sparse gauge networks covering less than 1% of the domain, interpolation essentially fabricates information that doesn't exist. We therefore require an architecture that can learn priors directly from each source's native sampling structure. This necessity drives our adoption of coordinate-based diffusion models, which represent precipitation as continuous spatial fields $x : \mathbb{R}^2 \to \mathbb{R}$ rather than tensors. In this formulation, both dense grids and sparse gauge observations are simply different sampling patterns of the same underlying field. PRIMER directly learns from arbitrarily and sparsely distributed points—each defined by its latitude, longitude, and precipitation intensity—without relying on spatial interpolation (see Fig.1a)—gauge observations influence the function locally while gridded data constrain large-scale structure. Our two-stage training strategy is thus a natural choice: we first learn the baseline priors $P_{\mathrm{ERA5}}(x)$ and $P_{\mathrm{IMERG}}(x)$, which represent the climatological distributions of precipitation fields x derived from climate reanalysis, i.e., fifth generation ECMWF atmospheric reanalysis (ERA5), and satellite-based retrieval dataset, i.e., Integrated Multi-satellitE Retrievals for GPM (IMERG). We then fine-tune the model using gauge observations to incorporate local accuracy, yielding the updated prior $P_*(x)$ (Fig. 1b; star indicates that it supposes to be a better prior). The coordinate-based representation ensures that

gauge information enhances rather than corrupts the prior, as each source contributes at its natural scale. Once trained, PRIMER supports diverse applications through principled posterior sampling: given observations $\mathcal{O}$—whether from biased satellites, sparse gauges, or coarse forecasts—we can sample from posterior $P_*(x \mid \mathcal{O})$ to produce improved ensemble estimates. Empirical evaluations demonstrate the effectiveness of our approach: when assessed against approximately 1,000 independent rain gauges across a diverse set of precipitation events, PRIMER reduces errors at the majority of sites. It also enhances the representation of extreme events and improves the realism of spatial structures, and generalizes effectively to previously unseen operational forecasts without retraining. By transforming the challenge of heterogeneous, imperfect data from a limitation into a strength, PRIMER establishes a new paradigm for precipitation data fusion that extends naturally to other Earth system variables plagued by observational trade-offs.

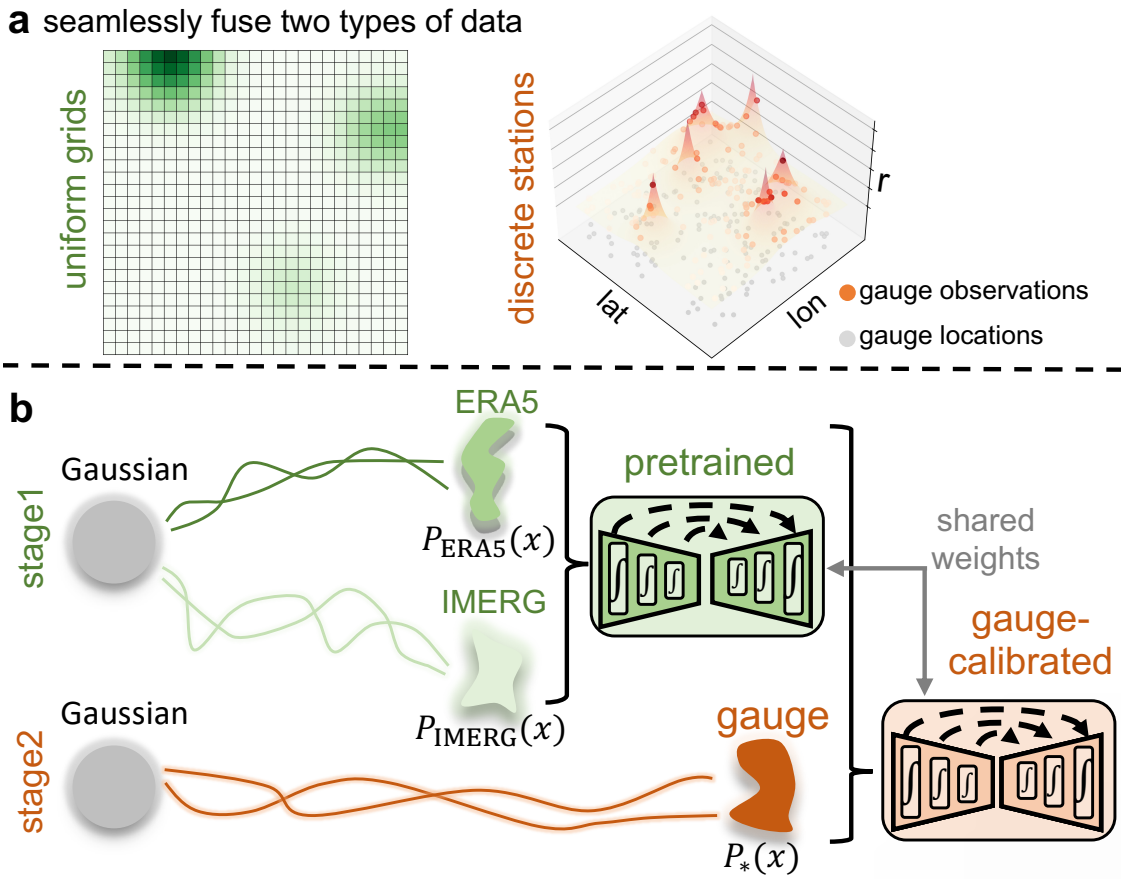

Fig. 1: **Overview of PRIMER.** (**a**) No single precipitation dataset provides uniformly reliable estimates across all spatial scales. PRIMER addresses this challenge by integrating heterogeneous data sources, including gridded reanalysis (e.g., ERA5), satellite-retrieved products (e.g., IMERG), and sparse but accurate in-situ gauge observations. (**b**) Our goal is to fuse information from these diverse datasets into a coherent and accurate prior distribution. PRIMER is trained in two stages. In **Stage 1**, the model is pretrained on gridded datasets to learn baseline priors $P_{\mathrm{ERA5}}(x)$ and $P_{\mathrm{IMERG}}(x)$. In **Stage 2**, it is fine-tuned using gauge observations and their corresponding locations to produce a updated prior $P_*(x)$. Weight sharing across data sources enables the model to leverage large-scale spatial patterns from gridded products while incorporating localized constraints from sparse gauge observations. In the following experiments, we will demonstrate that $P_*(x)$ yields superior accuracy compared to $P_{\mathrm{ERA5}}(x)$ and $P_{\mathrm{IMERG}}(x)$.

2 Results

2.1 Reproducing climatological distributions

The gist of the PRIMER methodology is to learn a trustworthy prior distribution of precipitation fields, thereafter applying it for a broad range of relevant probabilistic inference tasks, so as for accurate, high-resolution, full-coverage precipitation estimates and forecasts. Before verifying the probabilistic inference results, we should ensure the accuracy of the learned prior distribution. As directly evaluating such high-dimensional priors is intractable, we instead assess their statistical properties as proxies [45–47]. We compare unconditionally generated samples from $P_{\mathrm{IMERG}}(x)$, $P_{\mathrm{ERA5}}(x)$, and the final prior $P_*(x)$ against their respective reference datasets. In particular, we focus on the climatological mean and standard deviation of precipitation (Fig. 2). At the grid-point level, the agreement is clear. For mean precipitation (Fig. 2a–f), both $P_{\mathrm{IMERG}}(x)$ and $P_{\mathrm{ERA5}}(x)$ exhibit strong spatial correspondence with IMERG and ERA5, achieving Pearson correlation coefficients (PCCs) of 0.85 and 0.97, respectively. The standard deviation fields (Fig. 2g–l) are likewise well reproduced (PCC = 0.75 and 0.86), highlighting PRIMER's capacity to represent not just the average precipitation distribution but also its variance. Notably, we also introduce the updated prior $P_*(x)$, constructed by fine-tuning PRIMER using sparse but reliable gauge observations (data description is available in Method 4.6). Despite the limited spatial coverage of gauge observations, this calibration yields a climatologically consistent prior that preserves spatial structures learned from the gridded products while injecting localized realism. This "climatological jailbreak" illustrates how PRIMER can adapt to sparse in situ records without compromising coherence across scales. To further evaluate spatial structure, we perform a radially averaged power spectral density (RAPSD) analysis (Fig. 2m), which confirms that the learned priors accurately recover the multiscale spectral characteristics of the reference datasets, especially across mesoscale wavelengths, which are crucial for convective processes (see also Supplementary Information (SI) Fig. D6). Additional statistical evaluations—including precipitation frequency, extremes, skewness, and Empirical Orthogonal Function (EOF) modes—are provided in the SI Fig. D7.

2.2 Case study on high-impact events

The previous section evaluated PRIMER's ability to match climatological distributions. After Stage 2 fine-tuning, the updated prior $P_*(x)$ is expected to align more closely with gauge observations; however, its actual skill remains to be validated through posterior sampling experiments. To this end, we perform posterior sampling using different priors while conditioning on the same observations $\mathcal{O}$. By comparing the posterior samples against the held-out gauge data, we directly assess the impact of the prior on posterior accuracy, thereby quantifying how much fine-tuning improves alignment with real-world observations. We examine three representative high-impact events. These events were selected to span a wide range of precipitation regimes, including prolonged precipitation associated with the Meiyu front, heavy precipitation driven by landfalling typhoons, and localized convective extremes. The primary case, which occurred over Hubei Province, China, during the East Asian summer monsoon on 2 July 2016, is shown in Fig. 3; additional examples are provided in Fig. D9 and Fig. D10.

To evaluate the effectiveness of the posterior sampling, we define a relative skill metric, $\Delta\mathcal{M}$, based on standard performance scores, including mean absolute error (MAE) and the continuous ranked probability score (CRPS). The CRPS provides a probabilistic measure of an ensemble system's accuracy (see Method 4.5.1). For each metric, $\Delta\mathcal{M}$ quantifies the improvement relative to the original precipitation datasets (ERA5 or IMERG), with positive values indicating reduced error or enhanced skill. All evaluations are conducted

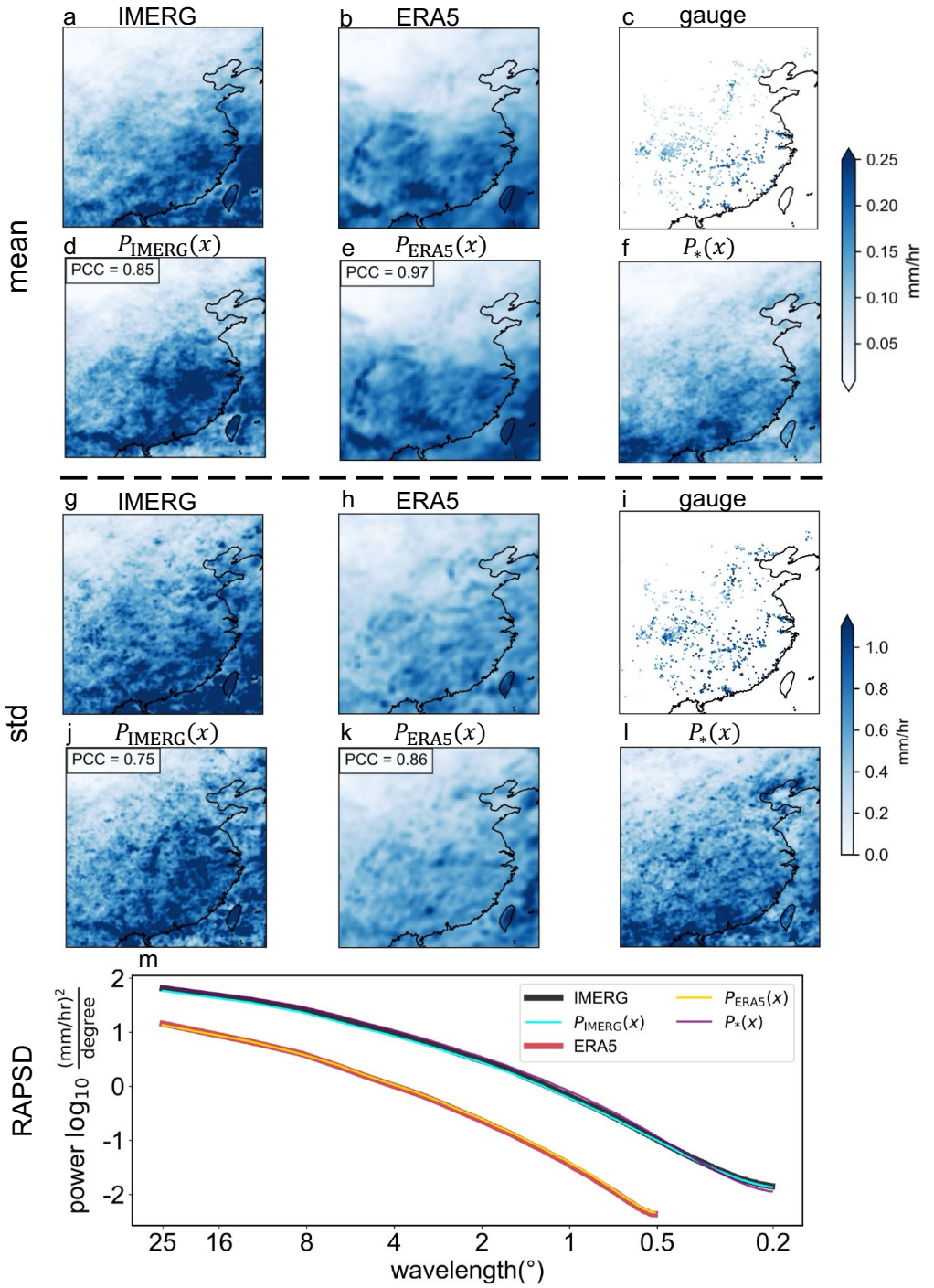

Fig. 2: **Climatological consistency between learned priors and reference datasets. a–f**, Spatial distributions of mean precipitation from IMERG (**a**), ERA5 (**b**), gauge observations (**c**), the learned prior $P_{\mathrm{IMERG}}(x)$ (**d**), $P_{\mathrm{ERA5}}(x)$ (**e**), and the final updated prior $P_*(x)$ (**f**). **g–l**, Standard deviation fields analogous to **a–f**. Pearson correlation coefficients (PCCs) between each learned prior and its corresponding reference (IMERG or ERA5) are indicated in the upper-left corner of relevant panels. **m**, Radially averaged power spectral density (RAPSD) as a function of spatial wavelength (in degrees), quantifying the spatial structure of precipitation fields. The learned priors $P_{\mathrm{IMERG}}(x)$ and $P_{\mathrm{ERA5}}(x)$ closely follow their references, and $P_*(x)$ captures consistent multiscale characteristics. All statistics are computed from 1,000 randomly sampled realizations. Colorbars represent the units for each corresponding row.

at a spatial resolution of 0.1°, where ERA5, IMERG, and posterior samples are compared against independent gauge observations treated as ground truth.

As shown in Fig. 3c and Fig. 3f, the updated prior $P_*(x)$ substantially outperforms baseline priors derived from reanalysis (ERA5) and satellite retrievals (IMERG). The ensemble-mean ΔMAE decreases from 0.46 mm/hr for $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ to 0.14 mm/hr for $P_{\text{ERA5}}(x \mid \mathcal{O}_{\text{ERA5}})$; a similar improvement is observed in the IMERG case, where the ΔMAE decreases from 0.29 mm/hr to 0.14 mm/hr. These gains extend beyond ensemble means: across individual samples, ΔMAE values for $P_{\text{ERA5}}(x \mid \mathcal{O}_{\text{ERA5}})$ are consistently lower than those for $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ (see SI Fig. D8). PRIMER allows the posterior sampling process to incorporate not only the background field but also additional gauge observations, if available. To evaluate this capability, we conduct an experiment where a subset (20%) of gauge observations are assimilated during sampling (denoted as "+ Inpaint"). This additional constraint significantly enhances accuracy, with posterior mean ΔMAE increasing to 1.11 mm/hr and 0.97 mm/hr for the ERA5 and IMERG cases, respectively. This highlights PRIMER's capacity to integrate background field with observational data. Spectral analysis further highlights distinctions among posterior samples (see SI Fig. D8). While $P_{\text{ERA5}}(x \mid \mathcal{O}_{\text{ERA5}})$ retains low-frequency biases, both $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ and its Inpaint variant enhance high-frequency components.

2.3 Statistical verifications

We applied PRIMER to a curated test set of 150 precipitation events from 2016, selected based on the criteria detailed in SI C.2. For each event, 50 posterior samples were drawn from $P_*(x \mid \mathcal{O})$, where $\mathcal{O}$ corresponds to raw data from either ERA5 or IMERG. In this process, PRIMER downscales ERA5 data to 0.1° resolution and performs bias correction, while directly correcting biases in IMERG. To evaluate the improved accuracy of the prior $P_*(x)$, we also conducted posterior sampling using the baseline priors ($P_{\text{ERA5}}(x)$ and $P_{\text{IMERG}}(x)$) under identical settings. At each gauge location, we computed the mean absolute error (MAE) and continuous ranked probability score (CRPS) of the posterior distributions. MAE was calculated using the ensemble mean of each posterior compared against the corresponding gauge observation, while CRPS assessed the full probabilistic accuracy. We then calculated differences in both metrics between the baseline posteriors—$P_{\text{ERA5}}(x \mid \mathcal{O})$ and $P_{\text{IMERG}}(x \mid \mathcal{O})$—and the posterior $P_*(x \mid \mathcal{O})$. Specifically, ΔMAE and ΔCRPS are defined as the baseline scores minus those of $P_*(x \mid \mathcal{O})$, such that positive values indicate improved performance.

Figures 4a–b reveal widespread reductions in mean absolute error (MAE), highlighting PRIMER's ability to systematically correct biases inherent in the original datasets, outperforming the baseline priors $P_{\text{ERA5}}(x)$ and $P_{\text{IMERG}}(x)$. Figures 4c–d show even deeper blue tones in CRPS, indicating more substantial improvements in probabilistic estimates. This suggests that PRIMER not only improves point estimates but also models the full posterior distribution more accurately, thereby reducing uncertainty and enhancing the reliability of ensemble-mean outputs. Notably, the largest improvements are observed in the Sichuan Basin and Pearl River Delta—regions with dense populations and strong economic activity—likely due to the higher density of gauge observations available for Stage 2 fine-tuning.

Beyond reducing pointwise error, PRIMER also enhances the physical realism of existing precipitation datasets. To comprehensively evaluate the performance of PRIMER, we adopt two complementary perspectives: the *member* view and the *envelope* view. The *member* view analyzes statistics from a single sample, representing one physically plausible realization. In contrast, the *envelope* is constructed by selecting, at each gauge location, for a given event, the maximum precipitation value across 50 posterior samples. As illustrated in Fig. 5a, both $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ and $P_*(x \mid \mathcal{O}_{\text{IMERG}})$ more accurately reproduce the frequency distribution of precipitation, particularly at higher intensities. Both perspectives

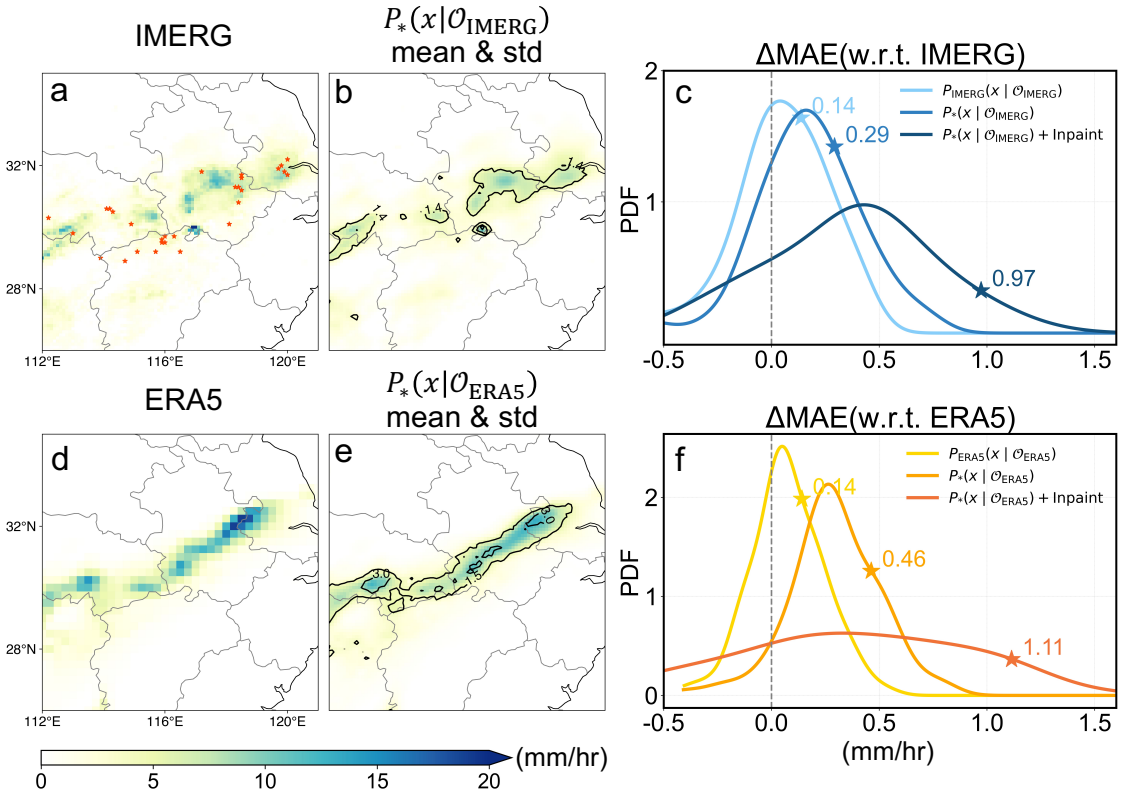

Fig. 3: **Case study of a Meiyu precipitation event on 2 July 2016 at 05 UTC.** **a**, Precipitation field from IMERG at the target time, with gauge locations shown as red dots (used as ground truth for evaluation). **b**, Posterior mean and standard deviation from $P_*(x \mid \mathcal{O}_{\mathrm{IMERG}})$ inferred by PRIMER. **c**, Probability density functions (PDFs) of changes in mean absolute error (ΔMAE), computed at gauge locations by comparing 100 posterior samples and the original IMERG data against observations. For each posterior sample, ΔMAE is calculated as the difference between the sample's MAE and that of IMERG, with positive values indicating effective bias correction by PRIMER. **d**, Precipitation field from ERA5. **e**, Posterior mean and standard deviation from $P_*(x \mid \mathcal{O}_{\mathrm{ERA5}})$. **f**, PDFs of ΔMAE relative to ERA5, analogous to **c**. In **c,f**, different curves represent various posterior distributions as labeled; ensemble means are marked with stars.

reveal improvements in the representation of heavy precipitation tails compared to the existing datasets, underscoring PRIMER's capacity to detect high-impact precipitation events that are often underrepresented in deterministic products. Improvements in spatial structure are further quantified using pixel-wise Pearson correlation coefficients (PCCs) with respect to gauge observations (Fig. 5b). $P_*(x \mid \mathcal{O}_{\mathrm{ERA5}})$ and $P_*(x \mid \mathcal{O}_{\mathrm{IMERG}})$ show markedly enhanced structural agreement relative to existing datasets, suggesting that PRIMER not only reduces local biases but also restores spatial coherence. While various methods have been proposed to assess spatial organization and feature propagation [48, 49], we employ a simplified yet informative diagnostic based on two-dimensional spatial lagged correlation coefficient (Method 4.5.2, Fig. 5c). Physically, these correlation characterizes how anomalies at a reference point are spatially linked to those at surrounding locations, thereby revealing key features of precipitation system organization. We approximate the 0.6 correlation contour with an ellipse and extract two geometric descriptors: the focal length (F), indicative of spatial extent, and the orientation (O), which captures the dominant directional alignment. Results show that both $P_*(x \mid \mathcal{O}_{\mathrm{ERA5}})$ and $P_*(x \mid \mathcal{O}_{\mathrm{IMERG}})$ produce orientations that are more consistent with reference orientations derived from

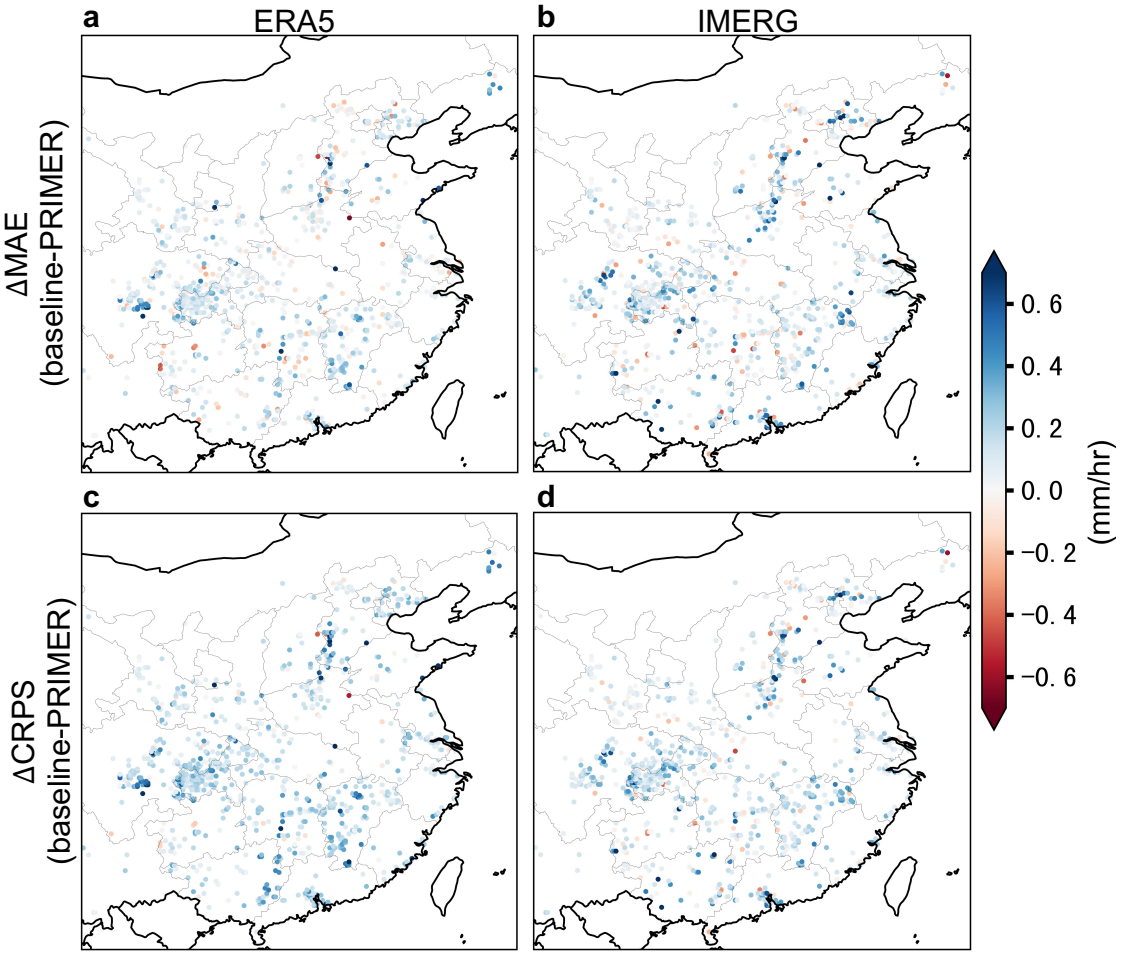

Fig. 4: **Improvement of PRIMER over the baseline in bias correction of existing precipitation datasets. a,b**, Improvements of PRIMER over the baseline after bias correction of ERA5 (**a**) and IMERG (**b**) evaluated by mean absolute error (MAE). **c,d**, Corresponding changes in continuous ranked probability score (CRPS) under the same settings. Each dot denotes a gauge station, with errors evaluated against gauge observations (serving as ground truth). Positive values (blue) indicate improved performance of PRIMER relative to the baseline model, while negative values (red) denote deterioration. The predominance of positive values suggests that PRIMER consistently achieves better bias correction effect, likely due to its ability to learn a more accurate prior distribution $P_*(x)$ by leveraging sparse, discrete gauge observations. Error statistics are based on 150 precipitation events from 2016. For spatial distributions of each model's MAE relative to ERA5 or IMERG, refer to SI Fig. D11.

gauge observations, indicating improved spatial alignment. In terms of focal length, $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ exhibits a clear reduction, while $P_*(x \mid \mathcal{O}_{\text{IMERG}})$ shows no substantial improvement. These results demonstrate PRIMER's effectiveness in correcting spatial anisotropy of precipitation systems.

2.4 Improving operational forecasts without retraining

PRIMER is not only effective for bias correction and downscaling of existing precipitation datasets, but also exhibits strong generalization. Figure 6 illustrates PRIMER's ability to correct biases in previously unseen operational precipitation forecasts, using the ECMWF High-Resolution Forecast (HRES) as a representative example [50]. Despite

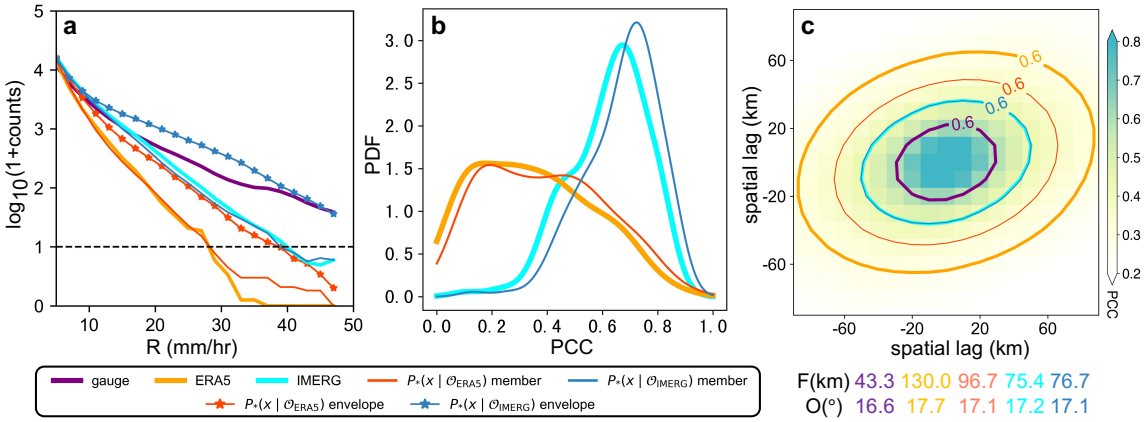

Fig. 5: **Improved intensity distribution and spatial coherence after bias correction of existing datasets. a**, Log-transformed histogram of precipitation intensity (2 mm/hr bins) at only gauge locations, aggregated over test sets. This panel highlights the ability of different datasets to reproduce the tail of the precipitation distribution (with purple line as the ground truth). **b**, Probability density functions (PDFs) of pixel-wise Pearson correlation coefficients (PCCs) between each dataset and the individual gauge observations. Higher PCC values indicate better structural fidelity to ground truth. **c**, Spatial lag correlation maps, with the 0.6 PCC contour visualized for each dataset. Elliptical fits to these contours are used to quantify the spatial coherence, including the major axis length (focal distance, F) and orientation angle (O), as summarized below **c**. Colors in panels **a**–**c** are illustrated in the below legend.

never being trained on HRES, PRIMER successfully corrects systematic biases in a typical precipitation event caused by typhoon landing (Fig. 6a,e). The ensemble mean of $P_*(x \mid \mathcal{O}_{\text{HRES}})$ (Fig. 6b,f) aligns with HRES, while each member (Fig. 6c,g) captures a diverse range of physically plausible precipitation scenarios, reflecting the model's ability to encode meaningful uncertainty. Maps of ΔCRPS (Fig. 6d,h) with widespread positive values (blue dots) indicate that PRIMER produces a reliable probabilistic ensemble system for HRES. These improvements arise from the Bayesian posterior sampling mechanism. By drawing samples from $P_*(x \mid \mathcal{O}_{\text{HRES}})$, we effectively use the learned prior distribution $P_*(x)$—which has been calibrated to match gauge statistics—to adjust the original HRES forecasts. This process mitigates systematic biases inherent in the original HRES forecasts. To illustrate these benefits more intuitively, we present time series at two representative gauge locations (Fig. 6i,j). The ensemble envelopes generated by PRIMER closely track observed precipitation peaks, offering a reliable uncertainty quantification for HRES. Taken together, these results underscore that PRIMER can perform physically consistent corrections on new forecast products without additional retraining (zero-shot adaptation) using its learned prior $P_*(x)$. This highlights the broader utility of PRIMER as a foundation model for downstream applications in precipitation prediction.

3 Discussion

Existing precipitation datasets exhibit a persistent trade-off among spatial coverage, temporal resolution, and measurement accuracy, with no single data source simultaneously meeting these criteria. This fundamental limitation necessitates sophisticated fusion methods capable of integrating heterogeneous observations while overcoming the deficiencies from each source. Generative AI, particularly probabilistic diffusion models, offers a powerful approach by capturing intricate distribution of precipitation patterns. However,

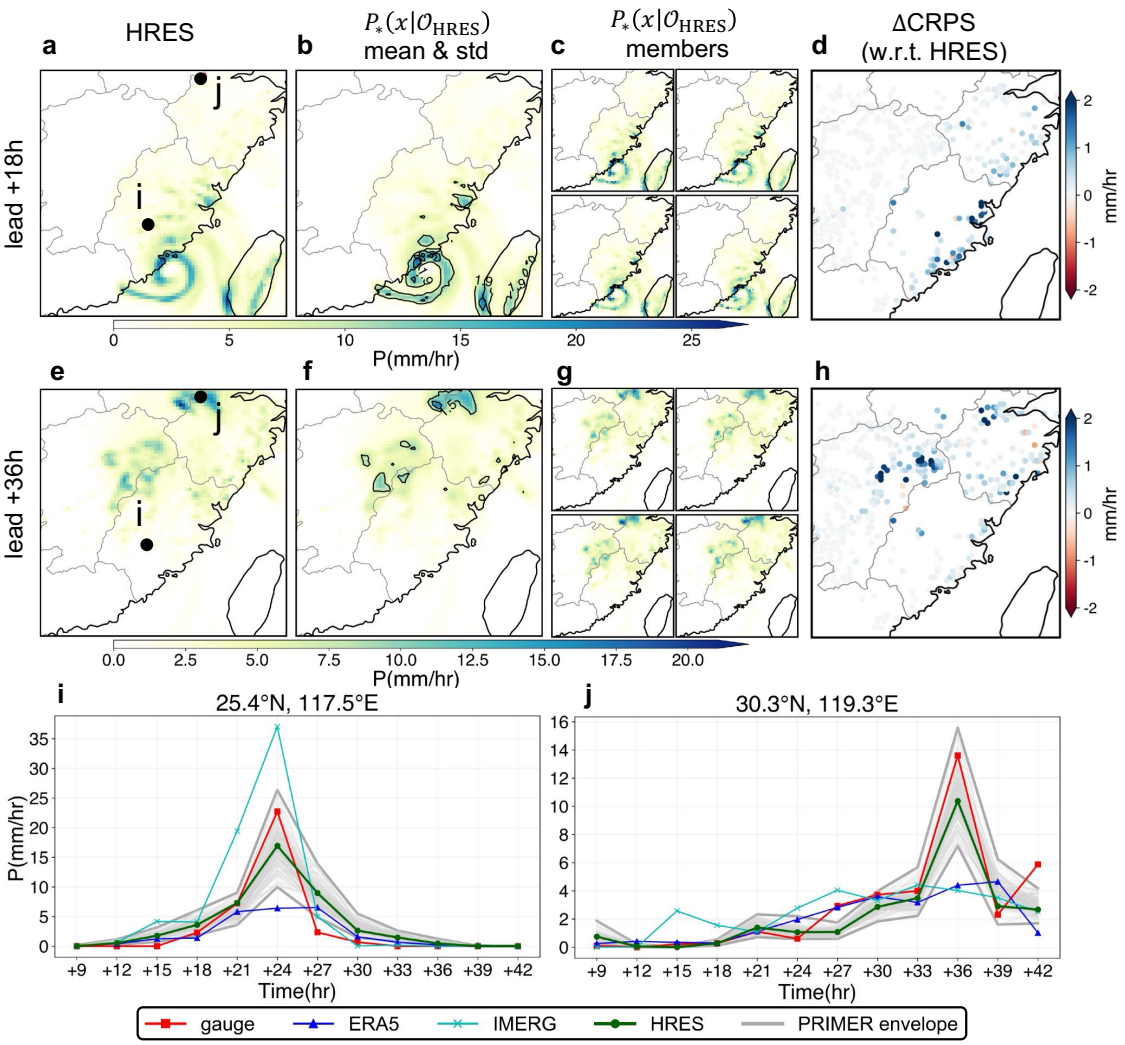

Fig. 6: **Bias-correction for operational forecasts without retraining.** **a**, **e**: HRES forecasts at 18-hr and 36-hr lead times (other lead times see Fig. D14), initialized at 00:00 UTC on 14 September 2016, coinciding with the landfall of Typhoon Meranti. **b**, **f**: Ensemble means. **c**, **g**: Four representative ensemble members, illustrating internal variability and structural diversity. **d**, **h**: Spatial distribution of ΔCRPS, with blue indicating improvement and red indicating deterioration. **i**, **j**: Precipitation time series at two representative gauge stations (more stations see Fig. D15); gray envelope denotes the spread across 100 ensemble members.

practical application has been severely limited by intrinsic challenges: the extreme variability and discontinuity of precipitation, and most important the paradox of establishing reliable priors from individually imperfect and incomplete datasets.

To overcome these barriers, we introduce PRIMER that directly represents precipitation as a continuous spatial field, seamlessly incorporating sparse gauge observations alongside dense gridded data without destructive interpolation. Our two-stage training procedure uniquely exploits the complementary strengths of different data sources: we initially establish robust climatological priors by leveraging broadly available gridded products, which, despite their wide coverage, exhibit considerable uncertainties. These priors are then refined using sparse but accurate gauge observations. Benchmark evaluations highlight PRIMER's capability to effectively integrate gauge observations with gridded data, providing localized realism without sacrificing large-scale spatial

coherence—a significant innovation termed *climatological jailbreak*. Experimental results demonstrate PRIMER's superiority in bias correction and super-resolution enhancement of existing precipitation datasets, consistently outperforming priors derived solely from single-source observations. Furthermore, experiments reveal that incorporating additional gauge observations during posterior sampling process significantly enhances accuracy, underscoring PRIMER's promising potential for future data assimilation applications. Crucially, PRIMER exhibits robust zero-shot generalization, maintaining physical consistency when applied to previously unseen operational forecasts. These findings underscore PRIMER's substantial potential as an advanced, principled approach for reliable and physically coherent precipitation data fusion.

Despite the impressive performance of PRIMER, one notable limitation is the lack of high-quality, in-situ gauge observations over oceanic regions. Sparse instrument coverage in oceanic regions presents a challenge. Another limitation of our study is that we focused on precipitation fusion within China, rather than globally. This decision was primarily driven by the substantial computational demands of performing global precipitation fusion, which would require resources far beyond our current capacity, given that we only have access to two A100 40G GPUs. Additionally, precipitation is one of the most complex and discontinuous variables in the climate system, which provides a stringent benchmark for validating our methodology before considering its application to broader climate domains.

Looking ahead, several compelling directions emerge from this study. First, integrating additional precipitation records [51, 52]—such as more gauge observations or even advanced ground-based radar observations—could further improve the learned prior. Second, the PRIMER framework is intrinsically extensible. Its architecture naturally supports the integration of auxiliary meteorological variables—such as temperature, wind, and humidity—as additional channels. This opens a promising pathway toward holistic representations of the atmospheric state. In particular, applying this framework to future climate simulations from CMIP [53, 54] offers a unique opportunity. By training on Earth system model outputs across multiple scenarios, a generalized version could be developed to learn coherent, external-forcing-aware distributions of the Earth system state. Such a generative model would facilitate projections of future Earth system evolution and deepen our understanding of its underlying statistics. These directions highlight the broader potential behind PRIMER as a scalable and principled foundation for advancing Earth system science.

4 Method

4.1 Problem formulation

A general formulation of the precipitation data fusion task involves two key components: (1) constructing an informative prior distribution over the precipitation field, and (2) performing posterior inference given new observations.

Let x denote the target precipitation field. Different data sources—including gridded products such as satellite-derived and reanalysis datasets, as well as sparse in-situ gauge measurements—provide multiple versions of x, each with varying spatial coverage and accuracy. Our goal is to effectively leverage these heterogeneous sources to construct a unified prior distribution $P(x)$. This prior plays a central role, as it is expected to integrate statistical characteristics of each source through a balanced fusion. A key innovation of this work lies in the design of a principled experimental framework for modelling such a prior.

Once an informative prior is established, posterior inference is conducted as new observational evidence $\mathcal{O}$ becomes available. Posterior distribution $P(x \mid \mathcal{O})$ can be factored into two components: the prior distribution $P(x)$, and the likelihood $P(\mathcal{O} \mid x)$. Another innovation of our work is the effective implementation of posterior inference that balances the prior and the observations, ensuring the inferred precipitation field reflects both the climatological variability and the specific constraints provided by $\mathcal{O}$. Consequently, this Bayesian framework naturally enables various downstream applications, such as super-resolution by conditioning on coarse-resolution data, bias correction by conditioning on biased estimates, and data assimilation by jointly conditioning on observations and background fields.

4.2 Preliminary on diffusion models

To construct a prior distribution, we employ score-based diffusion models within a principled learning strategy. To enable the model to distinguish between sources during training, we associate each sample with a corresponding entity embedding e_i ($e_1 = (1,0,0), e_2 = (0,1,0), e_3 = (0,0,1)$) [55], which is injected into the model. This embedding functions as a source identifier, enabling the model to learn distinct priors for different data sources. Specifically, e_1 corresponds to ERA5, e_2 to IMERG, and e_3 to gauge observations. Here, we first outline the foundations of the traditional diffusion framework before extending its conceptual scope. The forward diffusion process evolves the data distribution into a tractable Gaussian through a stochastic differential equation (SDE) [36, 37, 56, 57]:

$$dx_t = f(x_t, t)\, dt + g(t)\, dW_t, \tag{1}$$

where $x_t \in \mathbb{R}^n$ is the state at time t, $f(x_t, t)$ is the drift function, and W_t is a standard Wiener process. To generate samples from priors, we solve the reverse-time SDE [56, 58]:

$$dx_t = \left[f(x_t, t) - g^2(t)\nabla_{x_t} \log P_\theta(x_t \mid e_i)\right] dt + g(t)\, dW_t, \tag{2}$$

where the score function $\nabla_{x_t} \log P_\theta(x_t \mid e_i)$ denotes the gradient of the log-density with respect to different sources. Since this score is intractable, we approximate it using a neural network f_θ. We provide a theoretical justification for our proposed two-stage training strategy in SI Section B.1.

4.3 PRIMER

Traditional diffusion models typically rely heavily on U-Net architectures [44], which require inputs and outputs to be uniformly gridded data with fixed resolution. This architectural constraint limits their flexibility, particularly when processing discrete, sparse

gauge observations. PRIMER utilizes a new framework inspired by recent theoretical advances [59–62], which generalizes diffusion models from finite-dimensional Euclidean space to an infinite-dimensional Hilbert space $\mathcal{H}$, as illustrated in Figure B1 (see SI Section A for the origin of the name). In this setting, each element $x \in \mathcal{H}$ is a function $x : \mathbb{R}^n \to \mathbb{R}^d$, where $\mathbb{R}^n$ denotes coordinates and $\mathbb{R}^d$ represents physical quantities. Both dense gridded data and sparse gauge observations are treated as partial realizations of an underlying function, allowing PRIMER to natively integrate heterogeneous records. Following Bond *et al.* [59], we define $\mathcal{H}$ as $L^2([0,1]^n \to \mathbb{R}^d)$, where L^2 denotes the space of functions f such that $\int_{[0,1]^n} |f(x)|^2 \, dx < \infty$.

4.3.1 Mollification

While tempting, using white noise in the forward diffusion process poses a fundamental issue. Let $\epsilon(\mathbf{c})$ be a white noise where each $\mathbf{c} \in \mathbb{R}^n$ is sampled independently from $\mathcal{N}(0,1)$. For ϵ to lie in the Hilbert space $\mathcal{H}$, it must be square-integrable. However, $\epsilon(\mathbf{c})$ violates this, as its norm diverges. To address this, PRIMER applies a Gaussian kernel k to mollify the noise: $\xi(\mathbf{c}) = (k * \epsilon)(\mathbf{c}) = \int_{\mathbb{R}^n} k(\mathbf{c} - \mathbf{c}')\epsilon(\mathbf{c}') \, d\mathbf{c}'$. The resulting smoothed noise is square-integrable and thus belongs to $\mathcal{H}$, as rigorously proven in SI B.2. Similarly, PRIMER also mollifies x_0, which ensures that Lx_0 inherit the same smoothness properties. In practice, this operation is implemented efficiently using Discrete Fourier Transformations (DFT). In Fourier space, mollification corresponds to: $\epsilon(\boldsymbol{\omega}) = e^{\|\boldsymbol{\omega}\|^2 t} \, \xi(\boldsymbol{\omega})$, where $\boldsymbol{\omega} \in \mathbb{R}^n$ denotes the frequency vector, and $t = \sigma^2/2$, with σ being the standard deviation of kernel k (a detailed derivation is provided in SI B.3). Directly applying the inverse transformation is often numerically unstable, thus we employ Wiener filter, defined as [59, 63]: $\tilde{\epsilon}(\boldsymbol{\omega}) = \frac{e^{-\|\boldsymbol{\omega}\|^2 t}}{e^{-2\|\boldsymbol{\omega}\|^2 t} + \epsilon^2} \, \xi(\boldsymbol{\omega})$, where ϵ is a small positive regularization parameter.

4.3.2 Network architecture

Neural Operators are capable for learning a map between two functional spaces [62, 64–66]. Neural operators achieve discretization invariance by learning integral kernels parameterized via neural networks. Specifically, for an input function $x : \mathbb{R}^n \to \mathbb{R}^d$, with observations at m distinct spatial locations, the operator $K(x;\theta)$ is defined as:

$$\big(K(x;\theta)x\big)(\mathbf{c}) = \int_{\mathbb{R}^n} \kappa_\theta\Big(\mathbf{c}, \mathbf{b}, x(\mathbf{c}), x(\mathbf{b})\Big) \, x(\mathbf{b}) \, d\mathbf{b},$$

where $\kappa_\theta : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a kernel function parameterized by θ, which captures complex non-local dependencies. However, applying Neural Operators like FNO [66] directly to extensive spatial domains presents scalability and computational efficiency challenges [59]. PRIMER implements a hybrid multi-scale architecture that synthesizes the strengths of Neural Operators and convolutional networks. PRIMER first processes the input feature $x \in \mathbb{R}^{d \times m}$ using a series of SparseConvResBlocks, which primarily employ sparse depthwise convolutions [67], producing updated features with shape $\mathbb{R}^{D \times m}$, where $D \gg d$. This embedding step projects low-dimensional input features into a higher-dimensional space, a crucial operation in deep learning that enables the model to capture richer representations. For the motivation behind SparseConvResBlock, see SI. B.6. Since the features lie on an irregular set of discrete locations, we project them onto a coarse regular grid based on their spatial coordinates. This transformation aligns the features to a structured tensor. A U-Net is applied to this grid to capture multi-scale context. As we are ultimately interested in observations at the original irregular target locations, the processed grid features are reprojected to these coordinates via bilinear interpolation, yielding a feature tensor of shape $\mathbb{R}^{D \times m}$. Finally, a subsequent series of

SparseConvResBlocks refines the features to produce the output tensor of shape $\mathbb{R}^{d\times m}$. For details about network architecture, see SI B.5

4.3.3 Model training

The model is optimized by minimizing a simplified denoising objective [36, 56, 59] (derivation provided in SI Section B.4):

$$\mathcal{L} = \mathbb{E}_t \left[\| f_\theta(x_t, t, e_i) - \xi \|_{\mathcal{H}}^2 \right], \tag{3}$$

where x_t denotes the noisy input at time step t, e_i represents the embedding of data source, ξ is the ground-truth noise, and $\| \cdot \|_{\mathcal{H}}$ denotes the loss norm defined in Hilbert space $\mathcal{H}$. We adopt a two-stage training procedure. In Stage 1, the model is jointly trained on ERA5 (e_1) and IMERG (e_2) data. In Stage 2, we specialize the pretrained model to sparse gauge observations (e_3), following a personalization-inspired strategy akin to DreamBooth [68]. Specifically, we fine-tune the model using a shared-weight strategy, where training samples are proportionally drawn from multiple data sources. The total loss is computed as:

$$\mathcal{L}_{\text{fine-tuning}} = \alpha_1 \mathcal{L}_{\text{ERA5}} + \alpha_2 \mathcal{L}_{\text{IMERG}} + \alpha_3 \mathcal{L}_{\text{gauge}}, \tag{4}$$

with weights $\alpha_1 = 0.1$, $\alpha_2 = 0.4$, and $\alpha_3 = 0.5$. This strategy enables the model to preserve global climatological priors while adapting to high-fidelity signals, effectively grounding the generative manifold towards real-world observations.

All models are optimized using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and weight decay of 4×10^{-6}. The full training and inference pipelines are summarized in SI Algorithm 1 and SI Algorithm 2, with an overview schematic shown in SI Fig. B2. For the configuration of the hyperparameters, see SI Section B.7.

4.4 Posterior sampling

In tasks such as bias correction, downscaling, and data assimilation, the objective is to infer an unknown target state x from observations $\mathcal{O}$. A Bayesian framework enables the incorporation of prior knowledge through a prior distribution $P(x)$, facilitating posterior inference via Bayes' theorem: $P(x|\mathcal{O}) \propto P(\mathcal{O}|x)P(x)$. When employing PRIMER as priors, the standard reverse-time SDE can be adapted to sample from the posterior distribution. The modified reverse diffusion process takes the form:

$$dx_t = \left[f(x_t, t) - g^2(t) \left(\nabla_{x_t} \log P_\theta(x_t \mid e_i) + \nabla_{x_t} \log P_\theta(\mathcal{O} \mid e_i, x_t) \right) \right] dt + g(t)\, dW_t. \tag{5}$$

This formulation requires two key components: the time-dependent score function $\nabla_{x_t} \log P_\theta(x_t \mid e_i)$, which can be approximated by a trained score network; and the gradient of the likelihood $\nabla_{x_t} \log P_\theta(\mathcal{O} \mid e_i, x_t)$, which remains challenging to estimate due to the generally intractable dependency between $\mathcal{O}$ and x_t. Several recent studies have proposed various strategies to address posterior sampling within the diffusion framework [40, 41, 69]. In light of the characteristics of our problem setting, we adopt two representative approaches: Inpainting [70–72] and SDEdit [73].

Inpainting in diffusion models reconstructs unobserved regions by conditioning on partial observations $\mathcal{O}$. A binary mask $\mathbf{m}$ indicates observed entries ($m_i = 1$ if observed). At each reverse-time step t, a denoised estimate $\hat{x}_t$ is first computed. To enforce consistency with known observations, we blend the latent state using

$$x_t = \mathbf{m} \odot q(x_t|\mathcal{O}) + (1 - \mathbf{m}) \odot \hat{x}_t,$$

where $\odot$ denotes element-wise multiplication. The term $q(x_t|\mathcal{O})$ is constructed by applying the same forward noise process to $\mathcal{O}$; that is, for each observed entry, we simulate its

noisy counterpart at step t under the forward SDE. This blending operation preserves observed values while allowing the model to impute missing regions, approximating the posterior distribution $p(x|\mathcal{O})$. SDEdit can be viewed as a special case of inpainting where the entire input field is treated as observed, i.e., $\mathbf{m} = \mathbf{1}$. However, a key distinction lies in its use of a noise level parameter τ, which determines the strength of forward noise applied to the input before denoising. This parameter controls the extent to which the model is allowed to deviate from the original input, balancing fidelity and diversity. To select an appropriate τ, we conduct a sensitivity analysis on IMERG precipitation data for 13 June 2016 at 23:00 UTC. For each noise level from 0.1 to 0.9 in steps of 0.1, we generate an ensemble of 50 samples from posterior $P_*(x \mid \mathcal{O}_{\text{IMERG}})$ and compute both the ensemble mean root mean square error (RMSE) and the continuous ranked probability score (CRPS) over 50 repeated subsampling trials, each selecting 10 members randomly. As shown in SI Fig. B4, performance improves with increasing τ up to around 0.6, beyond which both RMSE and CRPS begin to deteriorate. This suggests an optimal trade-off at 0.6 noise levels, where PRIMER maintains sufficient variability to explore plausible outcomes while preserving alignment with observational constraints.

4.5 Statistical methods

In the main text, we analyze the statistical properties of different prior distributions, including $P_{\text{ERA5}}(x)$, $P_{\text{IMERG}}(x)$, and the updated prior $P_*(x)$. These priors are then used for posterior sampling. To isolate the effect of the prior, all posterior distributions are conditioned on the same observational evidence $\mathcal{O}$. As a result, differences in posterior accuracy primarily reflect differences in the quality of the corresponding priors. To evaluate the performance of each posterior distribution, we adopt the following evaluation method, described in detail below.

4.5.1 Evaluation metrics

Deterministic accuracy.

To assess the accuracy of the ensemble mean forecast, we report the Mean Absolute Error (MAE) and the Pearson Correlation Coefficient (PCC). MAE captures the average absolute deviation between the predicted ensemble mean $\hat{x}$ and the observed value x:

$$\text{MAE} = \frac{1}{N} \sum_i |\hat{x}_i - x_i| \,. \tag{6}$$

where i indexes the grid points corresponding to the gauge locations. PCC measures the linear association between predicted and observed spatial fields:

$$\text{PCC} = \frac{\sum_i (\hat{x}_i - \bar{\hat{x}})(x_i - \bar{x})}{\sqrt{\sum_i (\hat{x}_i - \bar{\hat{x}})^2} \sqrt{\sum_i (x_i - \bar{x})^2}}. \tag{7}$$

Here, $\bar{\hat{x}}$ and $\bar{x}$ denote the spatial means of the predicted and observed fields, respectively. High PCC indicates strong spatial agreement.

Probabilistic skill.

We use the Continuous Ranked Probability Score (CRPS) [74], a proper scoring rule that measures the quality of probabilistic forecasts by comparing the predicted cumulative distribution function (CDF) F with the observation y. It is defined as:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} \left(F(x) - \mathbf{1}_{\{x \geq y\}}\right)^2 \, dx, \tag{8}$$

where $\mathbf{1}_{\{x \geq y\}}$ is the Heaviside step function centered at y. Lower CRPS value indicates a better-calibrated ensemble system.

4.5.2 Evaluation tool

Spatial lagged correlation coefficient

We evaluate the spatial dependency of a geophysical field $x \in \mathbb{R}^{H \times W}$ by computing its correlation with spatially shifted copies. For each fixed offset $(\Delta i, \Delta j)$, we compute the Pearson correlation between x and its lagged version $x_{\Delta i, \Delta j}$, using only the overlapping valid gauge observations. This metric quantifies the degree to which values at one location are linearly correlated with values at a fixed spatial offset (lag) from that location, thus capturing the spatial dependency structure.

Empirical Orthogonal Function (EOF) decomposition

Given an anomaly matrix $x \in \mathbb{R}^{N \times T}$, where each row corresponds to spatial points and each column represents time instances, EOF decomposition factorizes x via [75]:

$$x = LY, \tag{9}$$

where $L \in \mathbb{R}^{N \times N}$ contains orthonormal spatial modes (EOFs), and $Y \in \mathbb{R}^{N \times T}$ holds the corresponding time coefficients (principal components). EOFs are derived as eigenvectors of the covariance matrix $S = \frac{1}{N-1} x x^\top$, arranged in decreasing order of eigenvalues, which represent the explained variance of each mode.

Radially averaged power spectral density (RAPSD)

To quantify spatial variability, we compute the radially averaged power spectral density (RAPSD) using the open-source Pysteps library [76]. Given a 2D scalar field $f(x, y) \in \mathbb{R}^{H \times W}$, its discrete Fourier transform is $F(k_x, k_y) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f(x, y)\, e^{-2\pi i \left(\frac{k_x x}{H} + \frac{k_y y}{W}\right)}$, and the corresponding power spectral density is

$$P(k_x, k_y) = \frac{1}{HW} \left| F(k_x, k_y) \right|^2 . \tag{10}$$

RAPSD is obtained by averaging $P(k_x, k_y)$ over annular bins of constant radial wavenumber $k = \sqrt{k_x^2 + k_y^2}$:

$$\mathrm{RAPSD}(k) = \frac{1}{N_k} \sum_{(k_x, k_y) \in \mathcal{A}_k} P(k_x, k_y), \tag{11}$$

where $\mathcal{A}_k$ denotes the components in each bin. We express RAPSD as a function of wavelength $\lambda = 1/k$ to highlight scale-dependent variability.

4.6 Data

Pretraining uses two gridded datasets: Integrated Multi-satellitE Retrievals for GPM (IMERG) [77] and ERA5 [78]. IMERG provides global precipitation estimates at 0.1° spatial and 30-minute temporal resolutions, derived from GPM satellite observations. To match ERA5's hourly resolution, pairs of consecutive 30-minute IMERG intervals are averaged to produce hourly estimates. The study focuses on East Asia (20-45°N, 100-125°E), a region of high population density. After cropping, IMERG data form 250×250 grids, with 2000-2020 (excluding 2016) used for training. ERA5, from ECMWF, provides hourly precipitation at 0.25° resolution, yielding 100×100 grids over the same domain. Both datasets are log-transformed as $x' = \log_{10}(0.1 + x)$ and standardized using IMERG statistics. For fine-tuning, we use dataset from Shen et al. [29], constructed using over

30,000 Automatic Weather Stations (AWS) across China. The gridded dataset has a spatial resolution of 0.1° and a temporal resolution of 1 hour, covering 2015 and 2017 for training, and 2016 for testing. We select grid cells with at least one assimilated AWS observation for training, and use a subset with no fewer than four AWS observations as ground truth for evaluation, assuming higher reliability (see SI Fig. C5 for the spatial distribution of these gauges). After identical cropping and preprocessing, the data are organized as two arrays: `gauge_observation` $(N, 1)$ for precipitation intensity and `gauge_coordinate` $(N, 2)$ for location (longitude, latitude), both of which are input into the model during fine-tuning.

IFS HRES is ECMWF's flagship deterministic highresolution model and is widely regarded as one of the best physics-based numerical-weather-forecast models in the world [79, 80]. HRES produces hourly forecasts at a 0.1° horizontal resolution. It is included in our experiments to demonstrate PRIMER's strong generalization capability even on datasets it was not trained on. For consistency, HRES forecasts undergo the same cropping and preprocessing steps as IMERG.

Declarations

- Data availability: ERA5 reanalysis were obtained from the Copernicus Climate Change Service's Climate Data Store (CDS) (https://cds.climate.copernicus.eu). For the quickest access, the WeatherBench2 data archive provides an efficient alternative (https://console.cloud.google.com/storage/browser/weatherbench2). The IMERG data can be accessed from https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHH_07/summary?keywords=imerg. Gauge observations were provided from the China Meteorological Administration (CMA) under license. However, restrictions apply to the availability of these data, which were used under license for the present study. Data are available from the authors upon reasonable request and with permission from the CMA. A small subset of gauge observations will be made available on GitHub to facilitate reproducibility and support code debugging. The high-resolution forecast data (HRES) from the Integrated Forecasting System (IFS) used in this study are produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). For more detailed information on HRES access, please refer https://www.ecmwf.int/en/forecasts/datasets/set-i.
- Code availability: The code implementing PRIMER will be available on the GitHub repository. Model configurations and training scripts used in this study will be made publicly available upon acceptance of this work.
- Acknowledgemenrs: This work was supported by the National Natural Science Foundation of China (42130603).

Appendix A Why "PRIMER"

The name PRIMER (Precipitation Records Infinite MERging) is deliberately chosen—not only as an acronym, but also as a metaphor. In English, a "primer" refers to a preparatory coating applied before the final layer of paint or makeup, ensuring better adhesion, durability, and refinement. Similarly, our framework first performs extensive pretraining on gridded products like ERA5 and IMERG before fine-tuning with sparse, high-quality gauge observations. This staged approach allows PRIMER to seamlessly merge multi-sources of records. We envision this method as a general-purpose "foundation layer" for geoscientific modeling—particularly valuable in domains where accurate downstream tasks rely on the fusing of heterogeneous records.

Appendix B Method details

B.1 Theoretical justification for dual-source integration

We provide a theoretical justification for our proposed two-stage integration framework by drawing parallels with recent advances in diffusion theory [43, 81–84]. Specifically, we aim to establish an upper bound on the Wasserstein-1 distance [85] between the true precipitation distribution $\mathcal{P}_{\text{true}}$ and the learned model distribution $\hat{\mathcal{P}}$. The model is trained in two stages: first on a large, noisy gridded dataset $\mathcal{D}$, and subsequently fine-tuned on a sparse but high-fidelity gauge dataset $\mathcal{D}^*$. In this setup:

- $\mathcal{D}$ comprises low-quality, relatively high-uncertainty samples;
- $\mathcal{D}^*$ comprises accurate, relatively low-uncertainty gauge observations.

Assume we observe precipitation samples $X_1, \ldots, X_n$, independently drawn from the following generative process:

$$X_i \sim \mathcal{P}_{\text{true}} * \mathcal{N}(0, \sigma_i^2 I), \tag{B1}$$

where $\mathcal{P}_{\text{true}} = \sum_{j=1}^{k} w_j \delta_{\mu_j}$ is a finite k-component mixture of point masses (or equivalently, a degenerate Gaussian mixture). Each δ_{μ_j} denotes a representative precipitation mode with weight w_j, and σ_i captures the noise level of the i-th observation. While this discrete formulation does not fully capture the full distribution of precipitation fields, it provides a tractable approximation that enables analytical insight into complex precipitation distributions, in line with common practice in diffusion-based modeling. Our objective is to learn a distribution $\hat{\mathcal{P}}$ that closely approximates $\mathcal{P}_{\text{true}}$ by minimizing their Wasserstein-1 distance:

$$W_1(\mathcal{P}_{\text{true}}, \hat{\mathcal{P}}) = \min_{C(\mathcal{P}_{\text{true}}, \hat{\mathcal{P}})} \mathbb{E}_{(X,X')\sim C}[\|X - X'\|], \tag{B2}$$

where C is a valid coupling between $\mathcal{P}_{\text{true}}$ and $\hat{\mathcal{P}}$ with marginals equal to $\mathcal{P}$ and $\hat{\mathcal{P}}$.

By invoking the theoretical framework established in [81], particularly Theorem 4.2, we obtain the following upper bound. Let n be the total number of samples (from both $\mathcal{D}$ and $\mathcal{D}^*$), d the dimensionality of the data space, and k the number of mixture components. Then, under mild regularity assumptions, there exists a procedure returning $\hat{\mathcal{P}}$ such that with probability at least $1 - \delta$:

$$W_1(\mathcal{P}_{\text{true}}, \hat{\mathcal{P}}) \leq C\left(k \left(\frac{d + \log(1/\delta)}{\sum_{i=1}^{n} 1/\sigma_i^4} \right)^{1/4} + k^3 \left(\frac{k \log k + \log(1/\delta)}{\sum_{i=1}^{n} 1/\sigma_i^{4k-2}} \right)^{1/(4k-2)} \right), \tag{B3}$$

The estimation error bound naturally decomposes into two principal components. The first term reflects the *dimensionality reduction error*, arising from the challenge of projecting

high-dimensional precipitation fields (with dimension d) into a lower-dimensional subspace of k modes. The second term quantifies the *low-dimensional estimation error*, which captures the precision of parameter estimation within this k-dimensional space. This decomposition mirrors the two-stage process used in our framework and aligns closely with recent analyses in ambient diffusion [81]. In the first stage, the model compresses the data into a reduced representation, where the estimation accuracy depends on the effective sample size, approximately represented by $\sum_{i=1}^{n} 1/\sigma_i^4$. Here, the gridded dataset $\mathcal{D}$—although characterized by high noise levels σ_i—offers substantial benefit due to its extensive spatial coverage and large number of samples. While noise rapidly degrades high-frequency information, it affects low-frequency components (i.e., structural patterns) to a lesser degree. Consequently, gridded datasets remain valuable for capturing the overall structure of the precipitation field, making $\mathcal{D}$ instrumental in reducing the dimensionality of the problem. In the second stage, the model performs fine-grained estimation within the reduced space, where performance becomes sensitive to uncertainties. The corresponding term in the bound depends on $\sum_{i=1}^{n} 1/\sigma_i^{4k-2}$, emphasizing the critical role of low-uncertainty gauge observations. Although our gauge dataset $\mathcal{D}^*$ is sparse, its markedly smaller noise variances make it disproportionately influential in this stage. Analogous to the clean data in [81], $\mathcal{D}^*$ effectively preserves high-frequency components. This theoretical framing highlights a key insight: low-quality (noisy) data are primarily useful for capturing structural information and aiding dimensionality reduction, while high-quality (clean) data are essential for refining local accuracy. By strategically combining $\mathcal{D}$ and $\mathcal{D}^*$, our framework balances this trade-off, leveraging the complementary strengths of each data source to robustly approximate the underlying precipitation distribution. To illustrate this point concretely, consider a simplified case where p fraction of the dataset is D^* (gauge observations), and $(1-p)$ fraction is D (reanalysis or satellite retrievals). Then, Eq. (B3) simplifies to (see corollary 4.3 in [81]):

$$W_1(\mathcal{P}_{\text{true}}, \hat{\mathcal{P}}) \leq C\left(k\left(\frac{d+\log(1/\delta)}{n\,(p+(1-p)/\sigma^4)}\right)^{1/4} + k^3\left(\frac{k\log k+\log(1/\delta)}{n\,(p+(1-p)/\sigma^{4k-2})}\right)^{1/(4k-2)}\right), \tag{B4}$$

revealing that the high-uncertainty samples are down-weighted by $1/\sigma^4$ (dimensionality reduction) and, more substantially, by $1/\sigma^{4k-2}$ (fine-grained estimation).

Overall, this theoretical framework underpins the rationale for our two-stage strategy:

1. Pretraining on $\mathcal{D}$ exploits its large sample size to establish robust large-scale spatial structure (reflected in the first term of the bound);
2. Fine-tuning on $\mathcal{D}^*$ leverages its high-fidelity observations to minimize local estimation error (reflected in the second term).

B.2 Proof that mollified white noise belongs to L^2 space

As established in the main text, white noise $\epsilon(\mathbf{c})$ is not an element of space $\mathcal{H} = L^2([0,1]^n \to \mathbb{R}^d)$. In this section, we formally show that its mollified version, obtained via convolution with a Gaussian kernel, is square-integrable and hence admissible within the Hilbert space.

Proof. Let $G(\mathbf{c}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\|\mathbf{c}\|^2}{2\sigma^2}}$ denote a Gaussian kernel with variance σ^2. We define the mollified signal as the convolution:

$$\xi(\mathbf{c}) = (\epsilon * G)(\mathbf{c}).$$

By the convolution theorem, the Fourier transform of ξ is the product of the transforms of its components:

$$\mathcal{F}[\xi](\boldsymbol{\omega}) = \mathcal{F}[\epsilon](\boldsymbol{\omega}) \cdot \mathcal{F}[G](\boldsymbol{\omega}),$$

where $\boldsymbol{\omega} \in \mathbb{R}^n$ denotes the frequency vector. The Fourier transform of the Gaussian is again Gaussian:

$$\mathcal{F}[G](\boldsymbol{\omega}) = e^{-\sigma^2 \|\boldsymbol{\omega}\|^2/2}.$$

Applying Parseval's theorem, the squared L^2 norm of ξ in physical space is equal to that in frequency space:

$$\|\xi\|^2 = \int_{\mathbb{R}^n} |\xi(\mathbf{c})|^2 \, d\mathbf{c} = \int_{\mathbb{R}^n} |\mathcal{F}[\xi](\boldsymbol{\omega})|^2 \, d\boldsymbol{\omega}.$$

Substituting the frequency-domain representation:

$$\|\xi\|^2 = \int_{\mathbb{R}^n} |\mathcal{F}[\epsilon](\boldsymbol{\omega})|^2 \cdot e^{-\sigma^2 \|\boldsymbol{\omega}\|^2} \, d\boldsymbol{\omega}.$$

Assuming $\epsilon(\mathbf{c})$ is white noise, its power spectral density is constant in expectation: $|\mathcal{F}[\epsilon](\boldsymbol{\omega})|^2 = C$. Thus:

$$\|\xi\|^2 \propto \int_{\mathbb{R}^n} e^{-\sigma^2 \|\boldsymbol{\omega}\|^2} \, d\boldsymbol{\omega}.$$

This is a standard Gaussian integral over $\mathbb{R}^n$, yielding:

$$\int_{\mathbb{R}^n} e^{-\sigma^2 \|\boldsymbol{\omega}\|^2} \, d\boldsymbol{\omega} = \left(\frac{\pi}{\sigma^2}\right)^{n/2} < \infty.$$

Therefore, the mollified signal $\xi(\mathbf{c})$ has finite energy and lies in $L^2([0,1]^n \to \mathbb{R}^d)$, satisfying the requirement for inclusion in the Hilbert space $\mathcal{H}$ used in PRIMER.

□

B.3 Proof of Fourier-domain relationship between original and mollified fields

We aim to show that mollifying a signal $x(\mathbf{c}) \in \mathcal{H}$ by convolving it with a Gaussian kernel results in a Fourier-domain relation:

$$\hat{x}(\boldsymbol{\omega}) = e^{\|\boldsymbol{\omega}\|^2 t} \hat{h}(\boldsymbol{\omega}),$$

where $\hat{x}(\boldsymbol{\omega})$ and $\hat{h}(\boldsymbol{\omega})$ denote the Fourier transforms of the original and mollified fields, respectively, and $t = \sigma^2/2$ is determined by the kernel σ. This relation is central to the spectral manipulation used in PRIMER.

Proof. Consider the Gaussian kernel in $\mathbb{R}^n$:

$$k(\mathbf{c}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \, e^{-\frac{\|\mathbf{c}\|^2}{2\sigma^2}}, \quad \mathbf{c} \in \mathbb{R}^n.$$

Its Fourier transform is given by:

$$\hat{k}(\boldsymbol{\omega}) = \int_{\mathbb{R}^n} k(\mathbf{c}) \, e^{-i\boldsymbol{\omega} \cdot \mathbf{c}} \, d\mathbf{c}.$$

Substituting for $k(\mathbf{c})$:

$$\hat{k}(\boldsymbol{\omega}) = \frac{1}{(2\pi\sigma^2)^{n/2}} \int_{\mathbb{R}^n} e^{-\frac{\|\mathbf{c}\|^2}{2\sigma^2}} e^{-i\boldsymbol{\omega}\cdot\mathbf{c}}\, d\mathbf{c}.$$

Complete the square in the exponent:

$$-\frac{\|\mathbf{c}\|^2}{2\sigma^2} - i\boldsymbol{\omega}\cdot\mathbf{c} = -\frac{1}{2\sigma^2}\left(\|\mathbf{c} + i\sigma^2\boldsymbol{\omega}\|^2 + \sigma^4\|\boldsymbol{\omega}\|^2\right).$$

Thus:

$$\hat{k}(\boldsymbol{\omega}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sigma^2\|\boldsymbol{\omega}\|^2}{2}} \int_{\mathbb{R}^n} e^{-\frac{\|\mathbf{c}+i\sigma^2\boldsymbol{\omega}\|^2}{2\sigma^2}} d\mathbf{c}.$$

The integral evaluates to the normalization constant due to translation invariance:

$$\int_{\mathbb{R}^n} e^{-\frac{\|\mathbf{c}+i\sigma^2\boldsymbol{\omega}\|^2}{2\sigma^2}} d\mathbf{c} = (2\pi\sigma^2)^{n/2}.$$

Therefore:

$$\hat{k}(\boldsymbol{\omega}) = e^{-\frac{\sigma^2\|\boldsymbol{\omega}\|^2}{2}}.$$

Setting $\sigma = \sqrt{2t}$ gives:

$$\hat{k}(\boldsymbol{\omega}) = e^{-\|\boldsymbol{\omega}\|^2 t}.$$

For the mollified signal $h(\mathbf{c}) = (x * k)(\mathbf{c})$, the convolution theorem implies:

$$\hat{h}(\boldsymbol{\omega}) = \hat{x}(\boldsymbol{\omega}) \cdot \hat{k}(\boldsymbol{\omega}) = \hat{x}(\boldsymbol{\omega}) \cdot e^{-\|\boldsymbol{\omega}\|^2 t},$$

so rearranging yields:

$$\hat{x}(\boldsymbol{\omega}) = e^{\|\boldsymbol{\omega}\|^2 t} \cdot \hat{h}(\boldsymbol{\omega}).$$

□

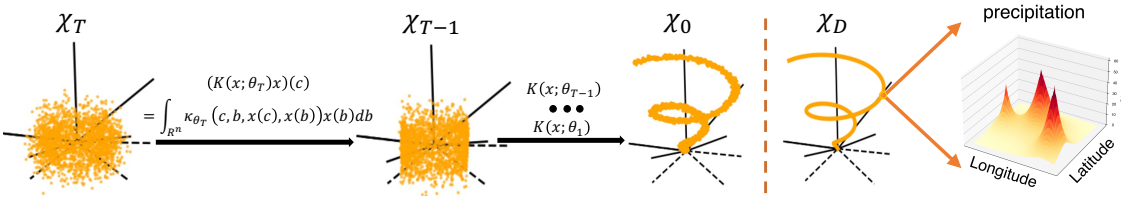

Fig. B1: Conceptual illustration of the denosing process in the PRIMER. Each point within the spaces χ_T, χ_{T-1}, …, χ_0, and χ_D represents a function residing in an infinite-dimensional Hilbert space, as indicated by the axes. Through iterative transformations governed by neural operators $K(x;\theta_t)$ parameterised by kernel κ_{θ_t}, PRIMER progressively transforms the initial χ_T toward the targeted distribution χ_0, closely approximating the desired distribution represented by χ_D. The rightmost panel visually illustrates such a function. The proximity between the distributions of χ_0 and χ_D highlights PRIMER's capability in modeling the phase space (distribution), serving as an useful prior.

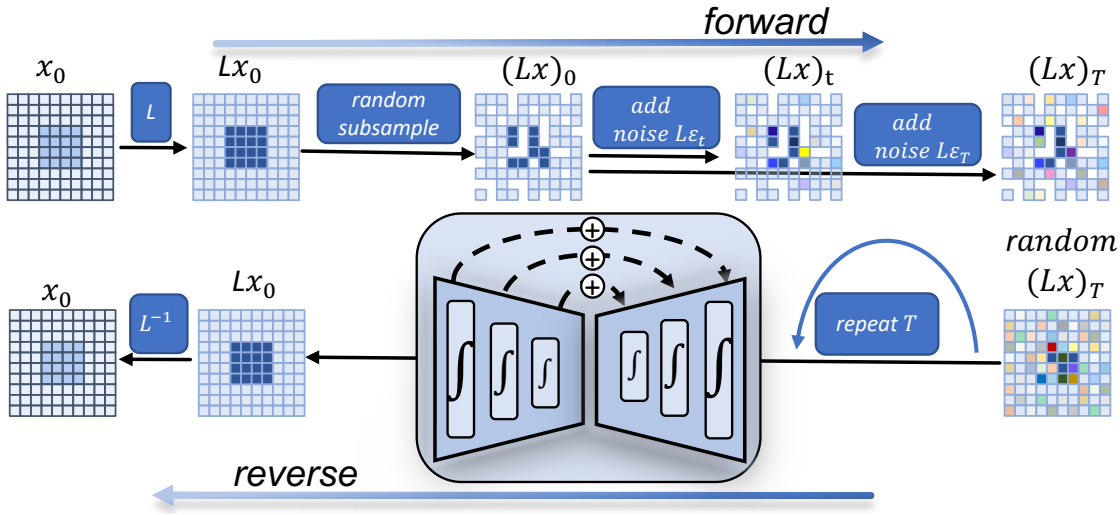

Fig. B2: **Schematic illustration of the algorithm.** The upper row depicts the forward process: starting with an initial state x_0, a smoothing gaussian kernel L is applied, followed by random subsampling to create $(Lx)_0$. Progressive noise addition generates intermediate states $(Lx)_t$ and the final state $(Lx)_T$. The lower row shows the reverse process: beginning with the noisy observation $(Lx)_T$, the neural network iteratively denoises the signal through T repetitions, ultimately recovering Lx_0. The inverse operator L^{-1} then reconstructs the original signal x_0 using Wiener filter.

B.4 Training and inference algorithm pseudocode

We introduce the transition distribution [59] in the forward process:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, Lx_0, (1-\bar{\alpha}_t)LL^*\right), \tag{B5}$$

where L denotes the mollification operator and constant coefficient $\bar{\alpha}_t \in [0,1]$ controls the balance between signal preservation and noise injection. The corresponding posterior distribution can be derived analytically [56, 59]:

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_t(x_t, x_0), \tilde{\beta}_t LL^*\right), \tag{B6}$$

where the mean and variance are given by

$$\tilde{\mu}_t(x_t, x_0) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\xi\right), \quad \xi \sim \mathcal{N}(0, LL^*), \tag{B7}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t. \tag{B8}$$

We use a neural network f_θ to predict ξ. The reverse transition is defined as [56, 59]:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \tilde{\beta}_t LL^*\right), \tag{B9}$$

with the predicted mean given by

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left[x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} f_\theta(x_t, t)\right]. \tag{B10}$$

Thus, PRIMER is trained by minimizing the following simplified objective [36, 56, 59]:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_t \left[\|\mu_\theta(x_t, t) - \tilde{\mu}_t(x_t, x_0)\|^2_{\mathcal{H}} \right] = \mathbb{E}_t \left[\|f_\theta(x_t, t) - \xi\|^2_{\mathcal{H}} \right]. \tag{B11}$$

Algorithm 1 Training Procedure of PRIMER

Require: Gauge observations $x_0 \in \mathbb{R}^{M \times d}$ (M gauges, each with a state vector in $\mathbb{R}^d$), gauge coordinates $C \in \mathbb{R}^{M \times n}$ (spatial locations in $\mathbb{R}^n$), Gaussian mollifier kernel k, diffusion schedule $\bar{\alpha}_t$

Sample white noise $\varepsilon : \mathbb{R}^n \to \mathbb{R}^d \sim \mathcal{N}(0, I)$
for all coordinates $c \in C$ **do**
    Compute mollified noise: $\xi(c) \leftarrow (L\varepsilon)(c) = \int_{\mathbb{R}^n} k(c - c')\varepsilon(c')\, dc'$
    Compute mollified data: $(Lx_0)(c) = \int_{\mathbb{R}^n} k(c - c')x_0(c')\, dc'$
    Compute diffused sample at time t: $x_t(c) \leftarrow \sqrt{\bar{\alpha}_t} \cdot (Lx_0)(c) + \sqrt{1 - \bar{\alpha}_t} \cdot \xi(c)$
end for
Predict mollified noise using neural network: $\hat{\xi} \leftarrow f_\theta(x_t, t)$
Compute loss: $\mathcal{L} \leftarrow \|\hat{\xi} - \xi\|^2_C$
Update model parameters θ using gradient descent on $\mathcal{L}$

B.5 Network architecture

The overall network architecture is designed to flexibly handle sparse and irregularly distributed observations, such as those from in-situ rain gauges, while maintaining strong representational capacity across heterogeneous data sources. As detailed in Section 4.3.2, the key distinction from a standard U-Net lies in the inclusion of multiple stacked *SparseConvResBlock* modules at both the input and output stages of the network. These modules are specifically designed to process inputs with sparse spatial distributions. The input to the model consists of feature representations $x \in \mathbb{R}^{B \times L \times C}$ along with their corresponding spatial coordinates in $\mathbb{R}^{B \times L \times 2}$, where B is the batch size, L is the number of gauge locations, and C is the number of feature channels. After being processed by a series of *SparseConvResBlock* modules, the features retain their shape while being adapted to the sparsity of the input. These processed features are then transformed onto a coarser, structured grid, which facilitates subsequent processing using a conventional U-Net. See Supplementary Information, Listing 1, for the PyTorch implementation of the sparse-to-grid transformation.

Algorithm 2 Inference Procedure of PRIMER

Require: Coordinates $C \in \mathbb{R}^{M\times n}$ (M gauges, each are located in $\mathbb{R}^n$), mollifier kernel k, trained network f_θ, diffusion schedule $\bar{\alpha}_t$, inverse signal-noise-ratio ϵ

Sample Gaussian white noise $\varepsilon \sim \mathcal{N}(0, I)$
for all coordinates $c \in C$ **do**
    Set initial sample:

$$x_T(c) \leftarrow \int_{\mathbb{R}^n} k(c - c')\varepsilon(c')\, dc'$$

end for
for $t = T, T-1, \ldots, 1$ **do**
    Predict mollified noise: $\hat{\xi} \leftarrow f_\theta(x_t, t)$
    Estimate denoised state:

$$\hat{x}_0 \leftarrow \frac{x_t - \sqrt{1-\bar{\alpha}_t}\cdot\hat{\xi}}{\sqrt{\bar{\alpha}_t}}$$

    if $t > 1$ **then**
        Sample new noise: $\varepsilon \sim \mathcal{N}(0, I)$
        for all coordinates $c \in C$ **do**
            Compute mollified noise:

$$\xi_{t-1}(c) \leftarrow \int_{\mathbb{R}^n} k(c - c')\varepsilon(c')\, dc'$$

        end for
        Update sample:

$$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\cdot\hat{x}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\cdot\xi_{t-1}$$

    end if
end for
Apply Wiener filtering:

$$x_0(\omega) \leftarrow \frac{e^{-\omega^2 t}}{e^{-2\omega^2 t} + \epsilon^2}\cdot\hat{x}_0(\omega)$$

return $x_0(c)$

```python
from pytorch3d.ops import knn_points, knn_gather

def knn_interpolate_to_grid(x, coords, uno_coords, knn_neighbours):
    """
    Interpolates sparse features to a structured grid using KNN.

    Args:
        x: Tensor of shape (B, L, C), input features at irregular
            locations
        coords: Tensor of shape (B, L, 2), spatial coordinates of x
        uno_coords: Tensor of shape (y_length, 2), coordinates of target
            structured grid
        knn_neighbours: int, number of nearest neighbors to use
    Returns:
        Tensor of shape (B, y_length, C), interpolated features
    """

    B = x.size(0)
    target_coords = uno_coords.unsqueeze(0).repeat(B, 1, 1)  # (B,
        y_length, 2)

    with torch.no_grad():
        _, assign_index, neighbour_coords = knn_points(
            target_coords, coords, K=knn_neighbours, return_nn=True
        )

        # neighbour_coords: (B, y_length, K, 2)
        diff = neighbour_coords - target_coords.unsqueeze(2)
        squared_distance = (diff * diff).sum(dim=-1, keepdim=True)
        weights = 1.0 / torch.clamp(squared_distance, min=1e-15)

    neighbours = knn_gather(x, assign_index)  # (B, y_length, K, C)
    out = (neighbours * weights).sum(2) / weights.sum(2)

    return out.to(x.dtype)
```

Listing 1: Transform sparse gauge representation to a structured coarse grid.

Upon completion of the U-Net forward pass, the resulting features are bilinearly interpolated (using torch.nn.functional.grid_sample function) back to the original set of irregular input coordinates. Finally, multiple *SparseConvResBlock* modules are applied to further refine the outputs at target spatial locations $\mathbb{R}^{B\times L\times 2}$. The architecture of the *SparseConvResBlock* module is shown in Figure B3, highlighting its ability to integrate conditioning on both diffusion timestep and dataset source labels, enabling the model to operate seamlessly across multi-source inputs with varying spatial coverage.

B.6 Rationale for the use of sparse convolution in PRIMER

Let $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ denote a spatially continuous function defined over a bounded domain D, observed only at a finite set of locations $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^N \subset D$, corresponding to sparse gauge measurements. These observations form a discrete sample set $\mathcal{S} = \{(\mathbf{c}_i, f(\mathbf{c}_i))\}_{i=1}^N$. We assume that f lies in a Sobolev space $H^s(D)$, which consists of functions in $L^2(D)$ whose weak derivatives up to order s are also square-integrable:

$$H^s(D) = \left\{ f \in L^2(D) \mid \partial^\alpha f \in L^2(D),\ \forall |\alpha| \le s \right\},$$

where $s > d/2$ (with $d = 2$ in our case). This condition ensures that f is sufficiently smooth. Moreover, we assume that f is approximately band-limited in the spectral domain; that is, its Fourier transform $\hat{f}(\boldsymbol{\omega})$ satisfies $\hat{f}(\boldsymbol{\omega}) \approx 0$ for $\|\boldsymbol{\omega}\| > \Omega$, for some cut-off

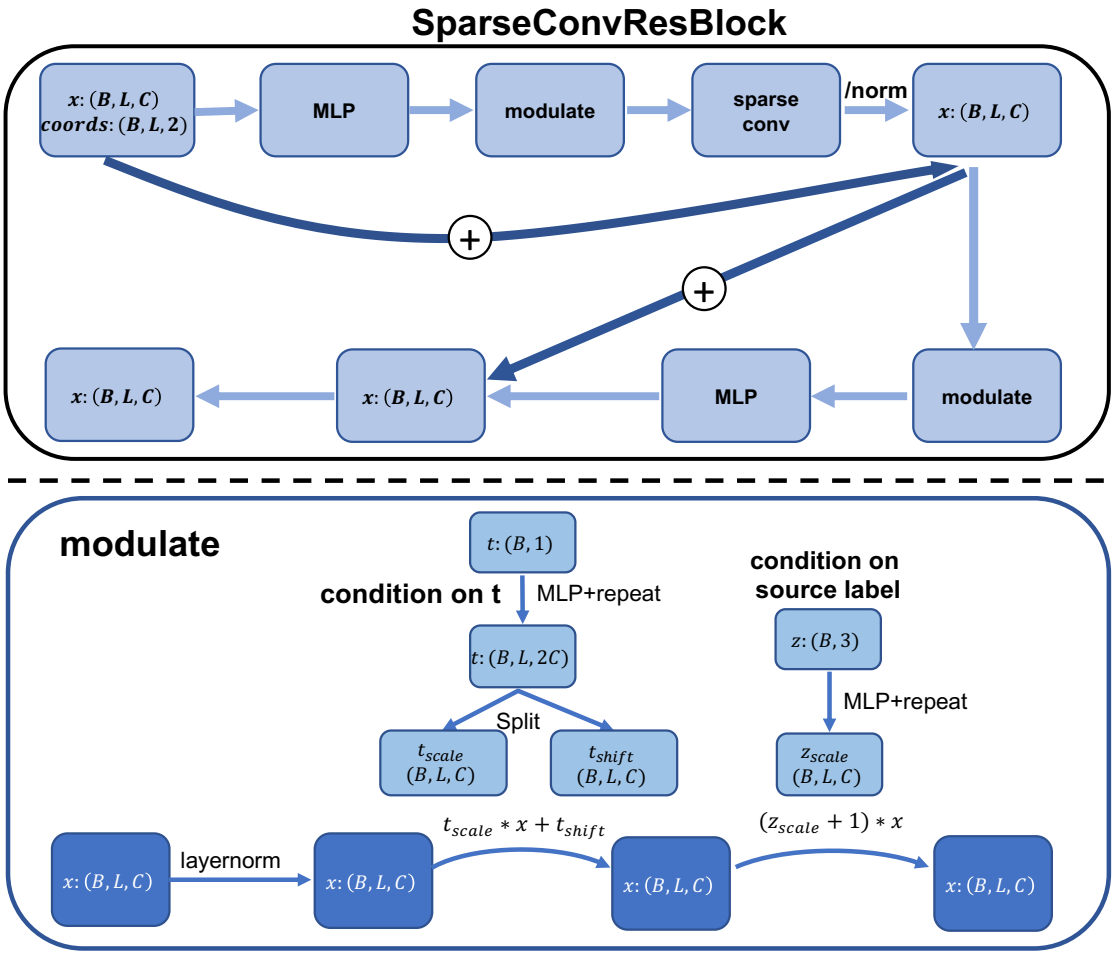

Fig. B3: **Architecture of the SparseConvResBlock and its modulation mechanism.** The top panel illustrates the overall structure of the *SparseConvResBlock*, which processes input features $x \in \mathbb{R}^{B\times L\times C}$ and associated coordinates $\mathbb{R}^{B\times L\times 2}$ through a residual block comprising one sparse depthwise convolution. The term *norm* shares the same shape as $x \in \mathbb{R}^{B\times L\times C}$ and is precomputed by convolving a unit-valued sparse tensor with fixed, non-trainable weights. The convolution kernel is initialized as a uniform averaging filter, where each weight is set to $1/K^2$ (K refers the kernel size). This operation estimates local support density, ensuring numerical stability. The bottom panel shows the internal design of the *modulate* module, which conditions the representation on two external variables: the diffusion timestep $t \in \mathbb{R}^{B\times 1}$ and the source label $z \in \mathbb{R}^{B\times 3}$ that denotes the dataset identity. Specifically, the source label is a one-hot vector indicating the origin of each sample, where ERA5 is represented as $(1, 0, 0)$, IMERG as $(0, 1, 0)$, and gauge observations as $(0, 0, 1)$. This encoding enables the model to learn dataset-specific feature modulations while maintaining a unified architecture across heterogeneous data sources. The timestep embedding is transformed by an MLP and split into scaling (t_{scale}) and shifting (t_{shift}) components, applied to the normalized input. Simultaneously, the source label contributes a scaling factor z_{scale} that further modulates the representation. This dual conditioning enables flexible control over the representation across both temporal and domain dimensions.

frequency $\Omega > 0$. This implies that the function's energy is primarily concentrated in a bounded low-frequency range. Consequently, even under sparse sampling, the dominant frequency characteristics of f are preserved, especially the low-frequency content that encodes large-scale spatial structure.

Traditional convolutional neural networks rely on regular grids, which impose translation-equivariant operations on dense Euclidean tensors. In contrast, sparse convolutional networks define a convolution operator K_θ directly on the irregular domain $\mathcal{C}$ without requiring interpolation or resampling to a dense grid. The sparse convolution operator K_θ is designed to operate on non-uniform point clouds by convolving features over local neighborhoods defined on the support $\mathcal{C}$. Given features h defined at sparse locations, the sparse convolution updates features by aggregating information from neighboring points via: $(K_\theta h)(\mathbf{c}_i) = \sum_{\mathbf{c}_j \in \mathcal{N}(\mathbf{c}_i)} \kappa_\theta(\mathbf{c}_j - \mathbf{c}_i) \cdot h(\mathbf{c}_j)$. Here $\mathcal{N}(\mathbf{c}_i)$ denotes the receptive field around $\mathbf{c}_i$, and κ_θ are learnable kernel weights that depend on relative spatial coordinates. This formulation naturally adapts to the irregular geometry of gauge networks and preserves local spatial relationships without imposing artificial gridding.

From a mathematical standpoint, the convolution operation can be interpreted as a linear integral operator acting on the input function f. By the convolution theorem, applying a spatial convolution is equivalent to performing a pointwise multiplication in the frequency domain: $f * \kappa \longleftrightarrow \hat{f}(\boldsymbol{\omega}) \cdot \hat{\kappa}(\boldsymbol{\omega})$ where $\hat{f}$ and $\hat{\kappa}$ denote the Fourier transforms of f and the kernel κ, respectively. This identity implies that convolutional neural networks fundamentally implement structured linear operators in the spectral domain, modulated by nonlinear activations in the spatial domain. In our setting, we consider a field f that is band-limited and belongs to a Sobolev space $H^s(D)$. The band-limited assumption ensures that the energy of $\hat{f}(\boldsymbol{\omega})$ is concentrated within a compact subset of the frequency domain. Furthermore, since the gauge observations $\mathcal{S} = \{(\mathbf{c}_i, f(\mathbf{c}_i))\}_{i=1}^N$ are assumed to sample f in a non-pathological manner—that is, the sampling locations $\{\mathbf{c}_i\}$ are well-distributed across the domain and do not systematically avoid critical regions—the low-frequency components of f are approximately preserved under such sparse sampling schemes. This stability implies that a sparse convolutional operator K_θ, though defined over an irregular set $\mathcal{C}$, still implements a meaningful approximation of the frequency-domain filtering process. Therefore, the sparse convolution module can be regarded as a discretized, sampling-robust spectral operator, providing the mathematical basis for its application to be a core module in PRIMER.

B.7 Choice of experiment parameter

PRIMER is built using the PyTorch framework [86]. We summarize here the key hyper-parameters used during training and inference. Due to the computational cost associated with training PRIMER, we did not perform an extensive hyper-parameter search. Instead, all values were chosen based on empirical experience. We expect that further tuning may yield improved performance. Notably, training was intermittently paused and resumed multiple times, during which model weights were checkpointed and certain hyper-parameters were adjusted to optimize convergence. Further systematic tuning may still improve overall performance.

Appendix C Details of data

C.1 Locations of gauges

C.2 Test set

The 150 representative precipitation events used for evaluation in Fig. 4 and Fig. 5 were carefully selected from hourly gauge observations collected across the study domain throughout 2016. At each timestamp, approximately 1,000 stations provided precipitation

Table B1: **Key parameters used in this study.** All values are empirically chosen without hyperparameter search.

Parameter	Description / Value
OS	Linux-5.10.0-34-cloud-amd64-x86_64-with-glibc2.31
Python version	3.10.0
GPU count	2
GPU type	NVIDIA A100-SXM4-40GB
CUDA version	11.7
Overall parameters	430,058,544 trainable parameters
diffusion_steps	1000
AdamW optimizer settings	
beta1	0.9 (1st moment decay rate)
beta2	0.99 (2nd moment decay rate)
weight_decay	4e-6
EMA (Exponential Moving Average) settings	
decay	0.995
update_every	every 10 batches
Batch size	Varied between 2–6 (per GPU) due to intermittent training interruptions
Learning rate	Varied between 10^{-4} and 10^{-6} due to intermittent training interruptions

measurements. To ensure a robust and representative test dataset, we employed two complementary intensity-based selection criteria. First, we identified the 100 timestamps exhibiting the highest individual station precipitation intensities, specifically highlighting localized extreme events. Second, we selected 50 additional timestamps characterized by the highest average precipitation intensity across all stations, thus capturing widespread precipitation scenarios. These distinct yet complementary selection strategies ensure comprehensive coverage of heavy precipitation event types, enhancing the generalizability and reliability of our model evaluations. Notably, there was no overlap between these two subsets, resulting in a final, unique set of 150 precipitation events.

The test set includes the following 150 timestamps (formatted as YYYYMMDDHH): 2016070822, 2016070821, 2016071704, 2016070820, 2016061323, 2016071907, 2016081020, 2016061902, 2016072616, 2016050923, 2016040316, 2016062312, 2016070317, 2016042000, 2016071108, 2016072507, 2016082715, 2016071910, 2016071814, 2016062419, 2016080214, 2016071813, 2016070522, 2016091212, 2016050922, 2016062313, 2016070101, 2016050921, 2016071912, 2016070521, 2016061501, 2016081019, 2016081821, 2016061901, 2016072421, 2016082506, 2016061322, 2016071915, 2016071709, 2016071914, 2016070919, 2016060119, 2016071909, 2016091420, 2016060120, 2016071901, 2016071107, 2016071904, 2016072611, 2016070716, 2016062310, 2016062200, 2016080216, 2016090916, 2016060611, 2016071900, 2016080307, 2016080923, 2016082714, 2016070819, 2016081008, 2016062302, 2016080200, 2016061214, 2016060612, 2016081021, 2016062304, 2016050607, 2016060414, 2016070102, 2016091112, 2016062201, 2016071908, 2016072121, 2016071913, 2016070405, 2016071521, 2016082516, 2016070323, 2016062718, 2016072612, 2016061104, 2016070400, 2016070823, 2016061913, 2016071312, 2016052005, 2016082517, 2016071702, 2016092808, 2016072122, 2016091600, 2016080509, 2016061903, 2016061820, 2016062005, 2016051423, 2016052002, 2016070519, 2016062802, 2016071406, 2016102019, 2016071407, 2016102018, 2016102021, 2016102020, 2016052218, 2016052104, 2016071405, 2016071408, 2016102023, 2016102022, 2016052217, 2016052219, 2016071409, 2016102100, 2016061108, 2016102117, 2016052220,

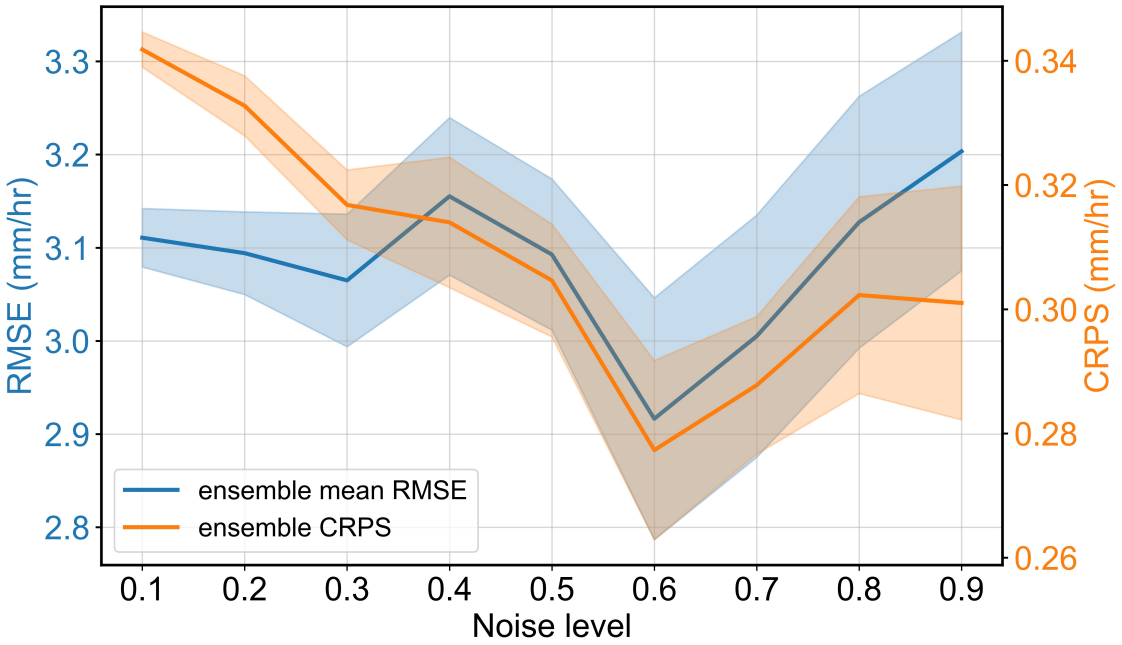

Fig. B4: **Sensitivity of ensemble RMSE and CRPS to the noise level parameter τ.** Evaluation is performed across a range of noise levels from 0.1 to 0.9. Ensemble members are sampled from $P_*(x \mid \mathcal{O}_{\text{IMERG}})$, and both the ensemble-mean root mean square error (RMSE; blue) and the continuous ranked probability score (CRPS; orange) are computed over 50 repeated subsampling trials, each using 10 randomly selected members. Shaded bands denote ± 1 standard deviation across repetitions. Both metrics show improvement as τ increases up to 0.6, reflecting a favorable balance between accuracy and diversity, but deteriorate beyond this point due to excessive stochasticity. These results support the choice of an intermediate noise level to balance observational fidelity with generative variability.

2016110719, 2016102015, 2016102118, 2016061109, 2016052211, 2016071106, 2016012817, 2016110720, 2016102121, 2016061107, 2016012816, 2016052103, 2016012803, 2016102120, 2016112305, 2016013112, 2016052023, 2016061111, 2016052100, 2016110823, 2016012815, 2016102122, 2016012813, 2016102119, 2016100622, 2016071410, 2016112220, 2016102103, 2016102102, 2016102101, 2016102116.

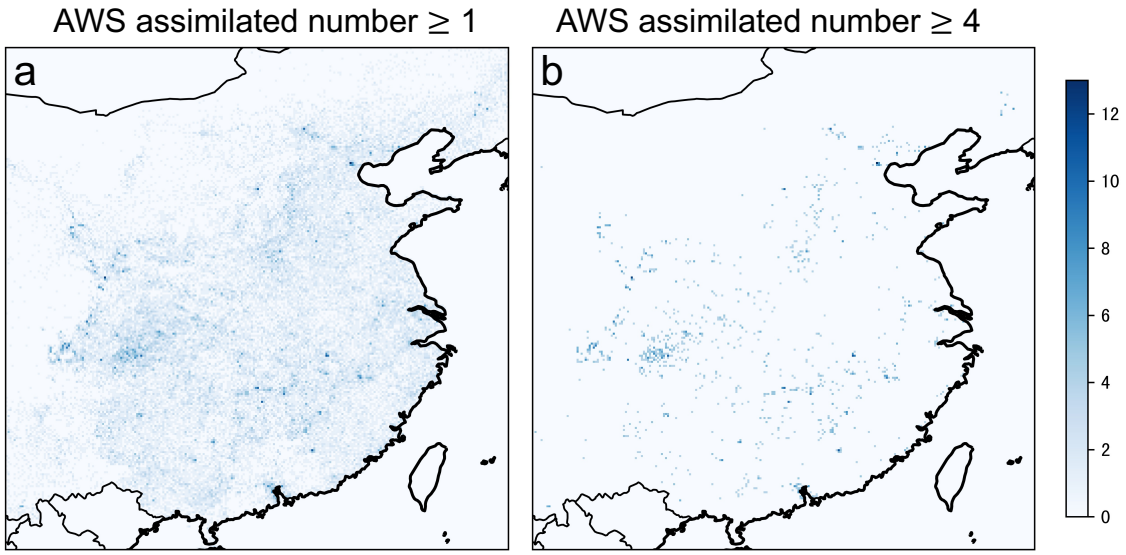

Fig. C5: **Spatial distribution of assimilated AWS counts.** Panels show the number of automatic weather stations (AWS) assimilated into each grid point of the high spatiotemporal gauge-satellite merged precipitation analysis from [29]. a, Grid points with at least one assimilated AWS observation, representing the full set of available data used for training. b, A more stringent subset showing grid points with four or more assimilated AWS observations, used exclusively for model evaluation. This design ensures that the test regions are better constrained by in-situ observations, enabling a robust assessment of model performance.

Appendix D Additional results

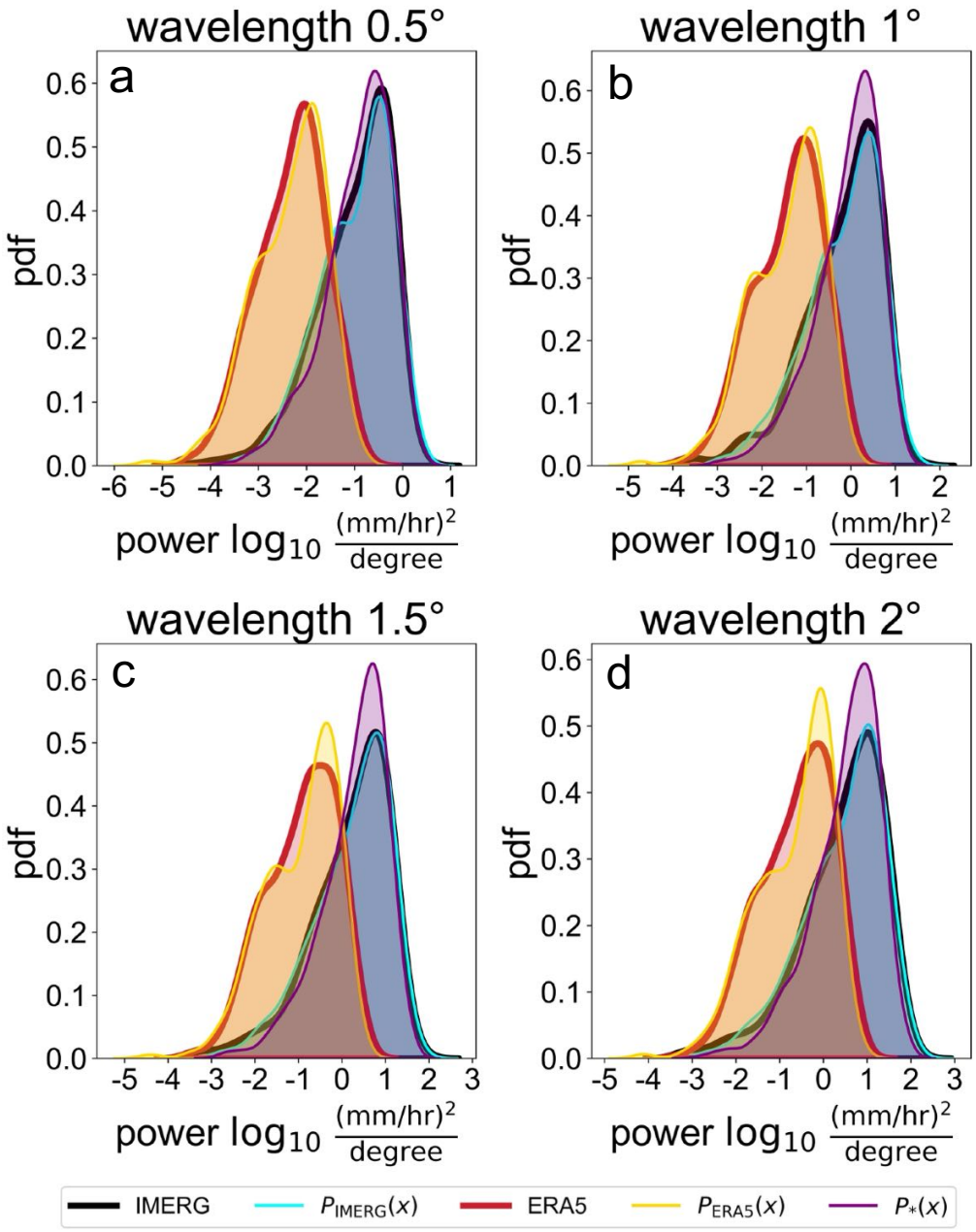

Fig. D6: **PDF of RAPSD between learned priors and reference datasets.** (**a-d**) shows the PDF of power at 0.5°, 1°, 1.5°, 2° wavelength respectively. All statistics are derived from 1,000 randomly sampled realizations of precipitation fields.

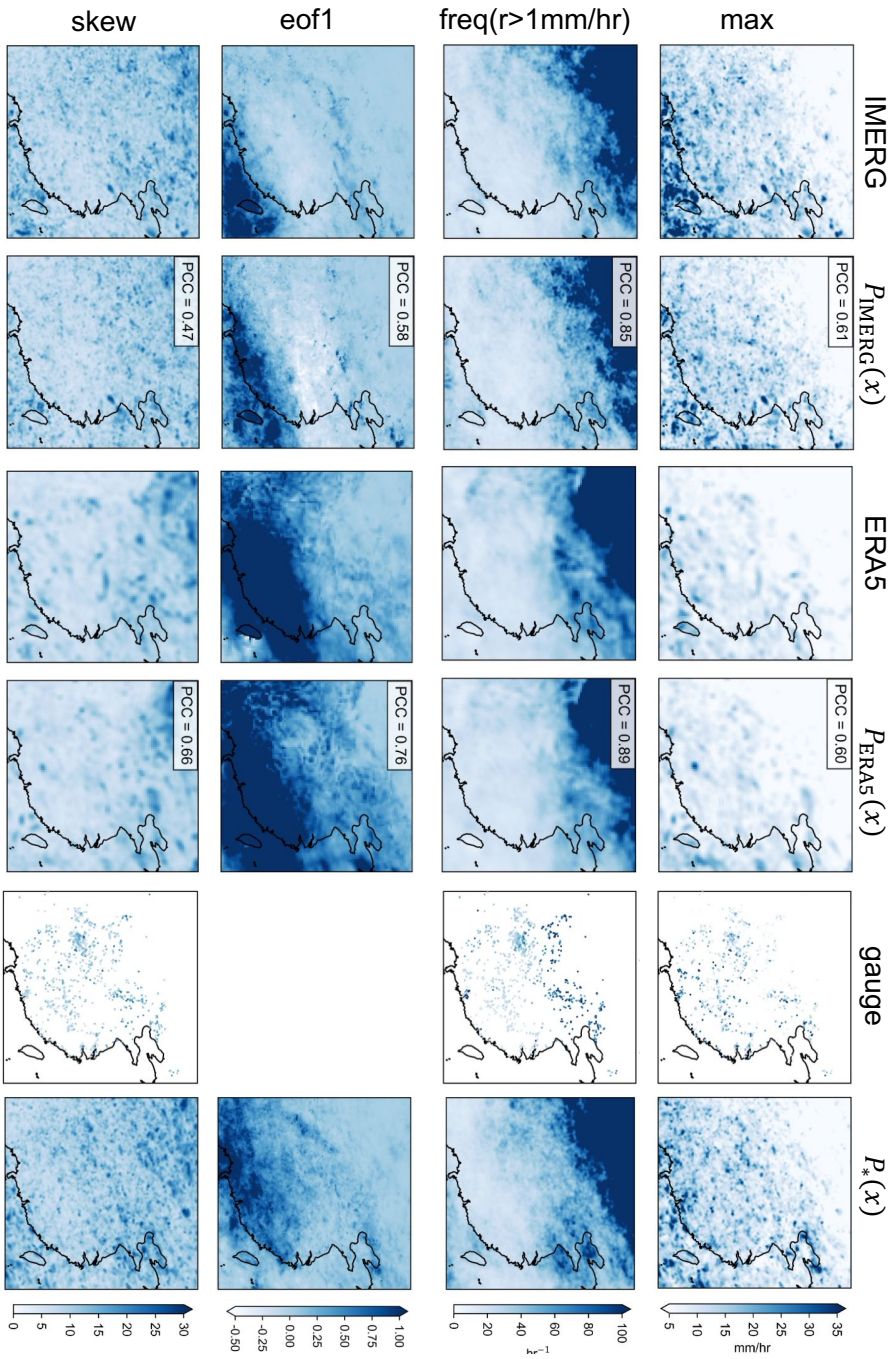

Fig. D7: **Climatological structure comparison between learned priors and reference datasets.** For clarity, the figure has been rotated 90° clockwise. Each row presents a distinct climatological statistic: **top to bottom**, spatial distribution of maximum precipitation rate, frequency of precipitation events (> 1 mm/hr), the leading empirical orthogonal function (EOF1), and skewness. Each column corresponds to a different data source: IMERG, unconditional samples from $P_{\mathrm{IMERG}}(x)$, ERA5, unconditional samples from $P_{\mathrm{ERA5}}(x)$, gauge observations, and samples from the final updated prior $P_*(x)$. Panels associated with $P_{\mathrm{IMERG}}(x)$ and $P_{\mathrm{ERA5}}(x)$ display Pearson correlation coefficients (PCCs) with their respective reference datasets (IMERG and ERA5), highlighting structural agreement. Colorbars denote the units of each diagnostic. All statistics are derived from 1,000 randomly sampled realizations of precipitation fields.

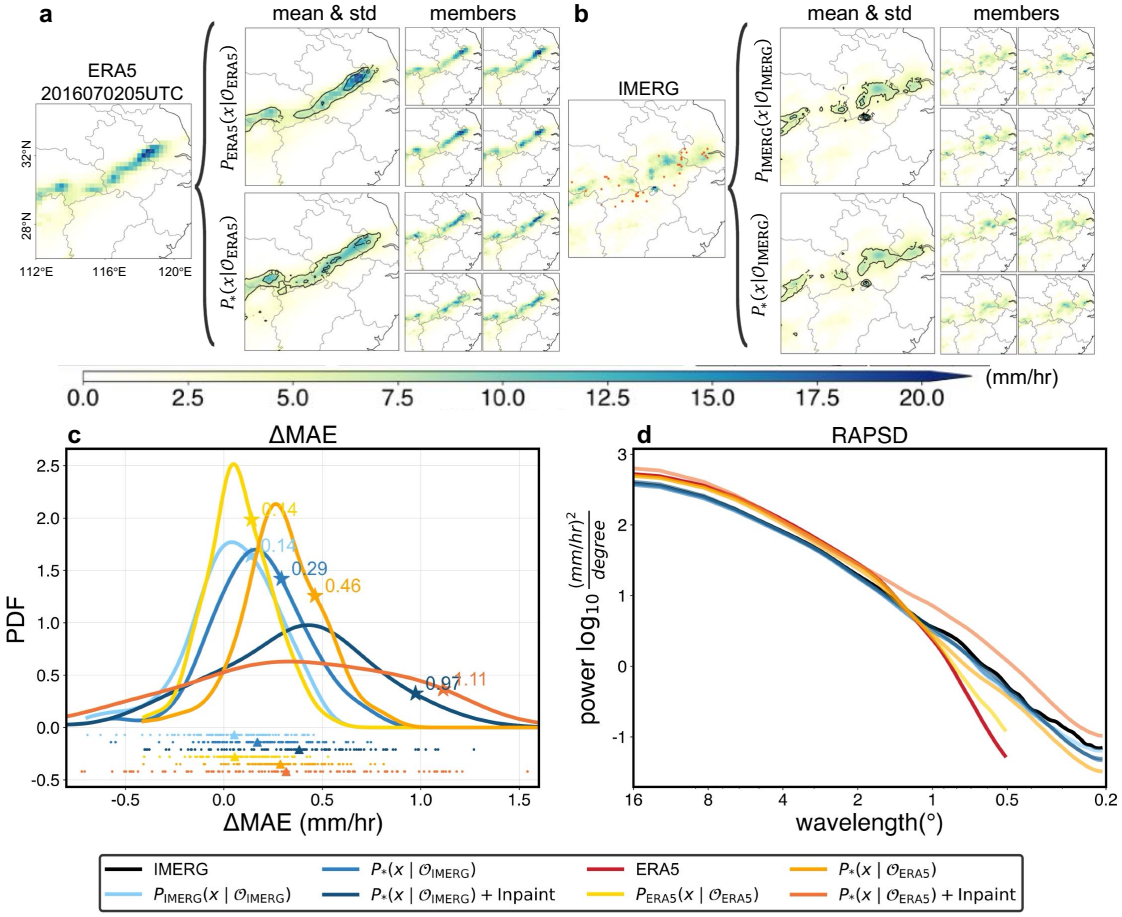

Fig. D8: **Case study of a Meiyu event.** This figure complements Fig. 3 by illustrating additional results sampled from alternative posterior distributions. **a,b**, Posterior samples based on ERA5 and IMERG as conditional input, respectively. We show the original precipitation field, the posterior mean and standard deviation, and four representative ensemble members. **c**, Probability density functions (PDFs) of changes in mean absolute error (ΔMAE) relative to original ERA5 or IMERG, with positive values indicating improved accuracy after bias correction. **d**, Radially averaged power spectral density (RAPSD) curves demonstrate that prior $P_*(x)$ effectively compensates for the underestimation of high-frequency spectral power in ERA5, thereby enhancing spatial structure realism. Note that we first interpolate ERA5 and samples from $P_{\mathrm{ERA5}}(x \mid \mathcal{O}_{\mathrm{ERA5}})$ to 0.1 degree before RAPSD caculation. Overall, this case highlights the flexibility of PRIMER in performing posterior sampling using diverse precipitation priors, including those derived from reanalysis, satellite. Among them, the prior $P_*(x)$ yields the most accurate reconstructions, underscoring the value of incorporating sparse yet reliable gauge observations for fine-tuning probabilistic models.

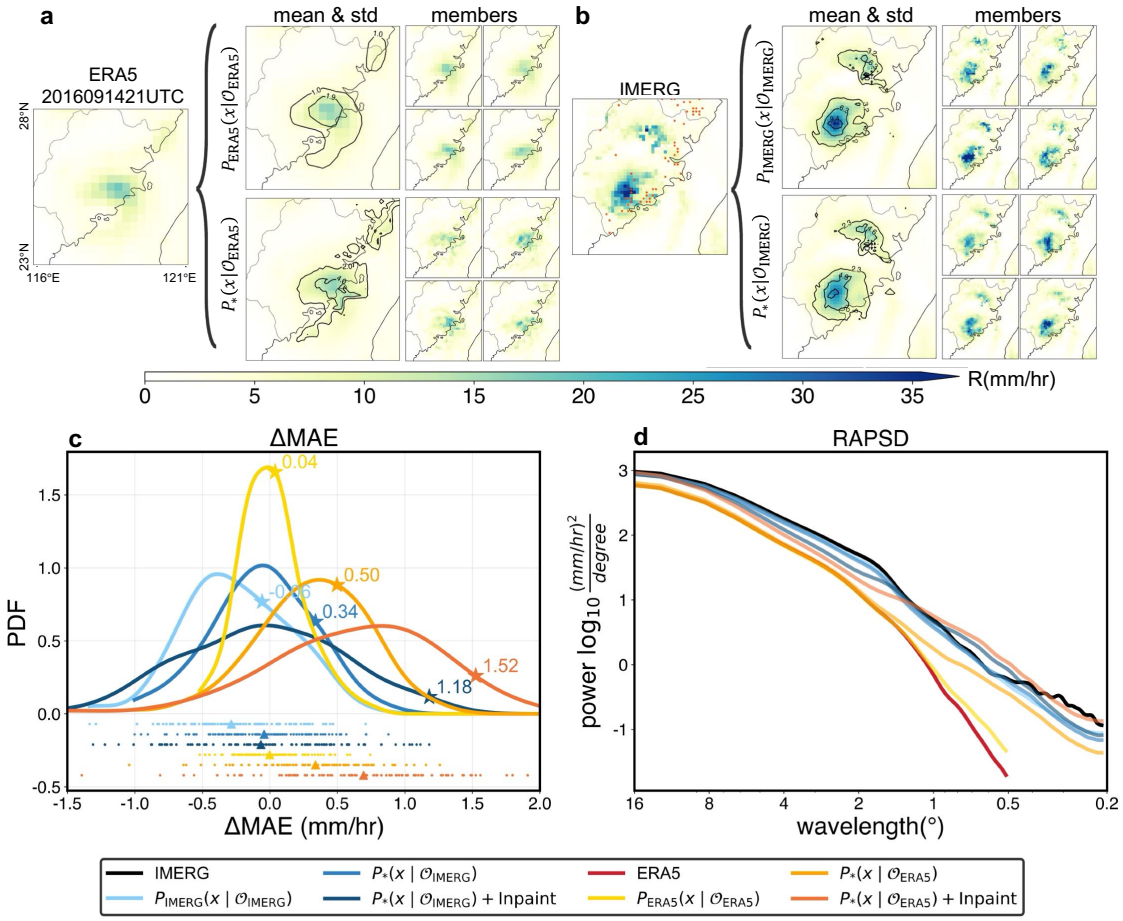

Fig. D9: **Case study of Typhoon Meranti (2016) precipitation event.** Typhoon Meranti, one of the most intense tropical cyclones recorded globally in 2016, made landfall in southeastern China in mid-September, causing widespread flooding and infrastructure damage. This figure is similar to Fig. D8.

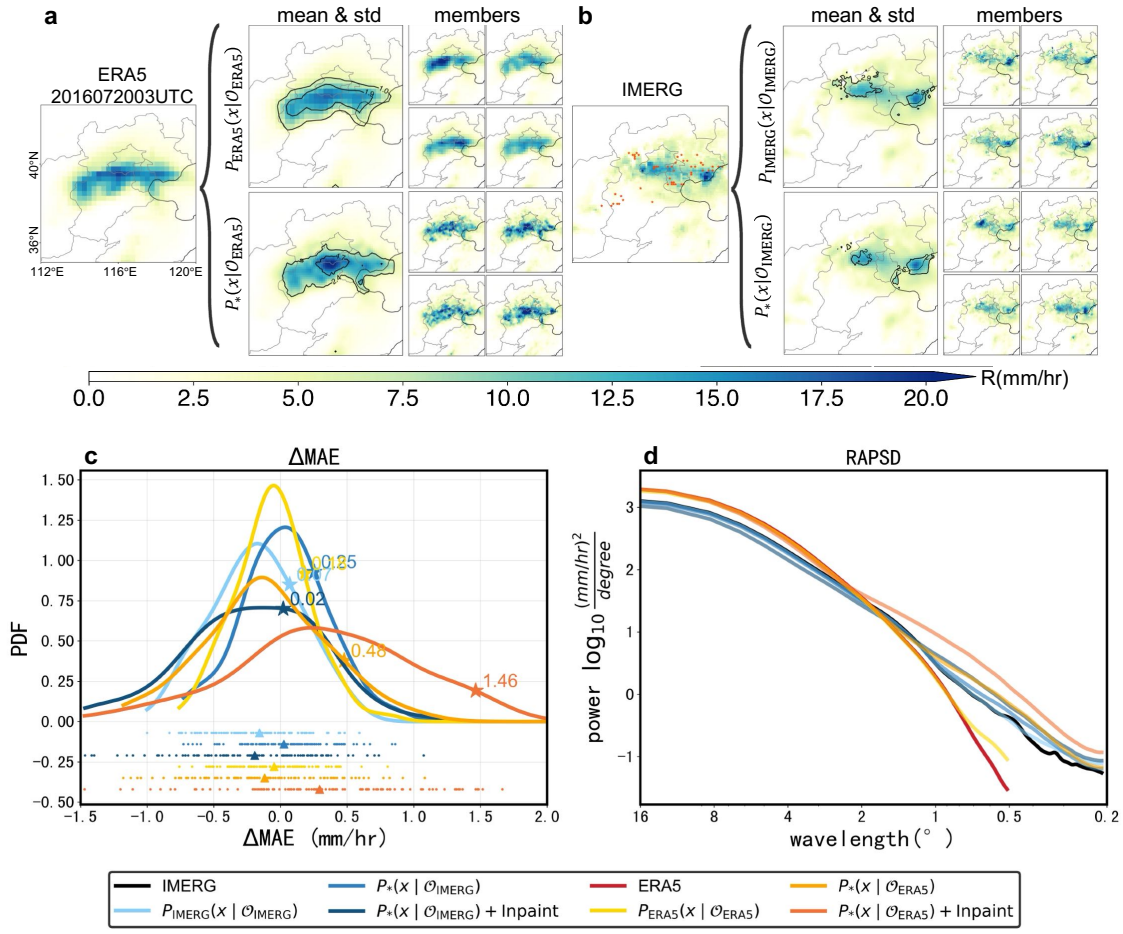

Fig. D10: **Case study of an extreme precipitation event near Beijing.** On 20 July 2016, an extratropical cyclone developed over North China, bringing prolonged and intense precipitation to the Beijing-Tianjin-Hebei region. This event, known as the "7·20" rainstorm. This figure is similar to Fig. D8.

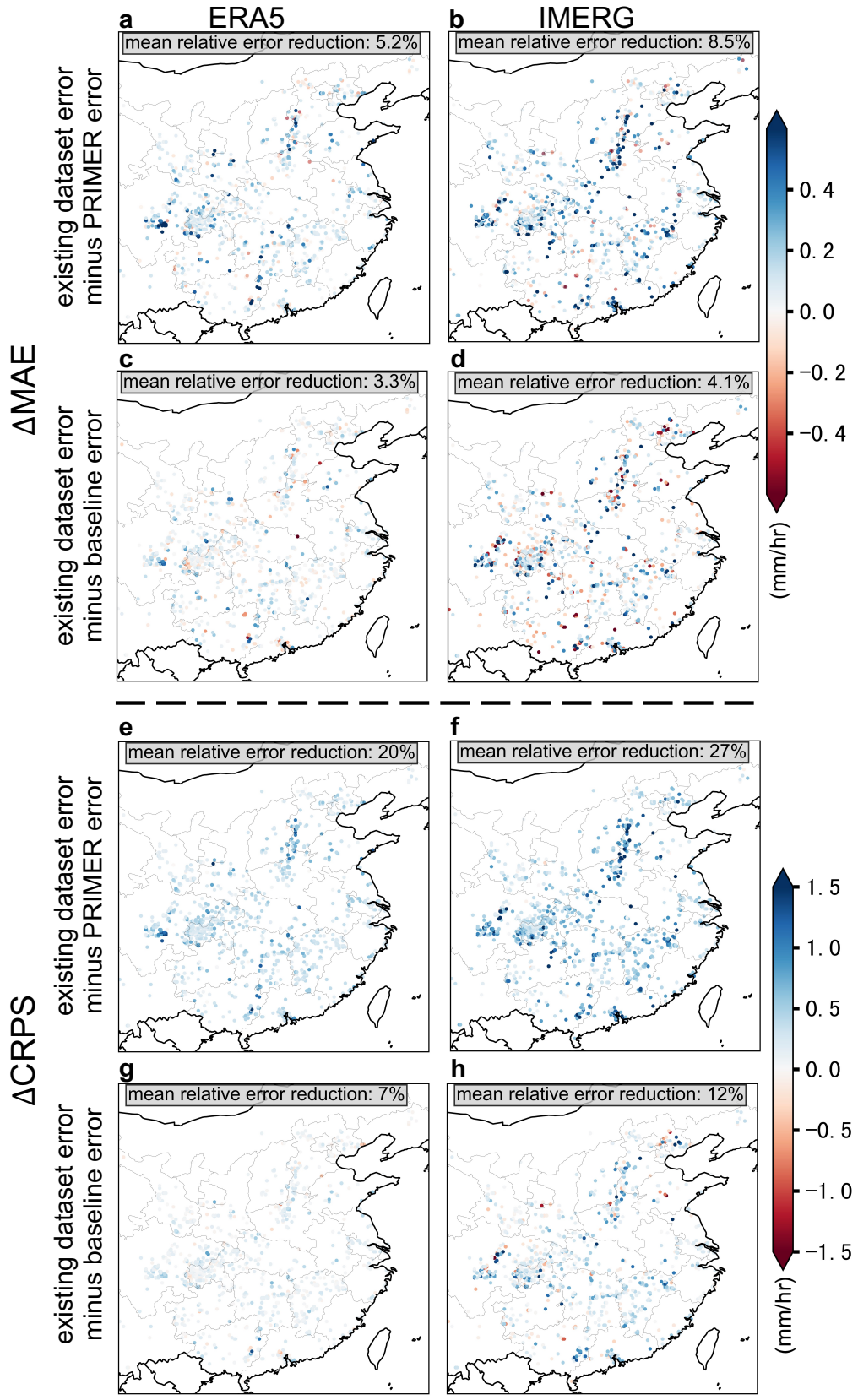

Fig. D11: **Spatial distributions of** ΔMAE **and** ΔCRPS**.** As in Fig. 4, but showing the reduction in MAE and CRPS after bias correction using PRIMER (prior $P_*(x)$) and the baseline priors ($P_{\text{ERA5}}(x)$ and $P_{\text{IMERG}}(x)$), applied separately to ERA5 and IMERG. The evaluation is based on 150 precipitation events that occurred in 2016. Overall, PRIMER outperforms the baseline method, as evidenced by larger mean relative error reductions (annotated in the top-left corner of each panel).

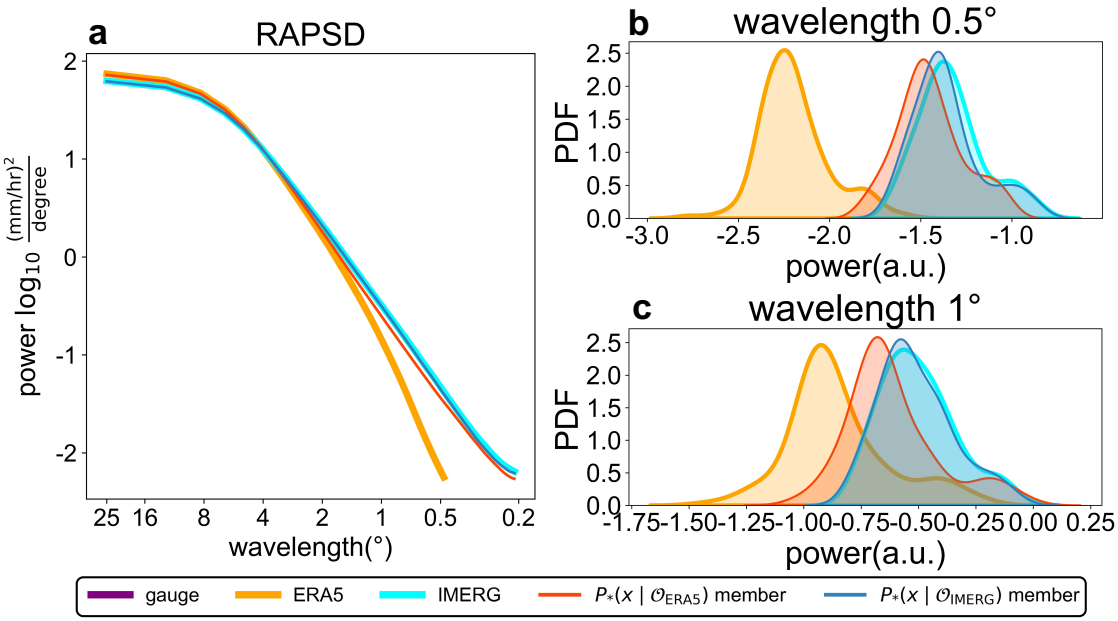

Fig. D12: **Enhancement of spatial variability in test datasets.** Due to the lack of power spectral references based on gauge observations, IMERG (0.1° resolution) is used as a proxy for evaluating fine-scale precipitation features. **a**, Radially averaged power spectral density (RAPSD) of log-transformed precipitation intensity, showing that PRIMER effectively restores high-frequency variability absent in the original ERA5 data. For consistency, ERA5 data are interpolated to a 0.1° grid prior to RAPSD computation. **b**, Probability density functions (PDFs) of spectral power at wavelengths of 0.5° (top) and 1° (bottom). While ERA5 (0.25° resolution) underrepresents spectral power at these smaller scales, samples that are generated from posteriors $P_*(x \mid \mathcal{O}_{\mathrm{ERA5}})$ shift the distribution toward higher power, indicating improved representation of fine-scale structure.

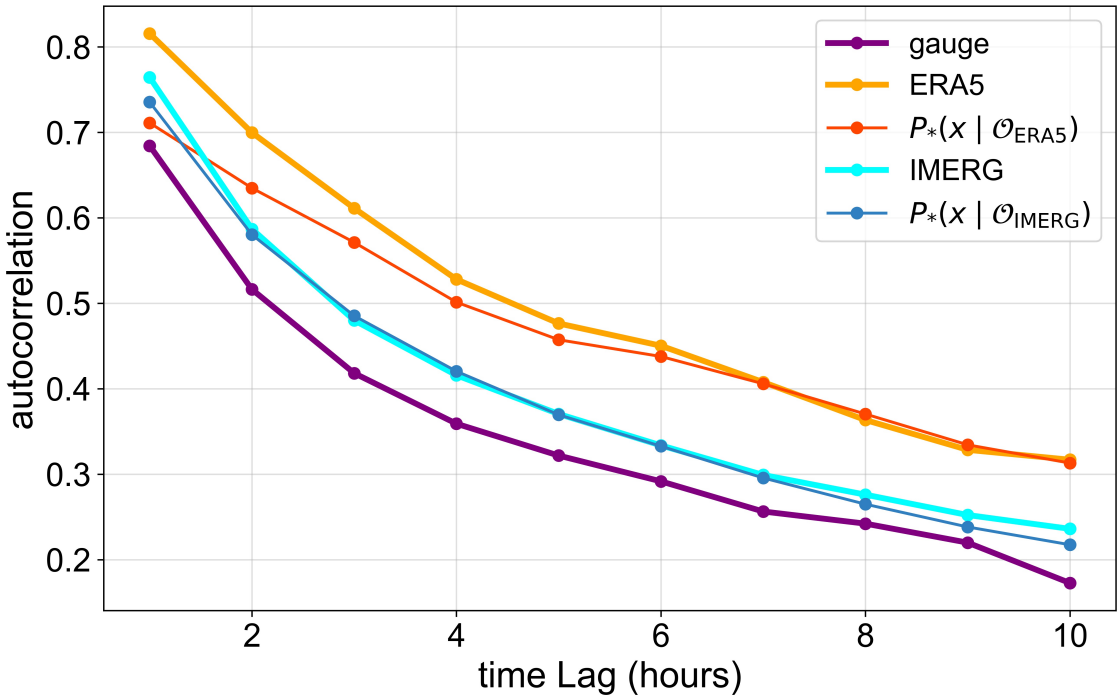

Fig. D13: **Temporal correlations in test datasets.** As a supplementary analysis to Fig. 5, temporal correlations are assessed by computing the autocorrelation with a lag of up to 10 hours. Correlations are computed exclusively at gauge locations and averaged over all paired precipitation events from a subset of 2016, for gauge observations, ERA5, $P_*(x \mid \mathcal{O}_{\text{ERA5}})$, IMERG, and $P_*(x \mid \mathcal{O}_{\text{IMERG}})$. Results show that applying PRIMER to ERA5 and IMERG preserves the intrinsic temporal dynamics, as evidenced by comparable autocorrelation structures before and after correction. Notably, original ERA5 and IMERG exhibit higher temporal correlations than gauge observations, reflecting artificial persistence introduced probably by numerical model, data assimilation and satellite retrieval processes. After PRIMER mollification, the temporal correlations of $P_*(x \mid \mathcal{O}_{\text{ERA5}})$ and $P_*(x \mid \mathcal{O}_{\text{IMERG}})$ decrease and become closer to those observed in the gauge observations, indicating improved physical realism.

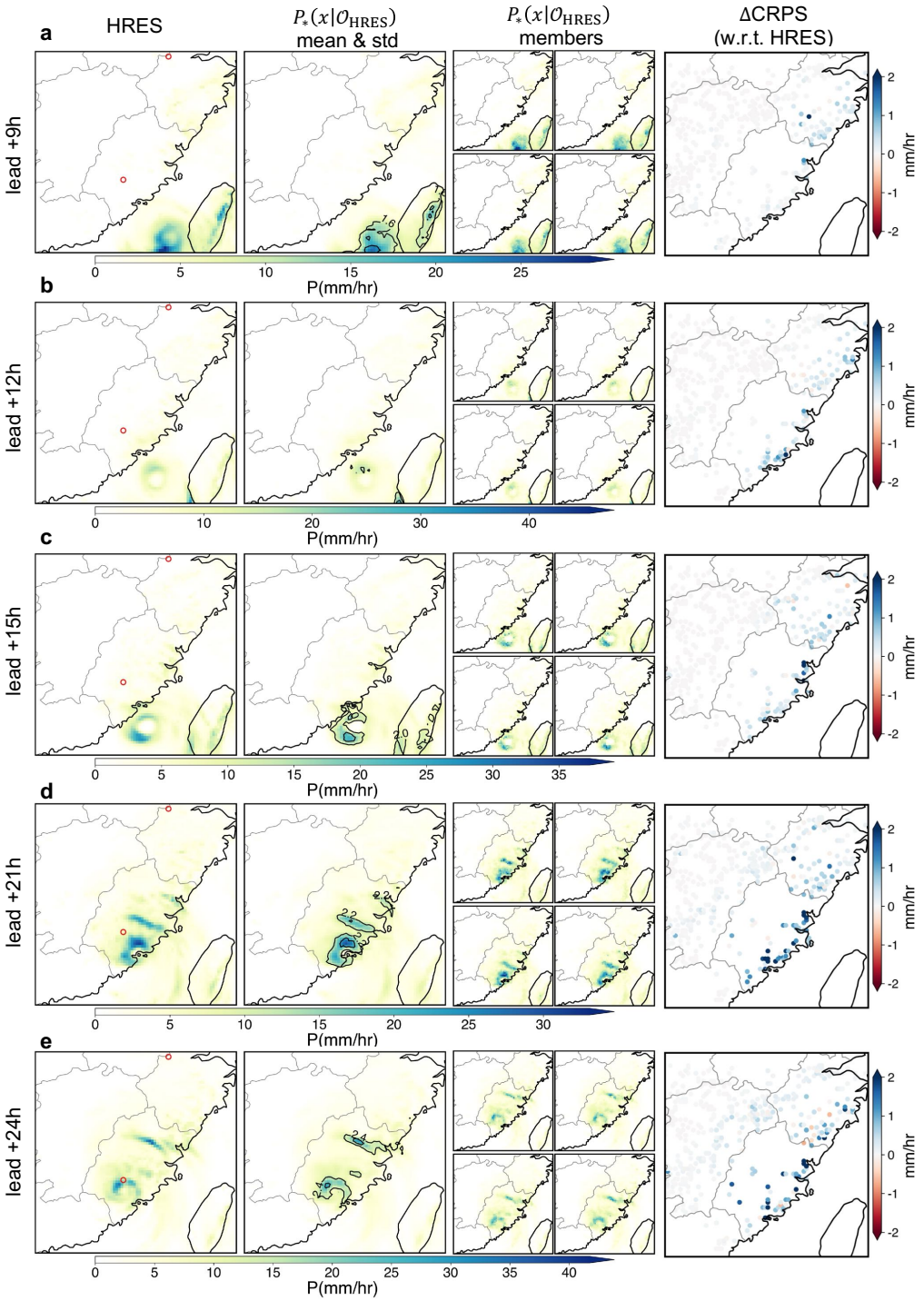

Fig. D14: **Zero-shot enhancement for operational forecasts.** Similar to Fig. 6, but for other lead times with a comprehensive illustration.

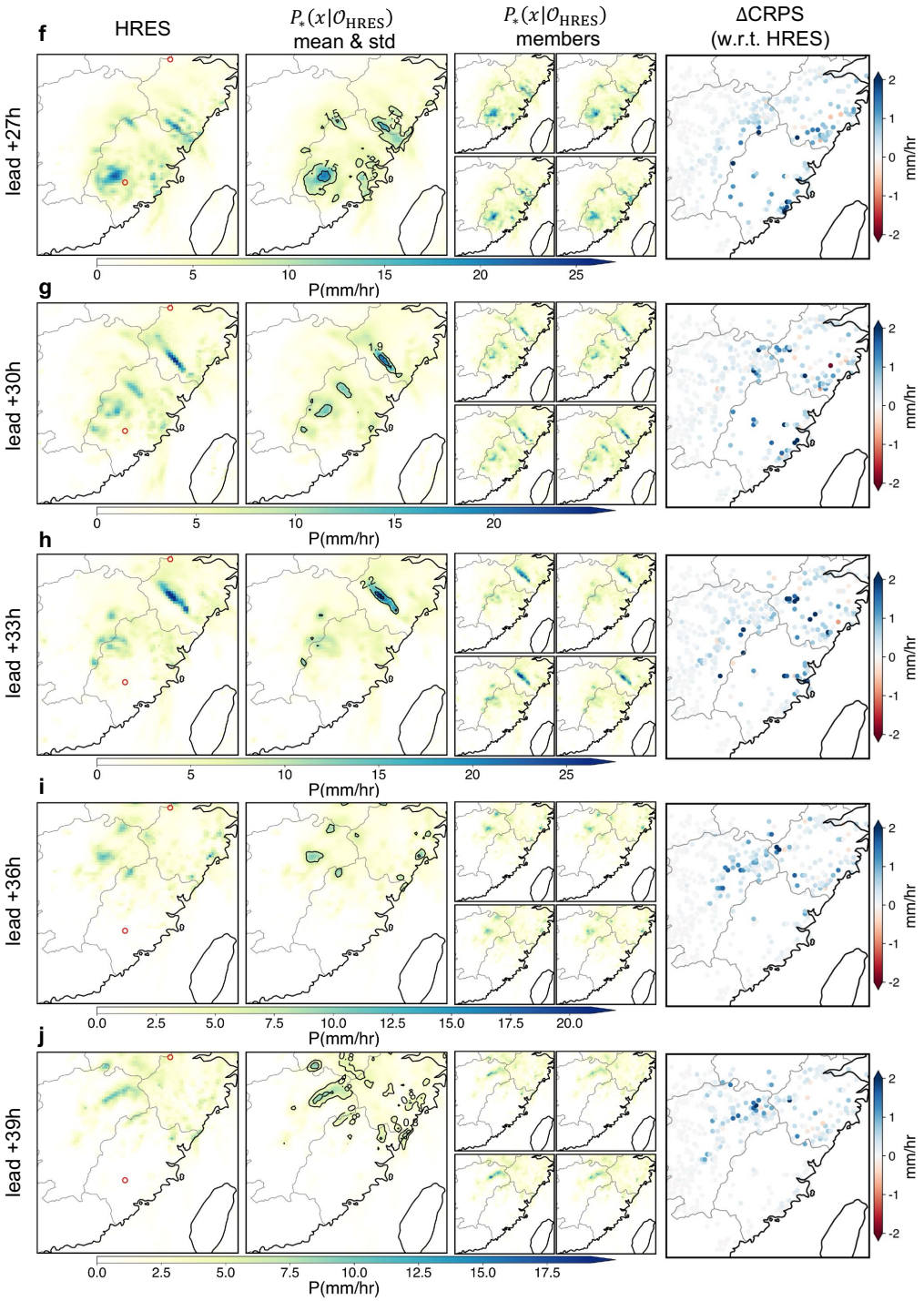

Fig. D14: Continued from previous page.

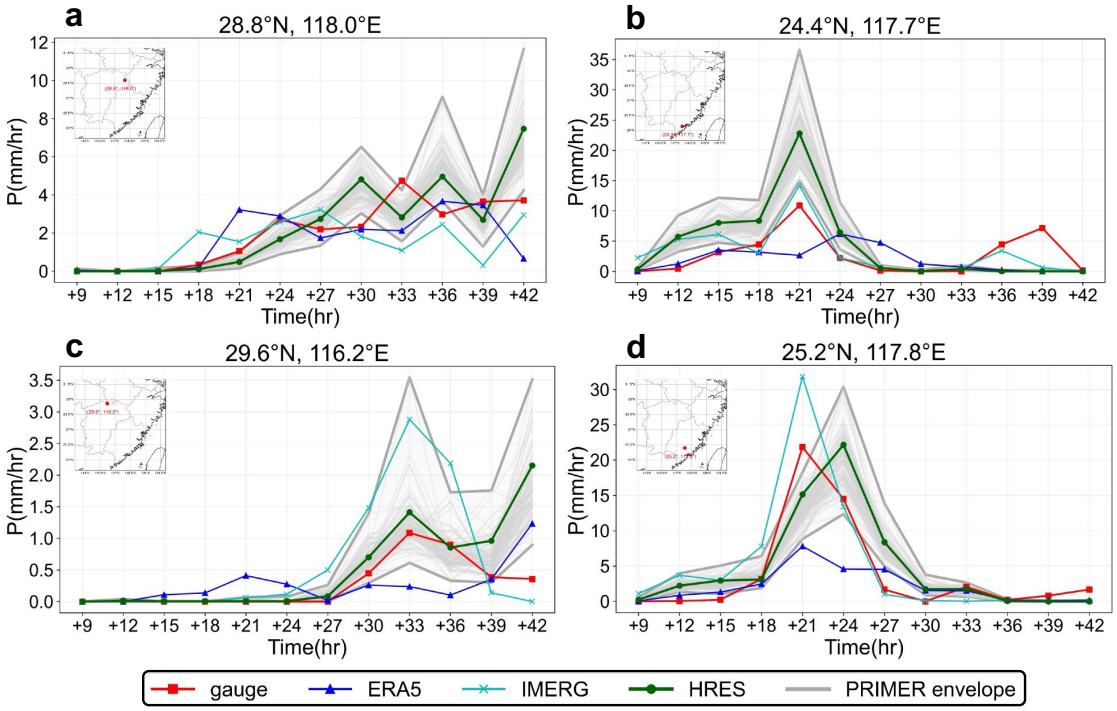

Fig. D15: **Zero-shot enhancement for operational forecasts.** Similar to Fig. 6, but for precipitation time series at other representative gauge stations; gray envelope denotes the spread across 100 ensemble members.

References

[1] Kotz, M., Levermann, A. & Wenz, L. The effect of rainfall changes on economic production. *Nature* **601**, 223–227 (2022).

[2] Sun, Y., Solomon, S., Dai, A. & Portmann, R. W. How often does it rain? *Journal of climate* **19**, 916–934 (2006).

[3] Pendergrass, A. G. & Knutti, R. The uneven nature of daily precipitation and its change. *Geophysical Research Letters* **45**, 11–980 (2018).

[4] Stevens, B. & Feingold, G. Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature* **461**, 607–613 (2009).

[5] Birch, C. *et al.* Impact of soil moisture and convectively generated waves on the initiation of a west african mesoscale convective system. *Quarterly Journal of the Royal Meteorological Society* **139**, 1712–1730 (2013).

[6] Prein, A. F., Mooney, P. A. & Done, J. M. The multi-scale interactions of atmospheric phenomenon in mean and extreme precipitation. *Earth's Future* **11**, e2023EF003534 (2023).

[7] Teixeira, J. *et al.* Parameterization of the atmospheric boundary layer: a view from just above the inversion. *Bulletin of the American Meteorological Society* **89**, 453–458 (2008).

[8] Lepore, C., Veneziano, D. & Molini, A. Temperature and cape dependence of rainfall extremes in the eastern united states. *Geophysical Research Letters* **42**, 74–83 (2015).

[9] Arakawa, A. The cumulus parameterization problem: Past, present, and future. *Journal of climate* **17**, 2493–2525 (2004).

[10] Houze Jr, R. A. Mesoscale convective systems. *Reviews of Geophysics* **42** (2004).

[11] Sun, Q. *et al.* A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics* **56**, 79–107 (2018).

[12] Kidd, C. & Huffman, G. Global precipitation measurement. *Meteorological Applications* **18**, 334–353 (2011). URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.284.

[13] Hou, A. Y. *et al.* The global precipitation measurement mission. *Bulletin of the American meteorological Society* **95**, 701–722 (2014).

[14] Levizzani, V., Amorati, R. & Meneguzzo, F. A review of satellite-based rainfall estimation methods. *European Commission Project MUSIC Report (EVK1-CT-2000-00058)* **66** (2002).

[15] Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).

[16] Tapiador, F. J. *et al.* Is precipitation a good metric for model performance? *Bulletin of the American Meteorological Society* **100**, 223–233 (2019).

[17] He, J. *et al.* The first high-resolution meteorological forcing dataset for land process studies over China. *Scientific data* **7**, 25 (2020).

[18] Ma, Y. *et al.* Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. *Journal of Geophysical Research: Atmospheres* **123**, 814–834 (2018).

[19] Baez-Villanueva, O. M. *et al.* Rf-mep: A novel random forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment* **239**, 111606 (2020).

[20] Ur Rahman, K., Shang, S., Shahid, M. & Wen, Y. An appraisal of dynamic Bayesian model averaging-based merged multi-satellite precipitation datasets over complex topography and the diverse climate of Pakistan. *Remote Sensing* **12**, 10 (2019).

[21] Yumnam, K., Guntu, R. K., Rathinasamy, M. & Agarwal, A. Quantile-based Bayesian model averaging approach towards merging of precipitation products. *Journal of Hydrology* **604**, 127206 (2022).

[22] Xie, P. & Xiong, A.-Y. A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *Journal of Geophysical Research: Atmospheres* **116** (2011).

[23] Woldemeskel, F. M., Sivakumar, B. & Sharma, A. Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. *Journal of Hydrology* **499**, 167–176 (2013).

[24] Fan, Z. *et al.* A comparative study of four merging approaches for regional precipitation estimation. *IEEE Access* **9**, 33625–33637 (2021).

[25] Zhang, L. *et al.* Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *Journal of Hydrology* **594**, 125969 (2021).

[26] Bhuiyan, M. A. E., Yang, F., Biswas, N. K., Rahat, S. H. & Neelam, T. J. Machine learning-based error modeling to improve GPM IMERG precipitation product over the brahmaputra river basin. *Forecasting* **2**, 248–266 (2020).

[27] Bhuiyan, M. A. E., Nikolopoulos, E. I., Anagnostou, E. N., Quintana-Seguí, P. & Barella-Ortiz, A. A nonparametric statistical technique for combining global precipitation datasets: Development and hydrological evaluation over the Iberian Peninsula. *Hydrology and Earth System Sciences* **22**, 1371–1389 (2018).

[28] Wu, H., Yang, Q., Liu, J. & Wang, G. A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. *Journal of Hydrology* **584**, 124664 (2020).

[29] Shen, Y., Zhao, P., Pan, Y. & Yu, J. A high spatiotemporal gauge-satellite merged precipitation analysis over china. *Journal of Geophysical Research: Atmospheres* **119**, 3063–3075 (2014).

[30] Box, G. E. & Tiao, G. C. *Bayesian inference in statistical analysis* (John Wiley & Sons, 2011).

[31] Wu, P., Imbiriba, T., Elvira, V. & Closas, P. Bayesian data fusion with shared priors. *IEEE Transactions on Signal Processing* **72**, 275–288 (2023).

[32] Price, I. *et al.* Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796* (2023).

[33] Goodfellow, I. *et al.* Generative adversarial networks. *Communications of the ACM* **63**, 139–144 (2020).

[34] Kingma, D. P. & Welling, M. Auto-encoding variational bayes (2022). URL https://arxiv.org/abs/1312.6114. arXiv:1312.6114.

[35] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research* **22**, 1–64 (2021).

[36] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020).

[37] Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models (2022). URL https://arxiv.org/abs/2010.02502. arXiv:2010.02502.

[38] Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis (2021). URL https://arxiv.org/abs/2105.05233. arXiv:2105.05233.

[39] Yim, J. *et al.* Diffusion models in protein structure and docking. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **14**, e1711 (2024).

[40] Daras, G. *et al.* A survey on diffusion models for inverse problems (2024). URL https://arxiv.org/abs/2410.00083. arXiv:2410.00083.

[41] Zheng, H. *et al.* Inversebench: Benchmarking plug-and-play diffusion priors for inverse problems in physical sciences (2025). URL https://arxiv.org/abs/2503.11043. arXiv:2503.11043.

[42] Hess, P., Aich, M., Pan, B. & Boers, N. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nature Machine Intelligence* 1–11 (2025).

[43] Dieleman, S. Diffusion is spectral autoregression (2024). URL https://sander.ai/2024/09/02/spectral-autoregression.html.

[44] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation (2015). URL https://arxiv.org/abs/1505.04597. arXiv:1505.04597.

[45] Klein, S. A. *et al.* Are climate model simulations of clouds improving? an evaluation using the ISCCP simulator. *Journal of Geophysical Research: Atmospheres* **118**, 1329–1342 (2013).

[46] Zhang, C. *et al.* The E3SM diagnostics package (E3SM diags v2. 6): A Python-based diagnostics package for Earth system models evaluation. *Geoscientific model development discussions* **2022**, 1–35 (2022).

[47] Lee, J. *et al.* Systematic and objective evaluation of Earth system models: PCMDI Metrics Package (PMP) version 3. *Geoscientific Model Development* **17**, 3919–3948 (2024).

[48] Guilloteau, C., Foufoula-Georgiou, E., Kirstetter, P., Tan, J. & Huffman, G. J. How well do multisatellite products capture the space–time dynamics of precipitation? part i: Five products assessed via a wavenumber–frequency decomposition. *Journal of Hydrometeorology* **22**, 2805–2823 (2021).

[49] Guilloteau, C., Foufoula-Georgiou, E., Kirstetter, P., Tan, J. & Huffman, G. J. How well do multisatellite products capture the space–time dynamics of precipitation? part ii: Building an error model through spectral system identification. *Journal of Hydrometeorology* **23**, 1383–1399 (2022).

[50] Buizza, R. *et al.* The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS) (2018). URL https://www.ecmwf.int/node/18658.

[51] Zhang, J. *et al.* Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society* **97**, 621–638 (2016).

[52] Beck, H. E. *et al.* Mswep: 3-hourly 0.25 global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences* **21**, 589–615 (2017).

[53] Stevens, B. A perspective on the future of CMIP. *AGU Advances* **5**, e2023AV001086 (2024).

[54] Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development* **9**, 1937–1958 (2016).

[55] Guo, C. & Berkhahn, F. Entity embeddings of categorical variables (2016). URL https://arxiv.org/abs/1604.06737. arXiv:1604.06737.

[56] Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).

[57] Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems* **32** (2019).

[58] Luo, C. Understanding diffusion models: A unified perspective (2022). URL https://arxiv.org/abs/2208.11970. arXiv:2208.11970.

[59] Bond-Taylor, S. & Willcocks, C. G. ∞-diff: Infinite resolution diffusion with subsampled mollified states (2024). URL https://arxiv.org/abs/2303.18242. arXiv:2303.18242.

[60] Pidstrigach, J., Marzouk, Y., Reich, S. & Wang, S. Infinite-dimensional diffusion models (2023). URL https://arxiv.org/abs/2302.10130. arXiv:2302.10130.

[61] Zhang, B. & Wonka, P. Functional diffusion (2023). URL https://arxiv.org/abs/2311.15435. arXiv:2311.15435.

[62] Azizzadenesheli, K. *et al.* Neural operators for accelerating scientific simulations and design. *Nature Reviews Physics* **6**, 320–328 (2024).

[63] Biemond, J., Lagendijk, R. L. & Mersereau, R. M. Iterative methods for image deblurring. *Proceedings of the IEEE* **78**, 856–883 (2002).

[64] Li, Z. *et al.* Neural operator: Graph kernel network for partial differential equations (2020). URL https://arxiv.org/abs/2003.03485. arXiv:2003.03485.

[65] Kovachki, N. *et al.* Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research* **24**, 1–97 (2023).

[66] Li, Z. *et al.* Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895* (2020).

[67] Tang, H., Liu, Z., Li, X., Lin, Y. & Han, S. Torchsparse: Efficient point cloud inference engine (2022). URL https://arxiv.org/abs/2204.10319. arXiv:2204.10319.

[68] Ruiz, N. *et al.* Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation (2022).

[69] Chung, H., Kim, J., Mccann, M. T., Klasky, M. L. & Ye, J. C. Diffusion posterior sampling for general noisy inverse problems (2024). URL https://arxiv.org/abs/2209.14687. arXiv:2209.14687.

[70] Chao, J. *et al.* Learning to infer weather states using partial observations. *Journal of Geophysical Research: Machine Learning and Computation* **2**, e2024JH000260 (2025).

[71] Lugmayr, A. *et al.* Repaint: Inpainting using denoising diffusion probabilistic models (2022). URL https://arxiv.org/abs/2201.09865. arXiv:2201.09865.

[72] Zhang, G. *et al.* Towards coherent image inpainting using denoising diffusion implicit models (2023). URL https://arxiv.org/abs/2304.03322. arXiv:2304.03322.

[73] Meng, C. *et al.* SDEdit: Guided image synthesis and editing with stochastic differential equations (2022). URL https://arxiv.org/abs/2108.01073. arXiv:2108.01073.

[74] Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* **102**, 359–378 (2007).

[75] Hannachi, A. A primer for eof analysis of climate data. *Department of Meteorology, University of Reading* **1**, 3 (2004).

[76] Pulkkinen, S. *et al.* Pysteps: An open-source Python library for probabilistic precipitation nowcasting (v1. 0). *Geoscientific Model Development* **12**, 4185–4219 (2019).

[77] Huffman, G. J. *et al.* Nasa global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm theoretical basis document (ATBD) version* **4**, 30 (2015).

[78] Hersbach, H. *et al.* The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society* **146**, 1999–2049 (2020).

[79] Rasp, S. *et al.* Weatherbench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems* **12**, e2020MS002203 (2020).

[80] Olivetti, L. & Messori, G. Do data-driven models beat numerical models in forecasting weather extremes? a comparison of ifs hres, pangu-weather, and graphcast. *Geoscientific Model Development* **17**, 7915–7962 (2024).

[81] Daras, G., Cherapanamjeri, Y. & Daskalakis, C. How much is a noisy image worth? data scaling laws for ambient diffusion (2024). URL https://arxiv.org/abs/2411.02780. arXiv:2411.02780.

[82] Daras, G. *et al.* Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems* **36**, 288–313 (2023).

[83] Daras, G., Dagan, Y., Dimakis, A. & Daskalakis, C. Consistent diffusion models: Mitigating sampling drift by learning to be consistent. *Advances in Neural Information Processing Systems* **36**, 42038–42063 (2023).

[84] Daras, G., Dimakis, A. G. & Daskalakis, C. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. *arXiv preprint arXiv:2404.10177* (2024).

[85] Panaretos, V. M. & Zemel, Y. Statistical aspects of wasserstein distances. *Annual review of statistics and its application* **6**, 405–431 (2019).

[86] Paszke, A. *et al.* Automatic differentiation in pytorch (2017).