

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

Received: 5 July 2025

Accepted: 15 December 2025

Published online: 29 December 2025

 Check for updates

Sencan Sun¹, Congyi Nai², Baoxiang Pan²✉, Wentao Li³, Lu Li⁴, Xin Li⁵, Efi Foufoula-Georgiou^{6,7} & Yanluan Lin¹✉

Precipitation remains one of the most challenging climate variables to observe and predict. Existing datasets face intricate trade-offs: gauges are relatively trustworthy but sparse, satellites provide near-global coverage with retrieval uncertainties, and numerical models offer physical consistency but are biased. Here we introduce PRIMER (Precipitation Records Infinite MERging), a framework that fuses these complementary sources. PRIMER employs a coordinate-based diffusion model that learns from arbitrary spatial locations and associated intensity values, enabling seamless integration of gridded data and irregular gauge observations. Through two-stage training—first learning large-scale patterns, then refining with gauge measurements—PRIMER captures both large-scale structure and local precision. Once trained, it can correct biases in existing datasets—yielding significant error reductions at most gauge sites—and downscale reanalysis. In addition, by combining background estimates with extra gauges, it produces analysis fields that further reduce errors. All tasks are achieved through posterior sampling utilizing the prior obtained by fusing multi-source records. Crucially, it generalizes without retraining, correcting biases in operational forecasts and downscaling future scenario precipitation fields. This demonstrates how PRIMER can transform imperfect data into a source of strength.

Precipitation—when, where, and how much water falls from the sky to the earth's surface—governs freshwater availability, agricultural productivity, flood hazards, and ecosystem health across the globe¹. 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 complex spatiotemporal organization depends crucially on small-scale cloud microphysics processes⁴ 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 turbulence⁷, atmospheric stability⁸, and mesoscale circulations^{9,10} that remain computationally prohibitive to simulate explicitly. These complexities create fundamental observational and predictive challenges.

¹Ministry of Education Key Laboratory for Earth System Modeling, Department of Earth System Science, Tsinghua University, Beijing, China. ²Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. ³State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering and College of Hydrology & Water Resources, Hohai University, Nanjing, China. ⁴School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China. ⁵Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing, China. ⁶Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA, USA. ⁷Department of Earth System Science, University of California Irvine, Irvine, CA, USA. ✉e-mail: panbaoxiang@lasg.iap.ac.cn; yanluan@tsinghua.edu.cn

Currently, we rely upon three sources to derive precipitation information, namely in situ gauge observations, remote sensing, and numerical simulation that often assimilate in situ and remote-sensed data¹¹. 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 $2.5^\circ \times 2.5^\circ$ grid cells contain less than two gauges on average¹², 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 sampling¹³. 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 regions¹⁴. Numerical weather prediction and reanalysis products provide physically consistent, complete spatiotemporal coverage by assimilating available observations into dynamical models¹⁵. 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 errors¹⁶, 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 precipitation¹¹, 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^{17–27}. Among data-fusion approaches, Bayesian methods offer a coherent and probabilistically grounded solution. The key insight is elegant: by deriving an informative prior 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 and observational modalities through tailored likelihood functions.^{28–31} The framework naturally weights observations by their reliability and propagates uncertainties to yield full posterior distributions³², essential for risk assessment.

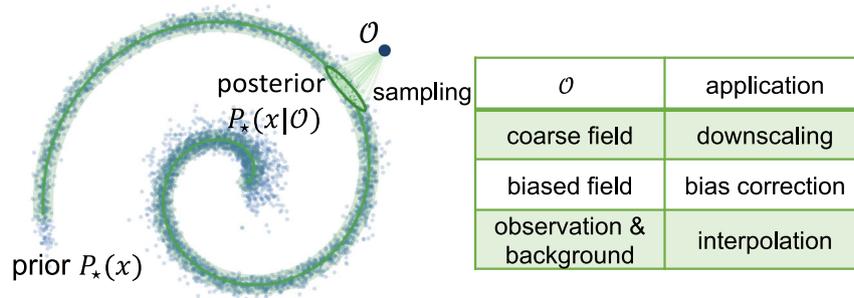
Recent advances in deep generative models, particularly probabilistic diffusion models^{33,34}, offer a transformative opportunity for implementing the above Bayesian framework. Diffusion models approximate target distributions by learning to reverse a gradual noising process. The forward process progressively perturbs data with Gaussian noise, while the reverse process, parameterized by a neural network, learns to invert this corruption to recover samples. This iterative denoising procedure enables diffusion models to generate high-quality and diverse samples, spanning domains from natural images³⁵ to protein structures³⁶, while also serving as priors for Bayesian inference, making them particularly well-suited for capturing the intricate patterns of precipitation. Once trained, they function as “plug-and-play” priors^{37–41}: 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 remains challenging because, from a machine learning perspective, a precipitation field is high-dimensional and the associated observation likelihood is complex. These barriers define the frontier for deploying generative AI in Earth-system science, demanding innovations that transcend conventional approaches.

To address these challenges, we introduce PRIMER (Precipitation Record Infinite MERging), a general framework that reconceptualizes how diffusion models can learn from imperfect, heterogeneous precipitation records. 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 Gaussian noise⁴². 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.

As emphasized by ref. 43, integrating data with varying degrees of sparsity—from sparse grids to dense fields—poses a major machine learning challenge. We acknowledge this issue and propose an approach to better handle such heterogeneity (see SI Section 2.8 for a comparison with ref. 43). Conventionally, diffusion models work on samples residing on fixed-resolution grids⁴⁴, 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 spatial fields $x : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 location and precipitation intensity—without relying on spatial interpolation (see Fig. 1b)—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 $P_{\text{ERAS}}(x)$ and $P_{\text{IMERG}}(x)$, which represent the climatological distributions of precipitation fields x derived from climate reanalysis, i.e., fifth-generation ECMWF atmospheric reanalysis (ERAS), and satellite-based retrieval dataset, i.e., Integrated Multi-satellite Retrievals for GPM (IMERG). We then fine-tune the model using gauge observational information at sparse grid locations (we refer hereafter to these densely observed grid cells as “gauge observations”; see Method 4.6 for data sources and detailed descriptions), so as to incorporate local accuracy, yielding the updated prior $P_*(x)$ (Fig. 1b; star indicates that it is supposed 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 $P_*(x|\mathcal{O})$ to produce improved ensemble estimates (Fig. 1a). Empirical evaluations demonstrate the effectiveness of our approach: it achieves statistically significant error reductions for grids which are densely observed with gauges, supplements high-frequency details through downscaling, and further reduces errors by merging gauge observations with the background, in a way similar to optimal interpolation, underscoring its

a posterior sampling by PRIMER and its application



b construct prior by fusing heterogeneous records

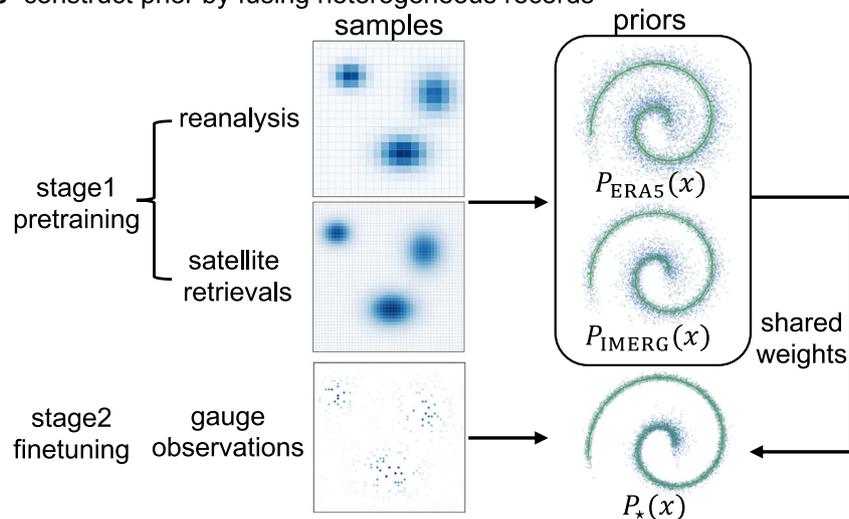


Fig. 1 | Overview of PRIMER. **a** Inference. PRIMER functions as a learned prior over the target precipitation field. Given a condition \mathcal{O} , PRIMER draws samples from the posterior $P_*(x|\mathcal{O})$. By changing \mathcal{O} , PRIMER samples from $P_*(x|\mathcal{O})$ under a shared prior, thereby unifying three applications in a single Bayesian generative framework. **b** Prior construction via principled data fusion. Because no single precipitation dataset is uniformly reliable across scales, PRIMER integrates heterogeneous records—reanalysis (e.g., ERA5), satellite retrievals (e.g., IMERG), and sparse gauge

observations—to obtain a more accurate prior. In Stage 1, the model is pretrained on gridded products to learn baseline priors $P_{\text{ERA5}}(x)$ and $P_{\text{IMERG}}(x)$. In Stage 2, it is fine-tuned with gauge observations under shared weights (section “Model training”) to produce a refined prior $P_*(x)$ that retains large-scale structure from gridded data while incorporating localized constraints from gauges. In the following experiments, we will demonstrate that $P_*(x)$ yields superior accuracy compared to baseline priors.

potential for operational use. It also generalizes to unseen operational forecasts without retraining and extends to downscaling future scenario precipitation fields in CMIP6. By transforming the challenge of heterogeneous, imperfect data from a limitation into a strength, PRIMER establishes a paradigm for precipitation data fusion that extends naturally to other Earth-system variables plagued by observational trade-offs.

Results

Reproducing climatological distributions

The gist of the PRIMER is to learn a trustworthy prior of precipitation fields, thereafter applying it for a broad range of relevant probabilistic inference tasks. 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 $P_{\text{IMERG}}(x)$, $P_{\text{ERA5}}(x)$, and the updated 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 $P_{\text{IMERG}}(x)$ and $P_{\text{ERA5}}(x)$ 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–i) are likewise well reproduced (PCC = 0.75 and 0.86), highlighting PRIMER’s capacity to represent not just the average precipitation spatial

structure but also its variance. Notably, we also introduce $P_*(x)$, constructed by fine-tuning PRIMER using sparse but reliable gauge observational information. 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 gauge 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. 9). Additional statistical evaluations—including precipitation frequency, extremes, skewness, and empirical orthogonal function (EOF) modes—are provided in the SI Fig. 10.

Case study on high-impact events

The previous section evaluated PRIMER’s ability to match climatology. 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

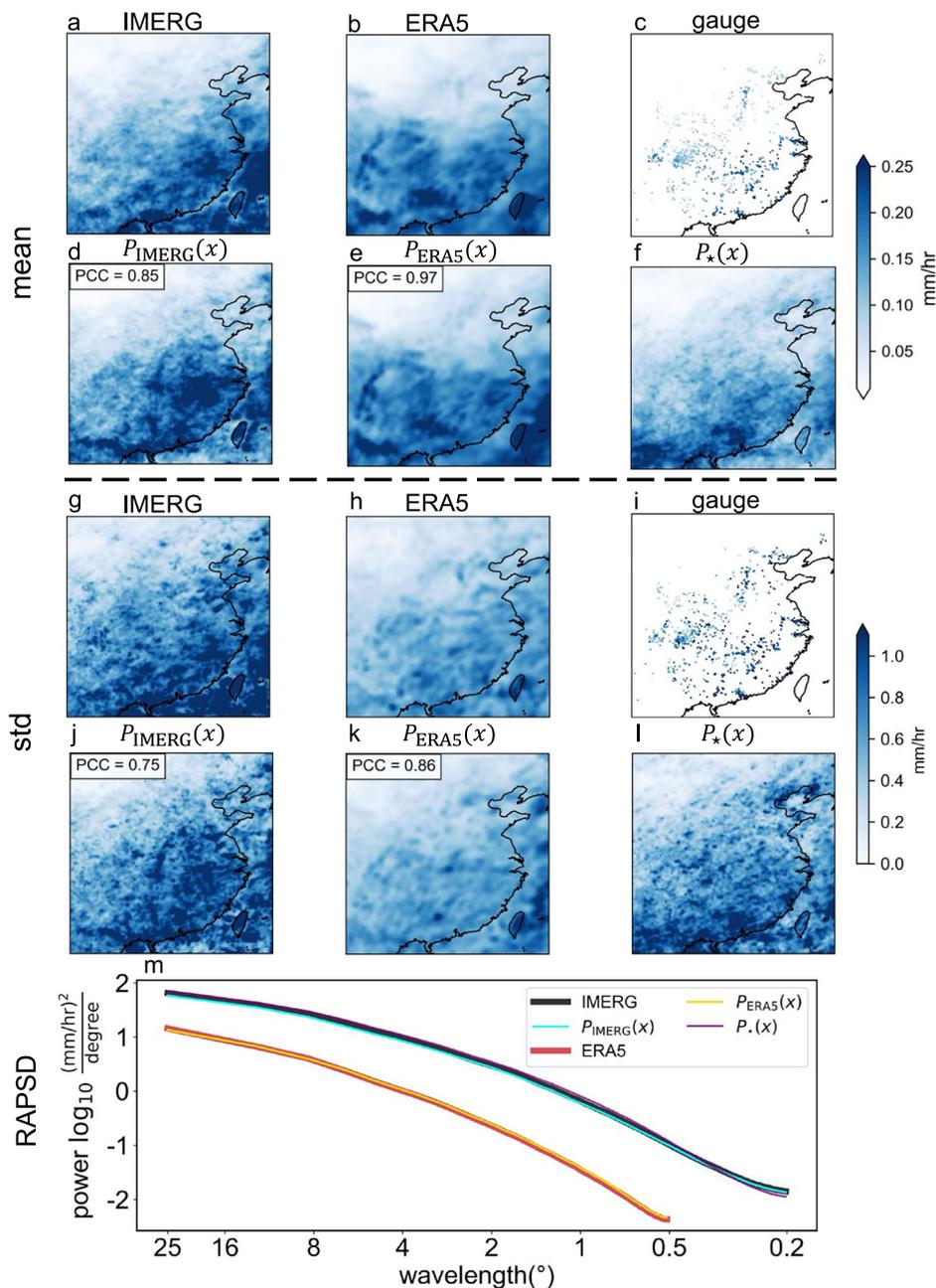


Fig. 2 | Climatological consistency between learned priors and reference datasets. **a–f** Spatial distributions of mean precipitation from IMERG (**a**), ERA5 (**b**), gauge observations at sparse grids which are dotted (**c**), $P_{\text{IMERG}}(x)$ (**d**), $P_{\text{ERA5}}(x)$ (**e**), and $P_*(x)$ (**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). The learned priors $P_{\text{IMERG}}(x)$ and $P_{\text{ERA5}}(x)$ closely follow their references, and $P_*(x)$ captures consistent multiscale characteristics. All statistics are computed from 1000 randomly sampled realizations. Maps were generated using Cartopy (<https://cartopy.readthedocs.io/stable/>).

posterior samples against the held-out gauge observations, 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 SI Figs. 12, 13.

To evaluate the effectiveness of PRIMER, we employ two standard performance metrics: the mean absolute error (MAE) and the continuous ranked probability score (CRPS), with the latter providing a probabilistic measure of an ensemble system’s accuracy (see Method “Evaluation metrics”). For each metric, we define a relative skill score, $\Delta\mathcal{M}$, as the non-negative error measure of the original precipitation dataset (ERA5 or IMERG) minus that of the posterior sample, so that positive values indicate reduced error and thus enhanced skill. All evaluations are conducted at a spatial resolution of 0.1°, where ERA5, IMERG, and posterior samples are compared against gauge observations treated as ground truth.

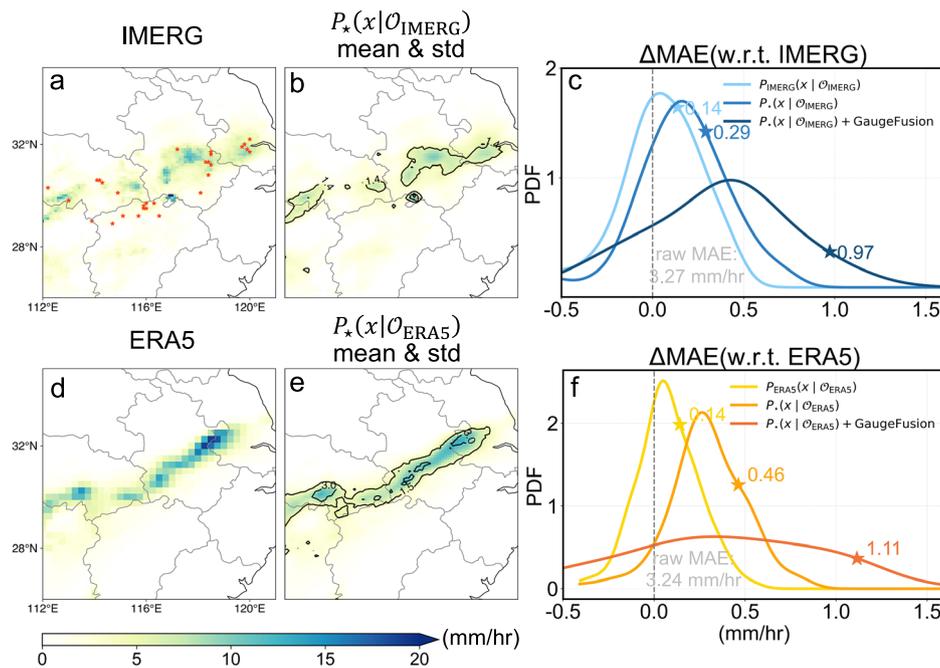


Fig. 3 | Case study of a Meiyu precipitation event on 2 July 2016 at 05 UTC. **a** IMERG at the target time, with gauge locations shown as red dots (used as ground truth for evaluation). **b** Posterior mean (shaded contours) and standard deviation (labeled contour lines) from $P_*(x|\mathcal{O}_{\text{IMERG}})$ inferred by PRIMER, based on 100 ensemble samples. **c** Probability density functions (PDFs) of relative changes in mean absolute error (ΔMAE). Here, ΔMAE is defined as the MAE of the original IMERG data minus each posterior sample, both evaluated against gauge observations. Positive values of ΔMAE indicate the amount by which PRIMER effectively

reduces the original IMERG errors at gauge locations. **d–f** analogous to **(a–c)** but for ERA5. In **c, f**, different curves represent posterior distributions as labeled, with ensemble means indicated by stars. The label “+GaugeFusion” refers to an experiment in which gauge observational information at sparse grids (20% of locations are incorporated during sampling, with errors evaluated on the remaining 80%) are combined with the background (raw IMERG or ERA5) in a manner analogous to optimal interpolation. This additional observational constraint markedly improves accuracy. Maps were generated using Cartopy.

As shown in Fig. 3c, f, the updated prior $P_*(x)$ substantially outperforms baseline priors derived from ERA5 and IMERG. The ensemble-mean ΔMAE decreases from 0.46 mm/hr for $P_{\text{ERA5}}(x|\mathcal{O}_{\text{ERA5}})$ to 0.14 mm/hr for $P_{\text{ERA5}}(x|\mathcal{O}_{\text{ERA5}})$; 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 $P_{\text{ERA5}}(x|\mathcal{O}_{\text{ERA5}})$ are consistently lower than those for $P_*(x|\mathcal{O}_{\text{ERA5}})$.

An important feature of PRIMER is its ability to incorporate additional gauge observations into the posterior sampling process, rather than relying solely on background fields. This capability resembles the requirements of operational analysis systems, where integrating sparse gauges can substantially improve quality. To evaluate this property, we design an experiment that mimics real-world conditions by including a subset (20%) of gauge observations (while errors are evaluated on the remaining 80%) during sampling (hereafter denoted as “+GaugeFusion”). The inclusion of these observational constraints yields a marked improvement in accuracy, with the ensemble-mean ΔMAE increasing to 1.11 and 0.97 mm/hr for the ERA5 and IMERG cases, respectively. In addition, it needs to be noted that spectral analysis further highlights distinctions among posterior samples (see SI Fig. 11). While $P_{\text{ERA5}}(x|\mathcal{O}_{\text{ERA5}})$ retains low-frequency biases, both $P_*(x|\mathcal{O}_{\text{ERA5}})$ and its GaugeFusion variant enhance high-frequency components.

Statistical verifications

We applied PRIMER to a curated test set of 150 precipitation events from 2016, selected based on the criteria detailed in SI Section 3. For each event, 50 posterior samples were drawn from $P_*(x|\mathcal{O})$, 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. At each gauge

location, we computed the MAE and CRPS of the posterior distributions. MAE was calculated using the ensemble mean of each posterior distribution compared against the corresponding gauge observation, while CRPS assessed the full probabilistic accuracy. We then calculated differences in both metrics between original datasets—and $P_*(x|\mathcal{O})$. In simple terms, a positive value at a gauge location means that PRIMER reduces the error of the original dataset after.

Figures 4a, b reveal widespread reductions in MAE, highlighting PRIMER’s ability to systematically correct biases inherent in the original datasets. Figure 4c, d shows pronounced reductions in CRPS, with deeper blue tones indicating substantial gains in probabilistic estimates. These results demonstrate that PRIMER captures the posterior distribution accurately, with the improvements confirmed as statistically significant by *t*-tests. In addition to PRIMER, we also evaluated baseline priors ($P_{\text{ERA5}}(x)$ and $P_{\text{IMERG}}(x)$) as well as baseline methods BCSD-EQM (Bias correction and spatial disaggregation–equitable quantile mapping)⁴⁸ and RM (random mixing)⁴⁹ (for notes of two methods, refer to SI Section 4). These baselines were subjected to the same application. As detailed in the SI Figs. 6–8, PRIMER generally outperforms these baselines. What’s more, the largest improvements are observed in the Sichuan Basin and the Pearl River Delta—regions with dense populations and strong economic activity. We further analyzed the correlation between gauge density and performance improvement (SI Section 5). Although a positive trend is apparent, the correlation is not statistically significant, indicating that PRIMER delivers relatively spatially consistent improvements irrespective of local gauge density.

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

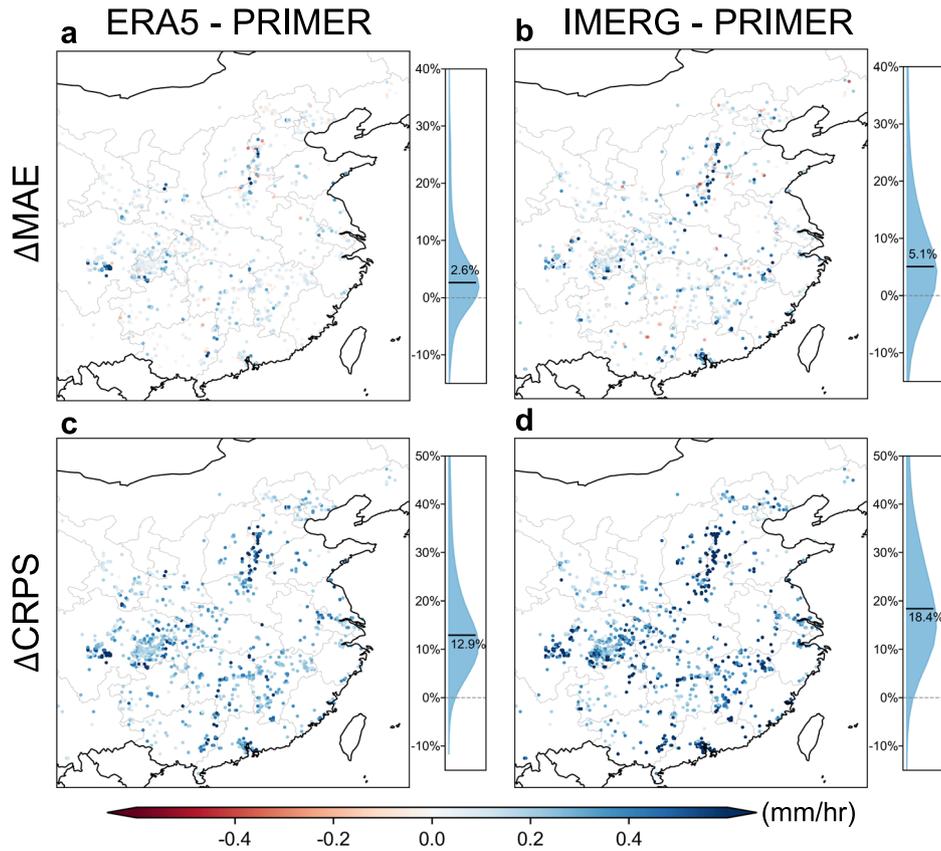


Fig. 4 | Bias correction of existing precipitation datasets. This figure shows the reduction in MAE (top row) and CRPS (bottom row) after bias correction (for ERA5 also downscaling) using PRIMER, applied separately to ERA5 (panel **a, c**) and IMERG (panel **b, d**). Each dot denotes the grid where gauge observational information is available. Positive values indicate a reduction in error relative to the original IMERG

or ERA5, while red points indicate deterioration. The violin plot on the right displays the PDF distribution of the relative error change across all locations, with the central black line representing the median of the relative error changes. The evaluation is based on 150 precipitation events that occurred in 2016. For the results of the three baseline methods, see SI Figs. 6–8. Maps were generated using Cartopy.

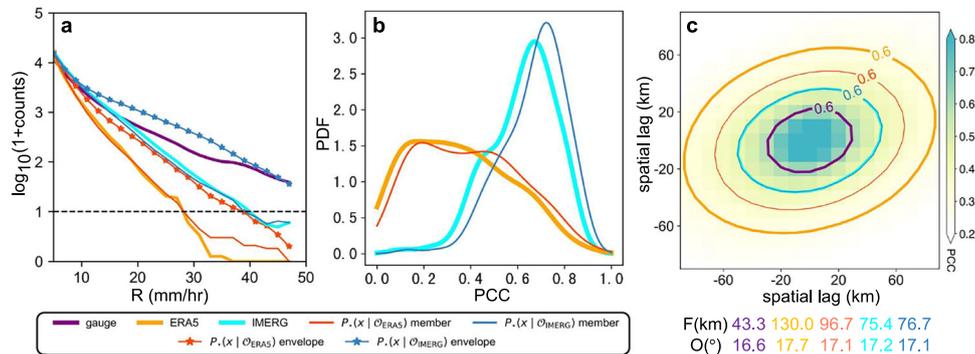


Fig. 5 | Improved physical reality of existing datasets. **a** Log-transformed histogram of precipitation intensity (2 mm/hr bins) at only gauge locations, aggregated over test sets. This panel highlights the ability of different datasets to reproduce the tail of the precipitation distribution (with the purple line as the ground truth). **b** Probability density functions (PDFs) of 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 legend below.

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 $P_*(x|O_{\text{ERA5}})$ and $P_*(x|O_{\text{IMERG}})$ more accurately reproduce the frequency distribution of precipitation, particularly at higher intensities. Both perspectives 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 original products. Improvements in spatial structure are further quantified using PCCs with respect to gauge observations (Fig. 5b). $P_*(x|O_{\text{ERA5}})$ and $P_*(x|O_{\text{IMERG}})$ 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

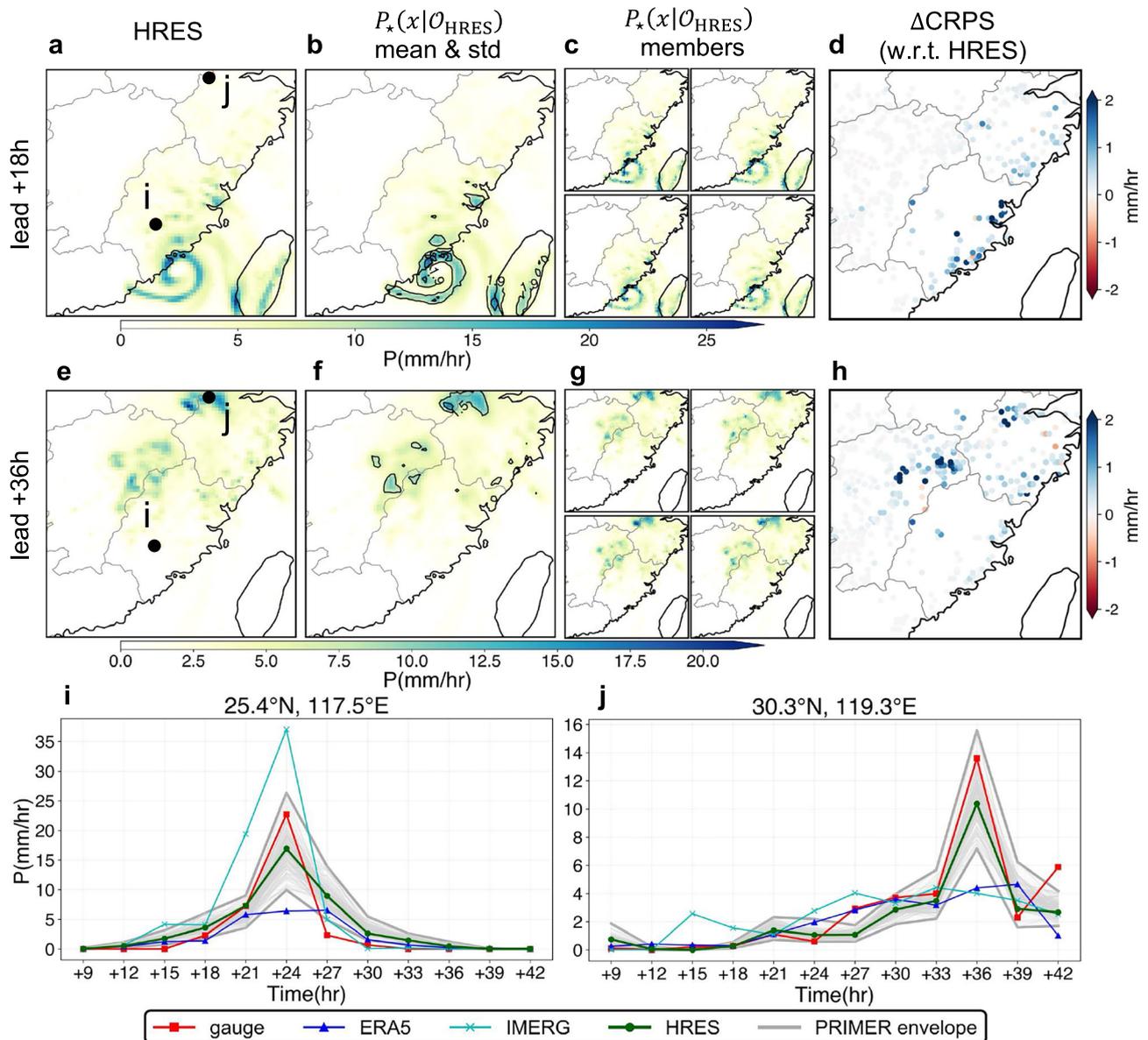


Fig. 6 | Bias-correction for operational forecasts without retraining. **a, e** HRES forecasts at 18-h and 36-h lead times (other lead times see SI Fig. 15), initialized at 00:00 UTC on 14 September 2016, coinciding with the landfall of Typhoon Meranti. **b, f** Ensemble means. **c, g** Four representative ensemble members, illustrating internal variability and structural diversity. **d, h** Spatial distribution of Δ CRPS, with

blue indicating improvement and red indicating deterioration. **i, j** Precipitation time series at two representative gauged grids (more stations see SI Fig. 16); gray envelope denotes the spread across 100 ensemble members. Maps were generated using Cartopy.

proposed to assess spatial organization and feature propagation^{50,51}, we employ a simplified yet informative diagnostic based on two-dimensional spatial lagged correlation coefficient (Method “Evaluation tool”, Fig. 5c). Physically, this 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 $P_*(x|O_{\text{ERA5}})$ and $P_*(x|O_{\text{IMERG}})$ produce orientations that are more consistent with reference orientations derived from gauge observations, indicating improved spatial alignment. In terms of focal length, $P_*(x|O_{\text{ERA5}})$ exhibits a clear reduction, while $P_*(x|O_{\text{IMERG}})$ shows no substantial improvement. These results demonstrate PRIMER’s effectiveness in correcting the spatial anisotropy of precipitation systems.

Generalization test

PRIMER is not only effective for existing precipitation datasets, but also exhibits a certain degree of 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 example⁵². Despite 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 $P_*(x|O_{\text{HRES}})$ (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 Δ CRPS (Fig. 6d, h) with widespread positive values (blue dots) indicate that PRIMER produces a reliable ensemble system for HRES. These improvements arise from the Bayesian posterior sampling mechanism. By drawing samples from $P_*(x|O_{\text{HRES}})$, we effectively use the learned prior distribution $P_*(x)$ –

which has been calibrated to match gauge statistics—to adjust the original HRES forecasts. 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, confirming that the HRES guidance is effectively incorporated. Occasional deviations arise when the HRES forecasts and the local gauge observations exhibit divergent trends—for instance, a slower decrease in HRES versus a sharper observed decline (panel i at +27 h), or mismatched peaks (panel j at +27 h and +42 h). In such cases, the posterior may not fully capture the observed variability.

To assess PRIMER under a future scenario, we selected a model simulation output from CMIP6 for 2050⁵³, when elevated CO₂ forcing is expected to alter precipitation regimes. Hourly precipitation was downscaled to 0.1° with PRIMER and compared against the raw model output. As shown in SI Fig. 17, the domain-mean precipitation curves from the raw output and the downscaled fields remain closely aligned, indicating that PRIMER preserves large-scale variability while adding fine-scale structure under a shifted climate state. Taken together, these results empirically suggest the broader utility of PRIMER as a foundation model for downstream applications without additional retraining (zero-shot adaptation).

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 of each source. Generative AI, particularly probabilistic diffusion models, offers a powerful approach by capturing the intricate distribution of precipitation patterns. However, practical application has been severely limited by 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 function in concept, 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 observational information. 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 and two baseline methods. Furthermore, experiments show that incorporating additional gauge observations into the posterior sampling markedly enhances accuracy, highlighting PRIMER's potential for optimal interpolation in operational contexts. Crucially, PRIMER exhibits a certain degree of zero-shot generalization, maintaining consistency when applied to previously unseen operational forecasts and even future scenario simulations.

Despite the impressive performance of PRIMER, several limitations remain. First, the scarcity of high-quality in situ gauge observations over oceanic regions constrains our ability to comprehensively evaluate model performance. Second, our current experiments are restricted to precipitation fusion within China rather than at the global scale. This decision was primarily driven by the substantial computational cost of global fusion, which exceeds our available resources. Third, from a methodological perspective, PRIMER does not provide a

theoretical guarantee of temporal continuity across posterior samples at different time steps. A promising direction is to extend the framework from frame-wise priors to videostyle priors that jointly model consecutive fields, thereby enhancing temporal consistency. Notwithstanding these limitations, precipitation itself is among the most complex and discontinuous variables in the climate system, which sets a particularly stringent benchmark for validating our methodology before extending it to other variables and broader climate domains.

In practice, PRIMER is readily deployable: when integrated into operational forecasting chains, it can perform real-time post-processing of precipitation fields from numerical or AI-based forecasts, delivering both bias correction and downscaling. It also integrates seamlessly with optimal interpolation by weighting gauge observations and the background, thereby yielding substantially improved states. Looking ahead, PRIMER advances three key principles for the community. First, recognizing that Earth-system data were inherently imperfect, AI for geoscience must be designed to be uncertainty-aware. By fusing heterogeneous precipitation records into a unified prior, PRIMER distills multi-source information into model parameters, in a manner analogous to how large language models compress corpus-level statistics—yielding greater accuracy than training on a single data source alone. Second, its flexible architecture and training framework naturally accommodate irregular observations alongside gridded products, providing a reusable template for broader geoscientific AI applications. Finally, PRIMER is intrinsically extensible: auxiliary variables (such as temperature, wind, and humidity) can be incorporated as additional input channels, enabling a more complete representation of the atmospheric state and ultimately strengthening both short-range weather forecasting and long-range climate simulation.

Methods

Problem formulation

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

Let x denote the target precipitation field. Different data sources—including gridded products such as satellite retrievals and reanalysis, as well as sparse gauge observations—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 $P(x)$. 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 framework for modeling such a prior.

Once an informative prior is established, posterior inference is conducted as new observational evidence \mathcal{O} becomes available. Posterior distribution $P(x|\mathcal{O})$ can be factored into two components: the prior distribution $P(x)$, and the likelihood $P(\mathcal{O}|x)$. 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} . This Bayesian framework naturally enables various downstream applications, such as super-resolution by conditioning on coarse data, bias correction by conditioning on biased estimates, and optimal interpolation by jointly conditioning on observations and background (Fig. 1a).

Preliminary on diffusion models

To construct a prior, we employ score-based diffusion models. To enable the model to distinguish between sources during training, we associate each sample with a corresponding entity embedding e_i ($e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$, $e_3 = (0, 0, 1)$)⁵⁴, 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)^{33,34,55,56}:

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

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, we solve the reverse-time SDE^{55,57}:

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

where the score function $\nabla_{x_t} \log P_\theta(x_t | e_i)$ denotes the gradient of the log-density. Since this score is intractable, we approximate it using a neural network f_θ .

PRIMER

Traditional diffusion models typically rely heavily on U-Net architectures⁴⁴, 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 framework inspired by recent theoretical advances^{58–61}, which generalizes diffusion models from finite-dimensional Euclidean space to an infinite-dimensional Hilbert space \mathcal{H} , as illustrated in SI Fig. 1. In this setting, each element $x \in \mathcal{H}$ is a function $x : \mathbb{R}^n \rightarrow \mathbb{R}^d$, 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 ref. 58, we define \mathcal{H} as $L^2([0, 1]^n \rightarrow \mathbb{R}^d)$, where L^2 denotes the space of functions f such that $\int_{[0, 1]^n} |f(x)|^2 dx < \infty$. The rationale for the name PRIMER is discussed in SI Section 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 Section 2.2. Similarly, PRIMER also mollifies x_0 , which ensures that Lx_0 inherits 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, and $t = \sigma^2/2$, with σ being the standard deviation of kernel k (a detailed derivation is provided in SI Section 2.3). Directly applying the inverse transformation is often numerically unstable; thus, we employ the Wiener filter, defined as^{58,62}: $\tilde{\epsilon}(\boldsymbol{\omega}) = \frac{e^{-|\boldsymbol{\omega}|^2 t}}{e^{-2|\boldsymbol{\omega}|^2 t} + \delta^2} \xi(\boldsymbol{\omega})$, where δ is a small positive regularization parameter.

Network architecture. Neural Operators are capable of learning a map between two functional spaces^{61,63–65}. Neural operators achieve discretization invariance by learning integral kernels parameterized via neural networks. Specifically, for an input function $x : \mathbb{R}^n \rightarrow \mathbb{R}^d$, with observations at m distinct spatial locations, the operator $K(x; \theta)$ is defined as:

$$(K(x; \theta)x)(\mathbf{c}) = \int_{\mathbb{R}^n} \kappa_\theta(\mathbf{c}, \mathbf{b}, x(\mathbf{c}), x(\mathbf{b})) x(\mathbf{b}) d\mathbf{b}, \tag{3}$$

where $\kappa_\theta : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a kernel function parameterized by θ , which captures complex non-local dependencies. PRIMER

implements a hybrid multiscale architecture that synthesizes the strengths of Neural Operators and convolutional networks. PRIMER first processes the input $x \in \mathbb{R}^{d \times m}$, together with their corresponding locations $c \in \mathbb{R}^{n \times m}$ using a series of SparseConvResBlocks, which primarily employ sparse depthwise convolutions⁶⁶, 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 that enables the model to capture richer representations. For the motivation behind SparseConvResBlock, see SI Section 2.6. Since the features lie on an irregular set of discrete locations, we project them onto a coarse regular grid based on their spatial coordinates (see SI code 1). This transformation aligns the features to a structured grid layout. A U-Net is applied to this grid to capture multiscale 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 are applied to produce the final output tensor of shape $\mathbb{R}^{d \times m}$. Details of the network are provided in SI Section 2.5.

Model training. The model is optimized by minimizing a simplified denoising objective^{33,55,58} (derivation provided in SI Section 2.4):

$$\mathcal{L} = \mathbb{E}_t \left[\| f_\theta(x_t, t, e_i) - \xi \|_{\mathcal{H}}^2 \right], \tag{4}$$

where x_t denotes the noisy input at time step t , e_i represents the entity embedding, 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). In Stage 2, we specialize the pretrained model to sparse gauge observations (e_3), following a strategy akin to DreamBooth⁶⁷. 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:

$$\mathcal{L}_{\text{fine-tuning}} = \alpha_1 \mathcal{L}_{\text{ERA5}} + \alpha_2 \mathcal{L}_{\text{IMERG}} + \alpha_3 \mathcal{L}_{\text{gauge}}, \tag{5}$$

with weights $\alpha_1 = 0.1$, $\alpha_2 = 0.4$, and $\alpha_3 = 0.5$. Assigning $\alpha_3 = 0.5$ prevents catastrophic forgetting of ERA5 and IMERG knowledge while ensuring strong gauge influence⁶⁸. Among the remaining weights, IMERG ($\alpha_2 = 0.4$) is favored over ERA5 ($\alpha_1 = 0.1$) given its finer resolution. Although not optimized through exhaustive search, this empirical configuration preserves climatological priors while adapting to high-fidelity signals, thereby grounding the generative manifold in real-world observations.

The full training and inference pipelines are summarized in SI Algorithm 1 and SI Algorithm 2, with an overview schematic shown in SI Fig. 2. For the configuration of the hyperparameters, see SI Section 2.7.

Posterior sampling

In tasks such as bias correction, downscaling, and optimal interpolation, the objective is to infer an unknown target state x given observations \mathcal{O} . PRIMER enables the incorporation of prior knowledge through a prior $P(x)$, facilitating posterior inference via Bayes' theorem: $P(x|\mathcal{O}) \propto P(\mathcal{O}|x)P(x)$. The standard reverse-time SDE can be modified to sample from the posterior distribution, yielding the following reverse diffusion process:

$$dx_t = \left[f(x_t, t) - g^2(t) \left(\nabla_{x_t} \log P_\theta(x_t | e_i) + \nabla_{x_t} \log P_\theta(\mathcal{O} | e_i, x_t) \right) \right] dt + g(t) dW_t. \tag{6}$$

This formulation requires two key components: the time-dependent score function $\nabla_{x_t} \log P_\theta(x_t | e_i)$, which can be approximated by a trained score network; and the gradient of the likelihood

$\nabla_{x_t} \log P_{\theta}(\mathcal{O}|e_i, x_t)$, 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^{37,38,69}. In light of the characteristics of our problem setting, we adopt two representative approaches: Inpainting^{70–72} and SDEdit⁷³.

Inpainting 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 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 $P_{*}(x|\mathcal{O}_{\text{IMERG}})$ and compute both the RMSE and CRPS over 50 repeated subsampling trials, each selecting 10 members randomly. As shown in SI Fig. 4, 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.

Statistical methods

Baseline methods. We employed two additional statistical methods for downscaling and bias correction, namely BCSD-EQM (bias correction and spatial disaggregation– equitable quantile mapping)⁴⁸ and RM (random mixing)⁴⁹. Owing to space limitations, the algorithmic flowcharts are provided in SI Section 4.

Evaluation metrics

Deterministic accuracy. To assess the accuracy of posterior sampling, 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 :

$$\text{MAE} = \frac{1}{N} \sum_i |\hat{x}_i - x_i|. \tag{7}$$

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

$$\text{PCC} = \frac{\sum_i (\hat{x}_i - \bar{\hat{x}})(x_i - \bar{x})}{\sqrt{\sum_i (\hat{x}_i - \bar{\hat{x}})^2} \sqrt{\sum_i (x_i - \bar{x})^2}}. \tag{8}$$

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

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:

$$\text{CRPS}(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}_{\{x \geq y\}})^2 dx, \tag{9}$$

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

Evaluation tool

Spatial lagged correlation coefficient. We evaluate the spatial dependency of a 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 PCCs 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.

EOF. 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 via⁷⁵:

$$x = LY, \tag{10}$$

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

RAPSD. To quantify spatial variability⁷⁶, we compute the radially averaged power spectral density (RAPSD) using the open-source Pysteps library⁷⁷. Given a 2D scalar field $f(x, y) \in \mathbb{R}^{H \times W}$, its discrete Fourier transform is $F(k_x, k_y) = \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f(x, y) e^{-2\pi i (\frac{k_x x}{H} + \frac{k_y y}{W})}$, and the corresponding power spectral density is

$$P(k_x, k_y) = \frac{1}{HW} |F(k_x, k_y)|^2. \tag{11}$$

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

$$\text{RAPSD}(k) = \frac{1}{N_k} \sum_{(k_x, k_y) \in \mathcal{A}_k} P(k_x, k_y), \tag{12}$$

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.

Normalized occurrence versus rank analysis. For each gauge and hour with ground truth y and an ensemble of N realizations $\{\widehat{y}^{(k)}\}_{k=1}^N$, we define $r = \frac{1}{N} \sum_{k=1}^N \mathbf{1}\{\widehat{y}^{(k)} \leq y\}$. If the ensemble is perfectly calibrated, $\{r\}$ are uniformly distributed on $[0, 1]$. We assess this by plotting a histogram of normalized occurrence versus rank and by comparing the empirical CDF of $\{r\}$ against the $y = x$ reference. Deviations from uniformity are diagnostic: U-shaped or dome-shaped histograms indicate under-dispersion or over-dispersion of the ensemble, respectively^{78,79}.

Data

Pretraining uses two gridded datasets: Integrated Multi-satellite Retrievals for GPM (IMERG)⁸⁰ and ERA5⁸¹. IMERG provides global precipitation estimates at 0.1° spatial and 30-min temporal

resolutions. To match ERA5's hourly resolution, pairs of consecutive 30-min 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 a gauge-assimilated gridded dataset from Shen et al. 27, constructed from over 30,000 Automatic Weather Stations (AWS) across China, with a spatial resolution of 0.1° and a temporal resolution of 1 hour. Since we do not have direct access to raw gauge measurements, we select only grid cells containing at least one assimilated AWS observation as a proxy for gauge observations. We use data from 2015 and 2017 for training, reserving 2016 for testing to align with the Typhoon Meranti forecasting experiment. For evaluation, we use a subset of grid cells containing at least four AWS observations, assuming these provide more reliable ground truth due to higher observation density. Throughout this work, we refer to these densely observed grid cells simply as “gauge observations.” (See SI Fig. 5 for the spatial distribution of grids with gauge observations). After identical cropping and preprocessing, the data were organized as two arrays: (N, 1) for precipitation intensity and (N, 2) for the grid indices (row, column) corresponding to each gauge's longitude-latitude location on the 0.1° target domain, both of which are input into the model during fine-tuning.

IFS HRES is ECMWF's flagship deterministic high-resolution model and is widely regarded as one of the best physics-based numerical-weather-forecast models in the world^{82,83}. HRES produces hourly forecasts at a 0.1° horizontal resolution. We further used simulation outputs from CAM-MPAS-HR under the HighResMIP forced-atmosphere (2015–2050) configuration, with SST and sea-ice prescribed from CMIP5 RCP8.5⁵³. The model has a nominal resolution of 0.25° (variant r1i1p1f1), and we used only the data for the year 2050. Precipitation fields were downsampled from 0.25° to 0.1° by PRIMER. These two datasets were included in our experiments to demonstrate PRIMER's good generalization capability on datasets it was not trained on.

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=HRES from the IFS used in this study are produced by the ECMWF. For more detailed information on HRES access, please refer <https://www.ecmwf.int/en/forecasts/datasets/set-i>. The CAM-MPAS-HR simulation data under HighResMIP forced-atmosphere (2015–2050) configuration are publicly available and can be conveniently accessed at <https://aims2.llnl.gov/search>.

Code availability

The code implementing PRIMER is openly available on GitHub at: <https://github.com/sunmoumou/PRIMER>.

References

- Kotz, M., Levermann, A. & Wenz, L. The effect of rainfall changes on economic production. *Nature* **601**, 223–227 (2022).
- Sun, Y., Solomon, S., Dai, A. & Portmann, R. W. How often does it rain?. *J. Clim.* **19**, 916–934 (2006).
- Pendergrass, A. G. & Knutti, R. The uneven nature of daily precipitation and its change. *Geophys. Res. Lett.* **45**, 11–980 (2018).
- Stevens, B. & Feingold, G. Untangling aerosol effects on clouds and precipitation in a buffered system. *Nature* **461**, 607–613 (2009).
- Birch, C. et al. Impact of soil moisture and convectively generated waves on the initiation of a west african mesoscale convective system. *Q. J. R. Meteorol. Soc.* **139**, 1712–1730 (2013).
- 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).
- Teixeira, J. et al. Parameterization of the atmospheric boundary layer: a view from just above the inversion. *Bull. Am. Meteorol. Soc.* **89**, 453–458 (2008).
- Lepore, C., Veneziano, D. & Molini, A. Temperature and cape dependence of rainfall extremes in the eastern United States. *Geophys. Res. Lett.* **42**, 74–83 (2015).
- Arakawa, A. The cumulus parameterization problem: past, present, and future. *J. Clim.* **17**, 2493–2525 (2004).
- Houze Jr, R. A. Mesoscale convective systems. *Rev. Geophys.* **42**, RG4004 (2004).
- Sun, Q. et al. A review of global precipitation data sets: data sources, estimation, and intercomparisons. *Rev. Geophys.* **56**, 79–107 (2018).
- Kidd, C. & Huffman, G. Global precipitation measurement. *Meteorol. Appl.* **18**, 334–353 (2011).
- Hou, A. Y. et al. The global precipitation measurement mission. *Bull. Am. meteorol. Soc.* **95**, 701–722 (2014).
- Levizzani, V., Amorati, R. & Meneguzzo, F. A review of satellite-based rainfall estimation methods. *European Commission MUSIC Report (EVK1-CT-2000-00058)* **66** (2002).
- Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
- Tapiador, F. J. et al. Is precipitation a good metric for model performance?. *Bull. Am. Meteorol. Soc.* **100**, 223–233 (2019).
- He, J. et al. The first high-resolution meteorological forcing dataset for land process studies over China. *Sci. Data* **7**, 25 (2020).
- Ma, Y. et al. Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. *J. Geophys. Res. Atmos.* **123**, 814–834 (2018).
- Baez-Villanueva, O. M. et al. Rf-mep: A novel random forest method for merging gridded precipitation products and ground-based measurements. *Remote Sens. Environ.* **239**, 111606 (2020).
- Yumnam, K., Guntu, R. K., Rathinasamy, M. & Agarwal, A. Quantile-based Bayesian model averaging approach towards merging of precipitation products. *J. Hydrol.* **604**, 127206 (2022).
- Xie, P. & Xiong, A.-Y. A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *J. Geophys. Res. Atmos.* **116**, D21106 (2011).
- Woldemeskel, F. M., Sivakumar, B. & Sharma, A. Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. *J. Hydrol.* **499**, 167–176 (2013).
- Fan, Z. et al. A comparative study of four merging approaches for regional precipitation estimation. *IEEE Access* **9**, 33625–33637 (2021).
- Zhang, L. et al. Merging multiple satellite-based precipitation products and gauge observations using a novel double machine learning approach. *J. Hydrol.* **594**, 125969 (2021).
- 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. *Hydrol. Earth Syst. Sci.* **22**, 1371–1389 (2018).
- Wu, H., Yang, Q., Liu, J. & Wang, G. A spatiotemporal deep fusion model for merging satellite and gauge precipitation in China. *J. Hydrol.* **584**, 124664 (2020).
- Shen, Y., Zhao, P., Pan, Y. & Yu, J. A high spatiotemporal gauge-satellite merged precipitation analysis over China. *J. Geophys. Res. Atmos.* **119**, 3063–3075 (2014).
- Box, G. E. & Tiao, G. C. *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, 2011).

29. Wu, P., Imbiriba, T., Elvira, V. & Closas, P. Bayesian data fusion with shared priors. *IEEE Trans. Signal Process.* **72**, 275–288 (2023).
30. Wikle, C. K. & Berliner, L. M. A Bayesian tutorial for data assimilation. *Phys. D Nonlinear Phenom.* **230**, 1–16 (2007).
31. Bonavita, M. Ensemble of data assimilations and uncertainty estimation. In *ECMWF Seminar on Data Assimilation for Atmosphere and Ocean* (2011).
32. Price, I. et al. Probabilistic weather forecasting with machine learning. *Nature* **637**, 84–90 (2025).
33. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).
34. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. Preprint at arXiv <https://arxiv.org/abs/2010.02502> (2022).
35. Dhariwal, P. & Nichol, A. Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* **34**, 8780–8794 (2021).
36. Yim, J. et al. Diffusion models in protein structure and docking. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **14**, e1711 (2024).
37. Daras, G. et al. A survey on diffusion models for inverse problems. Preprint at arXiv <https://arxiv.org/abs/2410.00083> (2024).
38. Zheng, H. et al. Inversebench: benchmarking plug-and-play diffusion priors for inverse problems in physical sciences. Preprint at arXiv <https://arxiv.org/abs/2503.11043> (2025).
39. Hess, P., Aich, M., Pan, B. & Boers, N. Fast, scale-adaptive and uncertainty-aware downscaling of earth system model fields with generative machine learning. *Nat. Mach. Intell.* **7**, 363–373 (2025).
40. Yang, S. et al. Generative assimilation and prediction for weather and climate. Preprint at arXiv <https://arxiv.org/abs/2503.03038> (2025).
41. Nai, C., Chen, X., Yang, S., Xiao, Z. & Pan, B. Boosting weather forecast via generative superensemble. *npj Clim. Atmos. Sci.* **8**, 377 (2025).
42. Dieleman, S. Diffusion is spectral autoregression. <https://sander.ai/2024/09/02/spectral-autoregression.html> (2024).
43. Andrychowicz, M. et al. Deep learning for day forecasts from sparse observations. Preprint at arXiv:2306.06079 (2023).
44. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional networks for biomedical image segmentation. *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.* 234–241 (2015).
45. Klein, S. A. et al. Are climate model simulations of clouds improving? An evaluation using the ISCCP simulator. *J. Geophys. Res. Atmos.* **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. *Geosci. Model Dev. Discuss.* **2022**, 1–35 (2022).
47. Lee, J. et al. Systematic and objective evaluation of Earth system models: PCMDI metrics package (PMP) version 3. *Geosci. Model Dev.* **17**, 3919–3948 (2024).
48. Lorenz, C., Portele, T. C., Laux, P. & Kunstmann, H. Bias-corrected and spatially disaggregated seasonal forecasts: a long-term reference forecast product for the water sector in semi-arid regions. *Earth Syst. Sci. Data* **13**, 2701–2722 (2021).
49. Yan, J., Li, F., Bárdossy, A. & Tao, T. Conditional simulation of spatial rainfall fields using random mixing: a study that implements full control over the stochastic process. *Hydrol. Earth Syst. Sci.* **25**, 3819–3835 (2021).
50. 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. *J. Hydrometeorol.* **22**, 2805–2823 (2021).
51. 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. *J. Hydrometeorol.* **23**, 1383–1399 (2022).
52. Buizza, R. et al. The development and evaluation process followed at ECMWF to upgrade the Integrated Forecasting System (IFS). <https://www.ecmwf.int/node/18658> (2018).
53. (PNL), P. N. N. L. PNNL-WACCEM CAM-MPAS-HR model output prepared for CMIP6 HighResMIP highresSST. <https://doi.org/10.22033/ESGF/CMIP6.14090> (2025).
54. Guo, C. & Berkhahn, F. Entity embeddings of categorical variables. Preprint at arXiv <https://arxiv.org/abs/1604.06737> (2016).
55. Song, Y. et al. Score-based generative modeling through stochastic differential equations. Preprint at arXiv:2011.13456 (2020).
56. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. *Adv. Neural Inform. Process. Syst.* **32**, 11918–11930 (2019).
57. Luo, C. Understanding diffusion models: a unified perspective. Preprint at arXiv <https://arxiv.org/abs/2208.11970> (2022).
58. Bond-Taylor, S. & Willcocks, C. G. ∞ -diff: Infinite resolution diffusion with subsampled mollified states. Preprint at arXiv <https://arxiv.org/abs/2303.18242> (2024).
59. Pidstrigach, J., Marzouk, Y., Reich, S. & Wang, S. Infinite-dimensional diffusion models. Preprint at arXiv <https://arxiv.org/abs/2302.10130> (2023).
60. Zhang, B. & Wonka, P. Functional diffusion. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 4723–4732 (2024).
61. Azizzadenesheli, K. et al. Neural operators for accelerating scientific simulations and design. *Nat. Rev. Phys.* **6**, 320–328 (2024).
62. Biemond, J., Lagendijk, R. L. & Mersereau, R. M. Iterative methods for image deblurring. *Proc. IEEE* **78**, 856–883 (2002).
63. Li, Z. et al. Neural operator: graph kernel network for partial differential equations. Preprint at arXiv <https://arxiv.org/abs/2003.03485> (2020).
64. Kovachki, N. et al. Neural operator: learning maps between function spaces with applications to PDEs. *J. Mach. Learn. Res.* **24**, 1–97 (2023).
65. Li, Z. et al. Fourier neural operator for parametric partial differential equations. Preprint at arXiv:2010.08895 (2020).
66. Tang, H., Liu, Z., Li, X., Lin, Y. & Han, S. TorchSparse: Efficient point cloud inference engine. *Proc. Mach. Learn. Syst.* **4**, 302–315 (2022).
67. Ruiz, N. et al. DreamBooth: Fine-tuning text-to-image diffusion models for subject-driven generation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 22500–22510 (2023).
68. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **114**, 3521–3526 (2017).
69. Chung, H., Kim, J., Mccann, M. T., Klasky, M. L. & Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. Preprint at arXiv <https://arxiv.org/abs/2209.14687> (2024).
70. Chao, J. et al. Learning to infer weather states using partial observations. *J. Geophys. Res. Mach. Learn. Comput.* **2**, e2024JH000260 (2025).
71. Lugmayr, A. et al. RePaint: Inpainting using denoising diffusion probabilistic models. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 11461–11471 (2022).
72. Zhang, G. et al. Towards coherent image inpainting using denoising diffusion implicit models. *Proc. Int. Conf. Mach. Learn.* 41164–41193 (2023).
73. Meng, C. et al. SDEdit: guided image synthesis and editing with stochastic differential equations. Preprint at arXiv <https://arxiv.org/abs/2108.01073> (2022).
74. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
75. Hannachi, A. A primer for EOF analysis of climate data. *Dep. Meteorol. Univ. Read.* **1**, 3 (2004).
76. Fabry, F. On the determination of scale ranges for precipitation fields. *J. Geophys. Res. Atmos.* **101**, 12819–12826 (1996).
77. Pulkkinen, S. et al. Pysteps: an open-source Python library for probabilistic precipitation nowcasting (v1. 0). *Geosci. Model Dev.* **12**, 4185–4219 (2019).

78. Harris, L., McRae, A. T., Chantry, M., Dueben, P. D. & Palmer, T. N. A generative deep learning approach to stochastic downscaling of precipitation forecasts. *J. Adv. Model. Earth Syst.* **14**, e2022MS003120 (2022).
79. Glawion, L., Polz, J., Kunstmann, H., Fersch, B. & Chwala, C. spate-GAN: spatio-temporal downscaling of rainfall fields using a cgan approach. *Earth Space Sci.* **10**, e2023EA002906 (2023).
80. Huffman, G. J. et al. Nasa global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG). *Algorithm Theor. Basis Doc. (ATBD) Version 4*, 30 (2015).
81. Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
82. Rasp, S. et al. Weatherbench: a benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.* **12**, e2020MS002203 (2020).
83. Olivetti, L. & Messori, G. Do data-driven models beat numerical models in forecasting weather extremes? A comparison of ifs hres, pangu-weather, and graphcast. *Geosci. Model Dev.* **17**, 7915–7962 (2024).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant 42130603), National Key R&D Program of China (Grant NO. 2024YFF0809004) and the US National Science Foundation (Grants IIS2324008 and RISE CAIG 2425748). We thank ColorfulClouds Tech. for providing computational support.

Author contributions

S.S. processed the data, developed the models, generated the figures, and wrote the manuscript. C.N. provided guidance on model training and contributed to the manuscript revision. B.P. conceived the overall research framework and provided overall project supervision. W.L. provided the gauge observational dataset. W.L., L.L., and X.L. provided valuable feedback on the manuscript and revisions. E.F.-G. contributed extensively to the manuscript, including early-stage discussions on the research focus and interpretation of results. Y.L. supervised the research, provided funding support, and contributed conceptual guidance to the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-67987-9>.

Correspondence and requests for materials should be addressed to Baoxiang Pan or Yanluan Lin.

Peer review information *Nature Communications* thanks Tim Higgins and the other, anonymous, reviewer for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025