Probabilistic diffusion model for stochastic parameterization – a case example of numerical precipitation estimation

Baoxiang Pan[1], Leyi Wang[2], Feng Zhang[3], Qingyun Duan[4], Xin Li[5], Xiaoduo Pan[5], Xi Chen[1], Fenghua Ling[6], Shuguang Wang[7], Ming Pan[8], Ziniu Xiao[1]

[1]Institute of Atmospheric Physics, Chinese Academy of Sciences
[2]Chongqing Institute of Big Data, Peking University
[3]Department of Atmospheric and Oceanic Sciences, Fudan University
[4]National Key Laboratory of Water Disaster Prevention, Hohai University
[5]Institute of Tibetan Plateau Research, Chinese Academy of Sciences
[6]Institute for Climate and Application Research, Nanjing University of Information Science and Technology
[7]School of Atmospheric Sciences, Nanjing University
[8]Scripps Institution of Oceanography, University of California San Diego

Key Points:

- To learn stochastic parameterization, one should steer generative models toward the requirements of ensemble forecasts.
- We propose criteria of READS (Realism, Efficiency, Adaptability, Diversity, Sharpness) for data-driven stochastic parameterization.
- In case example of numerical precipitation estimation, probabilistic diffusion model well meets the READS criteria.

Corresponding author: Baoxiang Pan, panbaoxiang@lasg.iap.ac.cn

Abstract

Estimating the unresolved geophysical processes from resolved geophysical fluid dynamics is the key for improving numerical weather-climate predictions. While data-driven parameterization for unresolved geophysical processes shows potential, most practices fail to capture the diversity of unresolved geophysical processes that agree with resolved geophysical fluid state. This pitfall undermines the likelihood or severity of simulated weather extremes, and erodes the fidelity of climate projections. We propose the criteria of READS (Realism, Efficiency, Adaptability, Diversity, Sharpness) for generative models to yield reasonable stochastic parameterization. We introduce probabilistic diffusion model, a non-equilibrium thermodynamics inspired deep generative modeling approach, to better meet these criteria. Using a case example of numerical precipitation estimation, we demonstrate the advantage of the proposed methodology in quickly delivering diverse and faithful estimates for the target unresolved process, as compared to other popular data-driven deterministic and stochastic methods (UNet, variational autoencoder, generative adversarial net), as well as dynamical downscaling method (WRF). We conclude that generative models, in particular, probabilistic diffusion model, can significantly enhance the representation of unresolved geophysical processes in numerical weather-climate predictions.

Plain Language Summary

"Life is a gorgeous robe, crawling with lice", so said Eileen Chang, a Chinese writer who enjoyed depicting the awkward discrepancies between ideal and reality. Same metaphor applies to climate models, rooted in physical principles of fluid dynamics and thermodynamics, rife with empirics making up the missing components. We use generative AI to make up the missing components in climate models, achieving realistic and informative simulations of unresolved climate processes, i.e., precipitation.

1 Introduction

Geophysical fluid dynamics operates across a continuous spectrum of spatiotemporal scales, ranging from micro-scale turbulences to synoptic-scale planetary waves. Their numerical solvers, coming with finite resolution, set a distinction between resolved dynamics and unresolved physical processes, with the latter being approximated as empirical functions of the former. This approximation, known as parameterization, is the source of error in numerical weather and climate predictions (Stensrud, 2009).

Typically, parameterization schemes are deterministic functions, providing a unique tendency accounting for the grid-scale impact of subgrid physical processes in numerical modeling of geophysical fluid dynamics. However, as we do not explicitly resolve the subgrid physical processes, a probabilistic formulation is advocated (Berner et al., 2017; T. Palmer, 2019): the impact of subgrid physical processes should be described by a probability distribution function conditioning on the resolved geophysical fluid dynamics. This probabilistic formulation enables a rigorous and consistent characterization of unresolved physical processes across model resolutions (Sakradzija et al., 2016). Also, it allows subgrid-scale noise to trigger crucial circulation regime transitions, supporting reliable probabilistic forecasts (T. N. Palmer et al., 2009).

To make accurate probabilistic representation of unresolved physical processes in numerical weather-climate models, existing efforts proceed along the following three lines. The first, which is straightforward yet lacks theoretical warrant, is to pre-define a perturbation to the parameters, functions, or outputs of deterministic parameterization schemes (Dorrestijn et al., 2013). The second, which is solid yet costly and restricted in scope, is to compute statistics of the equilibrium states of the considered process, via statistical mechanics analysis (Plant & Craig, 2008). The third, which promises remarkable ac-

curacy, efficiency, and flexibility, is to approximate the probability distribution of unresolved physical processes by *learning* from high fidelity data, such as high-resolution simulations or observations (Gagne et al., 2020; Ravuri et al., 2021; Harris et al., 2022).

We proceed along the third line as motivated by the recent breakthrough of probabilistic machine learning, in particular, probabilistic diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020). Probabilistic diffusion models learn to approximate probability distributions in an iterative manner, achieving unprecedented fitting capacity and controlling flexibility in generative modeling tasks. Using a case example of numerical precipitation estimation, we specify five requirements for developing data-driven stochastic parameterization schemes. We develop Diffusion based Precipitation estimator, dubbed DiP, and demonstrate its unique advantages in meeting these requirements, as compared to existing data-driven deterministic and stochastic parameterization schemes, as well as high resolution dynamical simulation method.

2 Problem setup and model requirements

We consider a case example of numerical precipitation estimation: given geophysical fluid dynamics resolved to a finite spatiotemporal resolution, the goal is to estimate the accompanying precipitation process. The challenge lies in that, precipitation results from a complicated chain of processes that are mostly unresolved in numerical models (Tapiador et al., 2019). Any error along this simulation chain may distort the location, timing, or quantity of the precipitation estimate, rendering the estimate useless, even misleading (Pan, Hsu, AghaKouchak, & Sorooshian, 2019; Pan, Hsu, AghaKouchak, Sorooshian, & Higgins, 2019; Chen & Wang, 2022).

Here we consider the region of East and Southeast Asia ($0°-40°$N, $100°$E$-140°$E), where precipitation is driven by diverse circulation regimes. We use the following resolvable dynamical variables to infer precipitation: key primitive variables (meridional and zonal wind velocity, temperature, specific humidity, and geopotential height) at 3 pressure levels (1000/850/500 hPa), and crucial surface level variables (sea level pressure, surface pressure, surface temperature, and total column precipitable water). These data are obtained by blending observations with short-range weather forecasts to faithfully represent historical circulation states. The data are from the Climate Forecast System Reanalysis project (Saha et al., 2006), coming at spatiotemporal resolution of $0.5°/1$ hour for Year 1979-2022. Besides these dynamical variables, we also consider $0.1°$ elevation data as extra, static predictor. These dynamical and static predictor variables are together denoted as $\mathbf{x}$. Precipitation, as our predictand variable, is denoted by $\mathbf{y}$. The data are from the Multi-Source Weighted-Ensemble Precipitation product (Beck et al., 2019), which merges gauge, satellite, and reanalysis precipitation records to achieve optimal quality. The data come at a $0.1°/3$-hourly resolution for same period.

Our objective is to approximate the conditional distribution of $p(\mathbf{y}|\mathbf{x})$, based on favorably large amount of $\{\mathbf{x}, \mathbf{y}\}$ paired data samples. This problem setup differs from a deterministic regression problem setup, which has been widely adopted for learning parameterization schemes (Yu et al., 2023; Wang & Tan, 2023). Specifically, in both deterministic and probabilistic formulations, we design a learning machine Θ, for which the optimal parameter θ^* is obtained by maximizing the overall likelihood of the reference data:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_i \log p_\theta(\mathbf{y}_i|\mathbf{x}_i) \tag{1}$$

In a deterministic regression problem setup, given any $\mathbf{x}$, the learning machine yields the most plausible $\mathbf{y}$: $\hat{\mathbf{y}} = \underset{\mathbf{y}}{\operatorname{argmax}} \log p_{\theta^*}(\mathbf{y}|\mathbf{x})$. This is often achieved by pre-assuming

the distributional form of p_θ. For instance, assuming $p_\theta(\mathbf{y}|\mathbf{x}) := \mathcal{N}\big(\mu_\theta(\mathbf{x}), \sigma^2_\theta(\mathbf{x})\big)$, we have $\hat{\mathbf{y}} = \mu_{\theta^*}(\mathbf{x})$, where $\theta^* = \underset{\theta}{\operatorname{argmin}} \sum_i \Big[\frac{\big(\mu_\theta(\mathbf{x}_i) - \mathbf{y}_i\big)^2}{\sigma^2_\theta(\mathbf{x}_i)} + \log \sigma^2_\theta(\mathbf{x}_i)\Big]$. Such an assumption offers huge computation convenience, yet, it comes with two deficits. First, a predefined distributional form often poorly fits a richly structured target physical process. Second, a deterministic formulation precludes interaction between subgrid noise and resolved dynamics, resulting in biased weather-climate predictions (Hardiman et al., 2022).

In a probabilistic modeling setup, given any $\mathbf{x}$, the learning machine outputs plausible $\mathbf{y}$ samples: $\hat{\mathbf{y}} \sim p_{\theta^*}(\mathbf{y}|\mathbf{x})$, where p_{θ^*} is a learned distribution subject to no predefined probability distribution form. This probabilistic formulation allows us to bypass the two deficits of a deterministic formulation, yet, it comes with its own challenges and requirements. To realize the potential, one must steer the learning machine toward verifiable goals of stochastic parameterization, which are quantified in ensemble forecast practices. We hence suggest the following five criteria for p_{θ^*} based on the requirements of ensemble forecast:

- **Realism**: samples from the estimated conditional probability distribution should be indistinguishable from observational samples, regarding either their structure or functionality. This requirement ensures that an accurate probability value can be assigned to the realized observations, either for training or evaluation purposes. Also, the generated samples can fit into the geophysical modeling pipeline, and be as useful as observations for a wide range of subsequent tasks.
- **Efficiency**: a solid approach for developing parameterization schemes is to consider each of the possible ways that the subgrid scale process evolve under the grid-scale constraint, to compute the probability of each such "configuration" in the equilibrium ensemble, and generate samples accordingly. This requires excessive human effort and computational resources. Here, we expect a well-trained p_{θ^*} to efficiently generate multiple samples of plausible subgrid physical processes, at least several orders faster than directly resolving the subgrid scale process.
- **Adaptability**: the interaction of subgrid scale physics and large scale dynamics often results in organized weather schemes across scales, ranging from local convection to weather fronts. Correspondingly, the model is preferred to automatically identify and apply to these organized weather schemes, rather than working at individual computing grids or fixed computing time steps.
- **Diversity**: the estimated conditional probability distribution should cover all plausible outcomes, rather than a limited subset of modes. This ensures that all observed states are within the *cone* of model simulations, particularly for extremes.
- **Sharpness**: the estimated conditional probability distribution should generate samples that are faithful to the conditioning information, maximizing the sharpness of the simulated distribution. Note that this requirement naturally confronts the diversity requirement: an overly constrained probability estimate may fail to encapsulate observations, which is unreliable; an overly dispersed probability estimate may lack clear distinction from climatology, which is uninformative. We must carefully balance sharpness and diversity, so that the probability estimate faithfully reflects the intrinsic stochasticity of the considered process.

We coin the term READS by concatenating the initial letters of the five criteria above. Below we introduce DiP, Diffusion based Precipitation estimator, and demonstrate its unique advantage in meeting the READS requirements, as compared to existing deterministic/probabilistic data-driven approaches, as well as high-resolution dynamical simulation approach.

3 Diffusion based Precipitation estimator (DiP)

3.1 A primer on probabilistic machine learning

The method we develop here falls into the scope of probabilistic machine learning, which applies probability theory to design learning machines that make predictions as probability distributions. Since the target distribution we try to approximate adheres to no pre-defined closed form, a common strategy is to learn a mapping between the target distribution and a tractable latent distribution, i.e., standard Gaussian. After learning the mapping from optimally large amount of data, we can pass samples from the latent distribution through the trained model to obtain target distribution samples, hence inferring this target distribution. The key challenge is that, we lack point-to-point correspondences between samples from the target distribution and the latent distribution, hence lacking straightforward supervision signals to enable learning (Ruthotto & Haber, 2021). A popular solution is to build bijective mapping between the target distribution and the latent distribution, therefore establishing correspondences. Probabilistic diffusion models excel in this task by establishing the bijection in an iterative manner. Below we outline how this is achieved. Mathematical and implementation details are given in Supporting Information S1.

3.2 Basics

Diffusion model approximates a target distribution $p(\mathbf{y})$ by reversing a Gaussian process (Fig. 1): the forward Gaussian process turns $p(\mathbf{y})$ into standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (Fig. 1a, Eq. 2); we learn to iteratively reverse this Gaussian process, mapping $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to $p(\mathbf{y})$ (Fig. 1b-d), hence achieving generative modeling. Following D. Kingma et al. (2021), the forward Gaussian process is pre-defined as:

$$p(\mathbf{z}_t|\mathbf{y}) := \mathcal{N}(\alpha_t \mathbf{y},\, \sigma_t^2 \mathbf{I}) \tag{2}$$

Here $\mathbf{z}_t$ is latent variable indexed by $t \in [0, 1]$, α_t/σ_t is monotonically decreasing/increasing function of t, strictly bounded by $[0, 1]$. Eq. 2 therefore bridges $p(\mathbf{y}) = p(\mathbf{z}_0|\mathbf{y})$ and $\mathcal{N}(\mathbf{0}, \mathbf{I}) = p(\mathbf{z}_1|\mathbf{y})$ (Fig. 1a). We reverse Eq. 2 to turn $\mathcal{N}(\mathbf{0}, \mathbf{I})$ into $p(\mathbf{y})$, using a chain of variational distributions (Fig. 1b):

$$p_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_{t_i}), \mathbf{\Sigma}_\theta(\mathbf{z}_{t_i})\big), \quad i \in [1, T] \tag{3}$$

Here $0 = t_0 < t_1 < t_2 < ... < t_T = 1$ is arbitrary discretization of time; $\{\mu_\theta, \mathbf{\Sigma}_\theta\}$ are learnable mean vector and covariance matrix, trained by maximizing the overall data likelihood (Supporting Information S1.1):

$$\log p_\theta(\mathbf{y}) = \mathbb{E}_{p(\mathbf{z}_0|\mathbf{y})} \log p(\mathbf{y}|\mathbf{z}_0) - D_{\mathrm{KL}}\big(p(\mathbf{z}_1|\mathbf{y})||p(\mathbf{z}_1)\big) - \sum_{i=1}^{T} \mathbb{E}_{p(\mathbf{z}_{t_i}|\mathbf{y})} D_{\mathrm{KL}}\big(p(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}, \mathbf{y})||p_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i})\big) \tag{4}$$

Given Eq. 2, to maximize Eq. 4 is approximately equivalent to minimizing the Fisher divergence between the data and model distributions (Supporting Information S1.2):

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log p_\theta(\mathbf{y}) \approx \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{T} \mathbb{E}_{p(\mathbf{z}_{t_i}|\mathbf{y})} \big\| \nabla \log p(\mathbf{z}_{t_i}|\mathbf{y}) - \epsilon_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}) \big\|_2 \tag{5}$$

Here $\epsilon_{\mathrm{NN}_\theta}$ is a neural network parameterization of $\nabla \log p(\mathbf{z}_{t_i}|\mathbf{y})$, known as the score function. Based on the learned score estimates, we can derive $p_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_{t_i}), \mathbf{\Sigma}_\theta(\mathbf{z}_{t_i})\big)$ (Supporting Information S1.2) and sample it, starting with $p(\mathbf{z}_1) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, ending with $p(\mathbf{z}_0) \approx p(\mathbf{y})$.

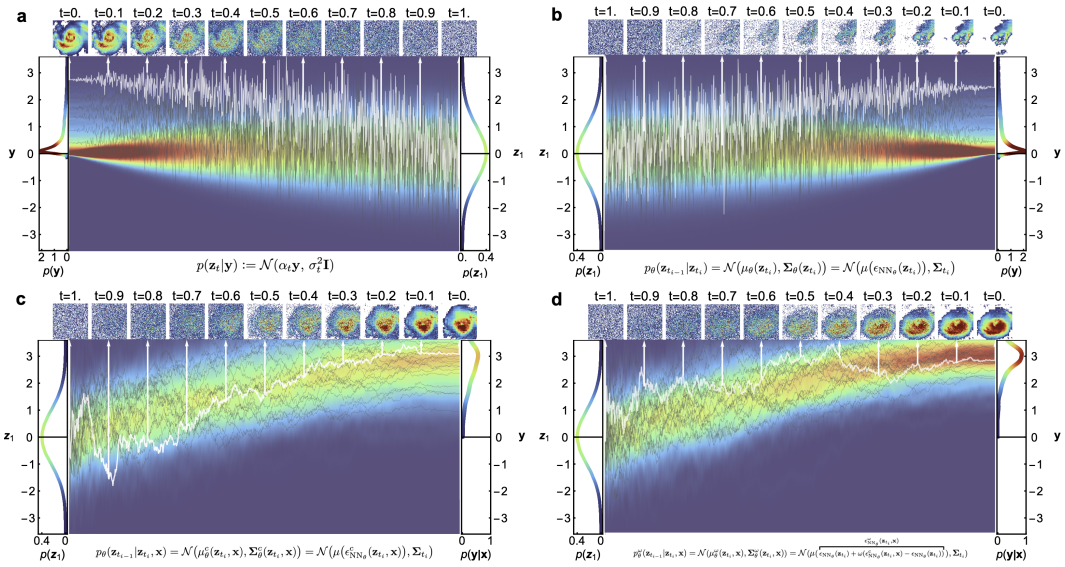


Figure 1. Overview of diffusion model. We map target distribution (synoptic-scale precipitation field, **a** left) to a same dimensional standard Gaussian distribution (**a** right) through a pre-defined Gaussian process (**a** bottom, Eq. 2). Color denotes probability distribution function value for an individual precipitation field pixel (here we select the center pixel) through diffusion time $t = [0, 1]$, lines show the diffusion trajectories of individual pixels for randomly selected samples, matrix plots show the noisified precipitation field (sample of Typhoon Lekima, 0000 UTC 09 August 2019, centered at 26.5°N, 114.4°E) across diffusion time (**a** top). We approximate the target distribution by reversing the Gaussian process, using a series of variational distributions (**b**, Eq. 3), which are trained by maximizing the data likelihood (Eq. 4-5). We include conditioning information to approximate conditional distribution of a same Typhoon event (**c**). We apply *classifier-free guidance* to control the impact of the conditioning information versus the latent variable in explaining the variability of the target variable for the same event (**d**, Eq. 6). By enhancing the guidance strength ω, we suppress the variance of the resulting conditional probability distribution (**c**/**d** right). The plots are supported by logarithm transformed precipitation observational data for Year 2019, and the trained diffusion models.

3.3 Conditioning

To generate $\mathbf{y}$ samples that are faithful to the conditioning information $\mathbf{x}$, we need to approximate the conditional distribution $p(\mathbf{y}|\mathbf{x})$. To achieve this, we include $\mathbf{x}$ during training and sampling (Fig. 1c-d). A direct inclusion of $\mathbf{x}$ does not specify the impact of $\mathbf{x}$ versus $\mathbf{z}_t$ in explaining the variability of $\mathbf{y}$ (Fig. 1c, Holmes & Walker 2017). To tackle the potential misspecification, and having $\mathbf{x}$ effectively control the learned distribution, we resort to *classifier-free guidance* (Ho & Salimans, 2022, Fig. 1d): we learn two sets of neural networks: $\epsilon_{\mathrm{NN}}(\mathbf{z}_{t_i})$ /$\epsilon^c_{\mathrm{NN}}(\mathbf{z}_{t_i}, \mathbf{x})$, so to approximate the unconditional/conditional scores: $\nabla \log p(\mathbf{z}_{t_i}|\mathbf{y})$ /$\nabla \log p(\mathbf{z}_{t_i}|\mathbf{y}, \mathbf{x})$. Based upon these two sets of score estimates, we compose score estimators for synthetic distributions $p_\omega(\mathbf{z}_{t_i}|\mathbf{x}, \mathbf{y}) \propto p(\mathbf{x}|\mathbf{y}, \mathbf{z}_{t_i})^\omega p(\mathbf{z}_{t_i}|\mathbf{y})$:

$$
\begin{aligned}
\nabla \log p_\omega(\mathbf{z}_{t_i}|\mathbf{y}, \mathbf{x}) &= \nabla \log p(\mathbf{z}_{t_i}|\mathbf{y}) + \omega \nabla \log p(\mathbf{x}|\mathbf{y}, \mathbf{z}_{t_i}) \\
&= \nabla \log p(\mathbf{z}_{t_i}|\mathbf{y}) + \omega\big(\nabla \log p(\mathbf{z}_{t_i}|\mathbf{x}, \mathbf{y}) - \nabla \log p(\mathbf{z}_{t_i}|\mathbf{y})\big) \\
&\approx \epsilon_{\mathrm{NN}}(\mathbf{z}_{t_i}) + \omega\big(\epsilon^c_{\mathrm{NN}}(\mathbf{z}_{t_i}, \mathbf{x}) - \epsilon_{\mathrm{NN}}(\mathbf{z}_{t_i})\big)
\end{aligned}
\tag{6}
$$

Here ω is guidance scale coefficient, balancing the diversity and sharpness of the learned conditional distribution:

$\omega = 1$: assuming impact of $\mathbf{x}$ has been perfectly accounted by $\epsilon^c_{\mathrm{NN}}(\mathbf{z}_{t_i}, \mathbf{x})$ (Fig. 1c).
$\omega < 1$: suppressing impact of $\mathbf{x}$, pervading the distribution toward climatology.
$\omega > 1$: raising impact of $\mathbf{x}$, sharpening the distribution toward more likely values (Fig. 1d).

We now apply score estimates of $p_\omega(\mathbf{z}_{t_i}|\mathbf{x}, \mathbf{y})$ to sample $p(\mathbf{y}|\mathbf{x})$, following a same strategy described in Sec. 3.2. The value of ω is empirically determined based on the probabilistic forecasting skill of its resulting model.

3.4 Baselines and implementation details

We compare the DiP methodology with popular deterministic and stochastic data-driven methods and moderate/high resolution dynamical simulation method, including:

- **UNet**: a de-facto choice for image-to-image regression tasks, using neural network consisting symmetric convolution and deconvolution blocks (Ronneberger et al., 2015).
- **Conditional variational autoencoder (CVAE)**: a probabilistic deep learning method that maximizes a lower bound of data likelihood to learn latent variable model for a target conditional distribution (D. P. Kingma & Welling, 2013; Pan et al., 2022).
- **Conditional generative adversarial net (CGAN)**: a probabilistic deep learning method in which a generative network learns to approximate a target conditional distribution, under the guidance of a discriminative network that distinguishes generated samples and true samples (Goodfellow et al., 2014; Pan et al., 2021; Ravuri et al., 2021).
- **CFS reanalysis precipitation product (CFSR)**: an optimized combination of CMAP (CPC Merged Analysis of Precipitation), daily gauge observations, and CFS background 6-hourly precipitation analysis (Saha et al., 2006).
- **Dynamical downscaling using WRF**: refining coarsely resolved climate processes via high resolution numerical geophysical fluid dynamics solver and accompanying parameterization schemes, using Advanced Research Version 4.2 of Weather Research and Forecasting (WRF-ARW V4.2, Skamarock et al. 2019).

For all the data-driven models, including DiP, we use data from 1979-2016/2017-2018/2019-2022 for training/validation/test. Considering the computation cost and the

characteristic scale of atmospheric dynamics, all the data-driven models operate at a synoptic scale (8°×8°): we randomly crop paired predictor and predictand field data within the study region for model training. The model structures, hyper-parameter setups, and training details are given in Supporting Information S2.

3.5 Evaluation

We verify models' performances using a suite of skill metrics corresponding to the READS criteria. We apply Human eYe Perceptual Evaluation (HYPE, Zhou et al. 2019) and power spectrum analysis to determine models' sample fidelity. We use Pearson correlation coefficient (r) and Root Mean Squared Error (RMSE) between observations and models' ensemble mean estimations to quantify models' deterministic prediction skills. We apply Continuous Ranked Probabilistic Skill (CRPS) to measure the accuracy of the predicted probabilities and the sharpness of the forecast distribution. We compute model's skill spread correlation (SSC) to quantify the reliability of a model's uncertainty estimates. We compute the ratio that observations falls into model's ensemble intervals (CR). We record the computing time of the considered models. All the skill metrics are computed across spatial scales from 0.1° to 2° by aggregating neighbourhood grids. For details, see Supporting Information S3.

4 Results

4.1 Case study

We start with a case example to compare models' performances. We consider the storm process associated with Typhoon Lekima, which ranks as the third costliest typhoon in Chinese history. We show $8° \times 8°$ observed and simulated precipitation rate maps along the typhoon trajectory (Fig. 2). Here, observations (Fig. 2a) present a clear ring structure of intense precipitation surrounding the typhoon eye before landing (0000 UTC 04 August 2019 - 0000 UTC 08 August 2019), with maximum precipitation rate reaching 100 mm/h. The eyewall structure gradually dissipates through two landings (1800 UTC 09 August and 1200 UTC 11 August), leaving a tightly curved rainband wrapping into a relatively well-defined centre.

The large-scale patterns of precipitation estimates from the data-driven models (Fig. 2b-e) and CFS reanalysis (Fig. 2f) roughly agree with observations (Fig. 2a), due to a shared circulation constraint from CFS reanalysis. For WRF dynamical downscaling (Fig. 2g), despite careful spectral nudging, the results do not strictly follow the observed typhoon trajectory, particularly after landing (1800 UTC 09 August). This is due to the chaotic nature of geophysical fluid dynamics. The fine-scale structure differs significantly among models: DiP (Fig. 2b) produces the most realistic small-scale details, creating a clear eyewall structure and associated spiral rainband, with intense precipitation matching observations at relatively correct locations. CGAN (Fig. 2c) can generate intense precipitations surrounding the typhoon eye. Yet, the estimates come with poor spatial structure, with neighboring grids loosely correlated, and the rainband barely depictable. CVAE (Fig. 2d) and UNet (Fig. 2e) offer similar, blurry estimates, failing to distinct characteristic typhoon eyewall and rainband structures. Besides, both models miss precipitation extremes, with maximum precipitation estimates below 30 mm/h. CFS reanalysis (Fig. 2f) shares similar drawbacks as CVAE and UNet, largely due to biases from the assimilated data sources and errors from precipitation related model parameterization schemes. WRF simulation (Fig. 2g) makes overly confined, extremely intense (approximately 150 mm/h) precipitation estimates, following the finely resolved, yet potentially misaligned circulation state estimates.

We further inspect the probabilistic models (DiP, CGAN, and CVAE) through the lens of the READS requirements (Sec. 2). For an individual snapshot of precipitation

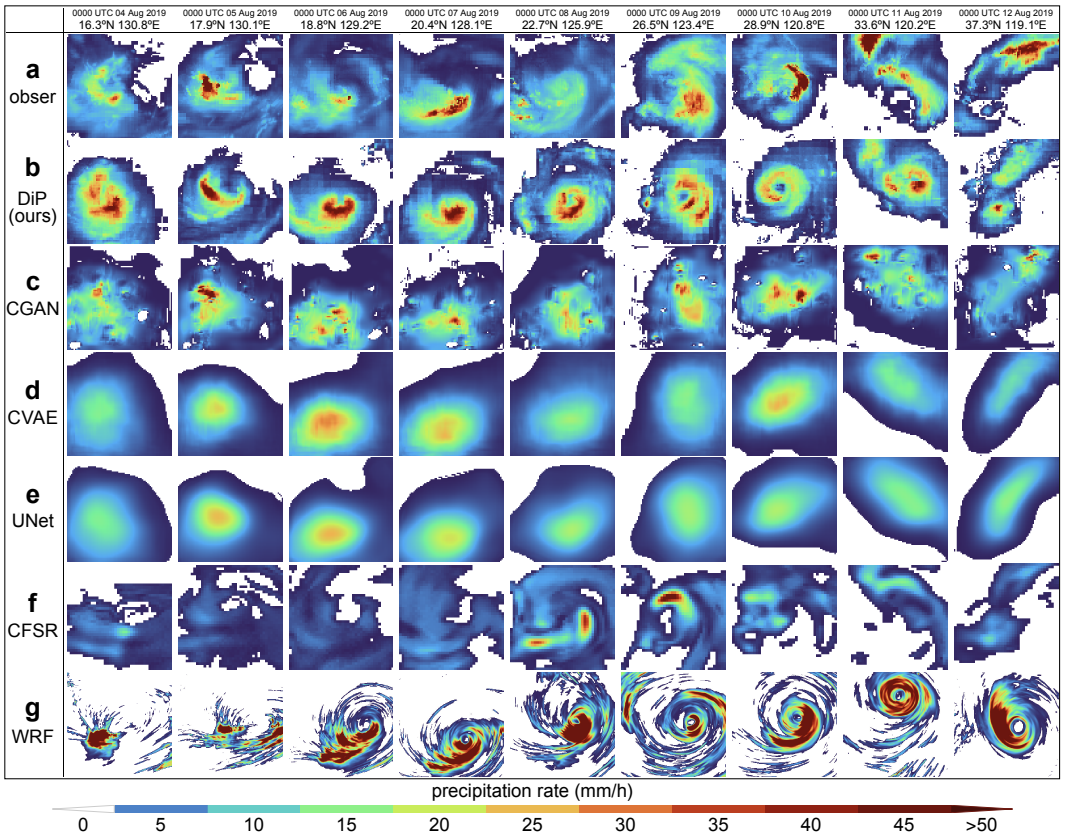


Figure 2. Observed and simulated 8° × 8° precipitation rate maps along the trajectory of Typhoon Lekima, from 0000 UTC 04 August 2019 to 0000 UTC 12 August 2019. a: precipitation observations from MSWEP. b-d: randomly selected samples of ensemble precipitation estimates using DiP/CGAN/CVAE. e: deterministic precipitation estimates using UNet. f: CFS reanalysis precipitation with resolution of 0.2°. g: precipitation estimates using WRF dynamical simulation, with resolution of ∼ 3 km. The typhoon trajectory from WRF simulation considerably diverges from observations after the first landing (1800 UTC 09 August). For after landing results, we show precipitation rate maps surrounding WRF simulated typhoon center.

estimate centering around 22.7°N, 125.9°E at 0000 UTC 06 August 2019, we show models' ensemble members, ensemble mean and standard deviation, ensemble mean absolute error, as well as radial/orientation averaged power spectrum (Fig. 3). We compute a suite of skill metrics corresponding to the READS requirements.

- **Realism**: we measure human climate experts' error rate in detecting observation from model estimates: for DiP/CGAN/CVAE, 3/1/0 out of 5 climate scientist evaluators fail to detect the observation from 15 randomly generated model estimates, suggesting the optimal spatial coherency of DiP estimates. Additionally, we inspect the spatial structure of precipitation estimates by computing their average spectrum power as function of spatial frequency and orientation: DiP and CGAN well reproduce the spatial variability across spatial scales and orientations. Meanwhile, WRF significantly overestimates spatial variability; CVAE, UNet and CFSR significantly underestimate spatial variability for high spatial frequency and all orientations.
- **Efficiency**: all the probabilistic models demonstrate advantageous efficiency compared to high-resolution numerical simulation: DiP/CGAN/CVAE generate 100-member ensemble estimates of 0.1° precipitation field within approximately 100/2/2 seconds on a NVIDIA GeForce RTX 4090 GPU. Here, DiP is two-orders slower than CGAN and CVAE due to its iterative generation nature. As a comparison, a deterministic WRF simulation takes around 5 hours in a 32-core CPU machine.
- **Adaptability**: data-driven models are often reported to struggle with extremes, due to unreasonable learning objective setups, as well as approximation, optimization, and statistical errors. While the typhoon case we consider here is featured by extreme precipitation, DiP successfully reproduces the maximum precipitation rate and characteristic typhoon rainfall structures, suggesting its adaptability for extreme cases. We further report models' performances for various weather schemes in Sec. 4.2.
- **Diversity-Sharpness tradeoff**: we measure the diversity of models' ensemble estimates by computing the percentage that a grid point observation falls into model's ensemble interval. Here, 80.5%/53.6%/29.7% grid point observations are within the 16-member ensemble interval from DiP/CGAN/CVAE. Grid points where observations fall above/below the ensemble interval are stippled with red/black. These results suggest the peculiar advantage of DiP in delivering broad range of plausible outcomes. We further investigate model's sharpness subject to a "proper" level of diversity. By "proper", we mean that the probability estimate accurately reflects the intrinsic stochasticity of the considered process, which is not directly measurable and requires statistical inference. A good indicator is how model's ensemble spread aligns with model's skill. DiP achieves the highest spread-skill correlation, assigning high/low forecast uncertainty estimates to predictions with high/low errors. We further consider the spatial correlation between the ensemble mean estimate and observation, as well as the mean absolute error between each ensemble member and observation. The high skill values of DiP suggest that its ensemble dispersion centers around observation, requiring no ensemble pruning. Finally, we report models' continuous ranked probability scores, which considers both prediction diversity and sharpness. DiP achieves the optimal performance under this proper scoring rule (Gneiting & Raftery, 2007).

4.2 Skill evaluation

We evaluated models' overall performances using test set data from 2019 to 2022. We report a suite of deterministic and probabilistic skill metrics for the considered models in Fig. 4.

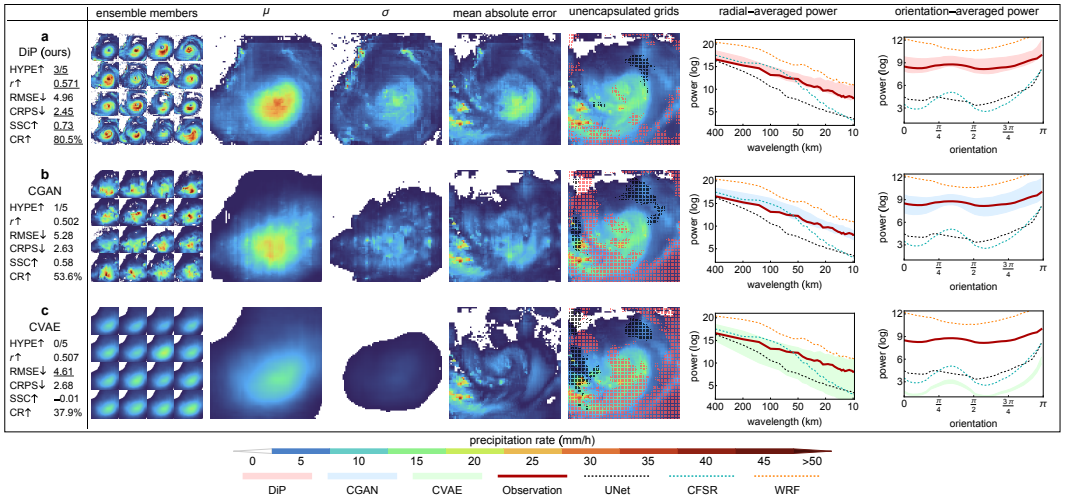


Figure 3. Precipitation estimates centering around 22.7°N, 125.9°E at 0000 UTC 06 August 2019, using DiP (a), CGAN (b), and CVAE (c). The columns show models' ensemble members, ensemble mean, ensemble standard deviation, ensemble mean absolute error, grid points where observation is not encapsulated by ensemble spread (red/black stipple for under/over estimation, background colored based on observation), and radial/orientation averaged power spectrum for observation and all the considered models, including DiP, CGAN, CVAE, UNet, CFS reanalysis, and WRF. The following skill metrics are computed. HYPE: human climate experts' error rate in detecting observation from model estimates; r: spatial correlation between model ensemble mean estimate and observation; RMSE: root mean squared error of model ensemble mean estimate; CRPS: continuous ranked probabilistic score of model ensembles; SSC: spread-skill correlation, where spread is represented using ensemble standard deviation, and skill is represented using model ensemble mean absolute error; CR: coverage ratio, which represent the percentage that grid observation falls into the coverage of ensemble spread.

For deterministic evaluation, we compute the correlation coefficient (r, Fig. 4a) and the root mean squared error (RMSE, Fig. 4b) between observations and models' ensemble mean estimates. We consider spatial scales from 0.1° to 2°, and ensemble size from 8 to 128. For all the considered spatial scales, the data-driven models offer precipitation estimates that are significantly more accurate than the CFS reanalysis precipitation product (dashed lines). This highlights the necessity of learning from high-fidelity data (i.e., observations or high-resolution simulations) to represent unresolved processes in climate modeling. Specific to the data-driven models, DiP and CGAN demonstrates similar r and RMSE skill, matching or slightly falling behind UNet (solid lines). Meanwhile, CVAE offers optimal r and RMSE skill for spatial scales beyond grid-resolution level (0.1°). In principle, a supervised learning approach, i.e., UNet, should provide the optimal deterministic skill. Yet, our results highlight that, for spatial scales that models are not directly trained on, a probabilistic model that better exploit the spatial coherency can outperform a supervised learning model. While CVAE has demonstrated this potential, there is room of progress for DiP and CGAN to further improve their deterministic skills.

For probabilistic evaluation, we compute the continuous ranked probabilistic skill (CRPS, Fig. 4c), the skill-spread correlation (SSC, Fig. 4d), and the coverage ratio (CR, Fig. 4e) of models' ensemble estimates. For CRPS, the CRPS of a deterministic model, i.e., UNet and CFS reanalysis, is equivalent to the model's mean absolute error. Here, DiP, CGAN, and VAE significantly outperforms UNet and CFS reanalysis. At grid-resolution level, for ensemble size of 8, DiP and CGAN perform similarly, both outperforming CVAE by a large margin. As we gradually double the ensemble size, DiP demonstrates slight advantage over CGAN. This advantage becomes more obvious at larger spatial scales. This result suggests that, compared to CGAN, DiP offers more spatially-coherent probabilistic estimates. SSC quantifies the reliability of a model's uncertainty estimates: a higher SSC suggests that the model assigns higher/lower forecast uncertainty estimates to forecasts that turn out to have higher/lower biases, which is crucial for decision makings. DiP achieves the highest SSC for all spatial scales, followed by CGAN. An increase of ensemble size reduces the statistical error of model's uncertainty estimates, hence increases model's SSC. This effect is mostly evident for DiP. CR quantifies the ratio that an observation falls into model's ensemble interval, quantifying how well a probabilistic model is calibrated. Again, DiP achieves the highest CR among the considered models, providing a comprehensive range of plausible outcomes.

To sum up, DiP verifies competitively compared to alternative data-driven deterministic/probabilistic approaches, as well as reanalysis precipitation products: for spatial scales from 0.1° to 2°, DiP matches supervised learning approach in delivering deterministic precipitation estimates (on r and RMSE), and offers optimal probabilistic estimation skills (on CRPS, SSC, and CR). This methodology better meets the READS requirements: it allows us to efficiently generate realistic samples that are faithful to a broad range of resolved circulation schemes, and are diverse to cover most plausible outcomes.

5 Conclusions

Numerical weather-climate models resolve geophysical fluid dynamics to a finite resolution, necessitating probabilistic inference for unresolved processes. For example, what is the probability that, at millimeter scale, various hydrometeors interact, collide, coalesce to yield precipitation, given circulation status resolved to kilometer scale? If we could accurately and efficiently answer these questions, we could not only better understand, but also better predict the climate.

We follow the data-driven ideology to learn representations of unresolved climate processes from high fidelity data, such as high-resolution simulations and observations.

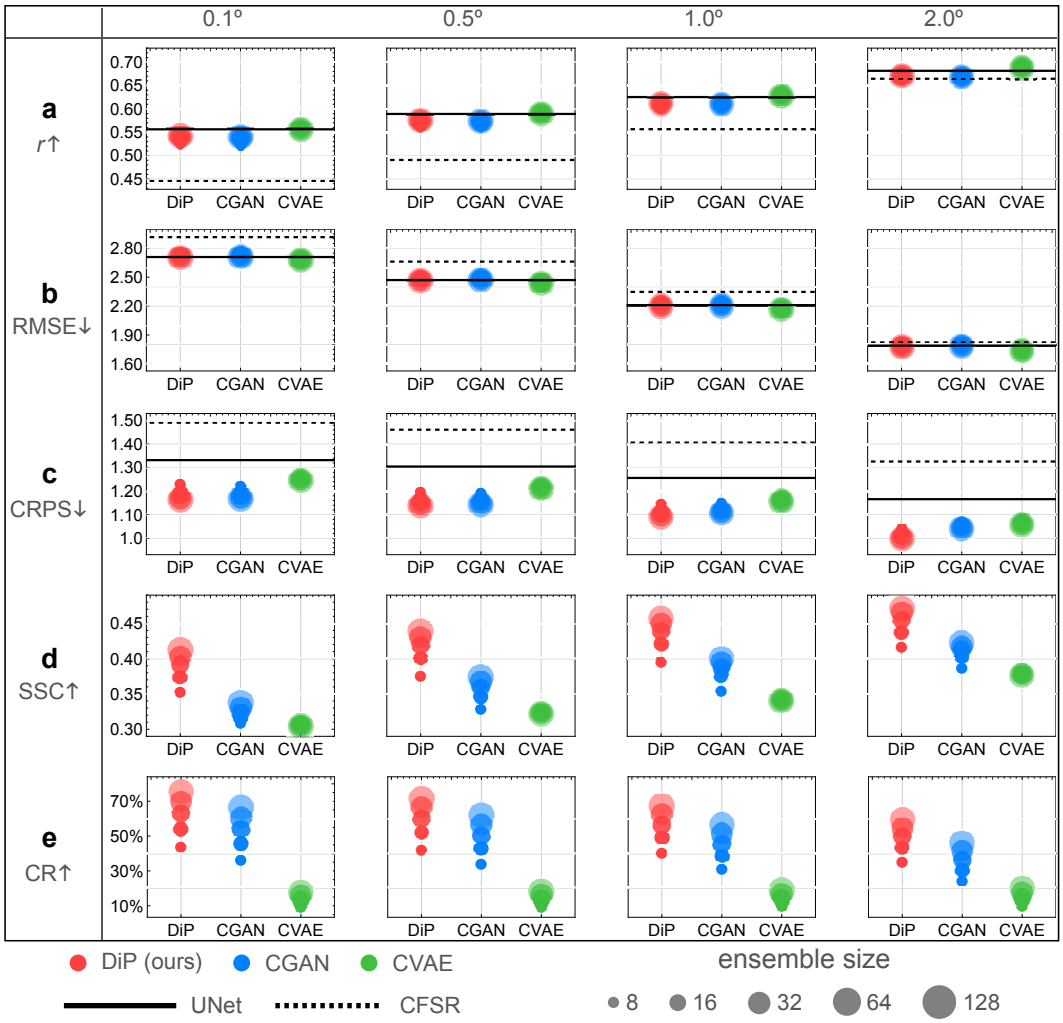


Figure 4. Performance evaluation using data from 2019 to 2022. The following skill metrics are considered. r: average correlation coefficient between model ensemble mean estimates and observations; RMSE: root mean squared error of model ensemble mean estimate; CRPS: continuous ranked probabilistic score of model ensembles; SSC: spread-skill correlation, where spread is represented using ensemble standard deviation, and skill is represented using model ensemble mean absolute error; CR: coverage ratio, which represents the percentage that grid observation falls into the coverage of ensemble spread. For the probabilistic models, we consider ensemble size from 8 to 128 to compute the skill metrics. All the skill metrics are computed across spatial scales from 0.1° to 2° by spatial pooling.

We point out the limitations of supervised learning approaches in such tasks, and advocate the potential advantages of generative modeling approaches.

To realize these potential advantages, we should steer the learning machine toward verifiable goals of stochastic parameterization, which are quantified in ensemble forecast practices. Hence, based on the requirements of ensemble forecast, we propose the READS (Realism, Efficiency, Adaptability, Diversity, and Sharpness) criteria for probabilistic representation of unresolved climate processes.

To solidify these arguments and provide practical solutions, we consider the problem of numerical precipitation estimation. We develop DiP, a probabilistic diffusion model based methodology to learn stochastic parameterization of precipitation. Compared to existing generative models, DiP approximates a target distribution in a principled, iterative manner, which offers it tremendous fitting capability and controlling flexibility.

Using a Typhoon storm case and four-year evaluation, we demonstrate the advantage of DiP in meeting the READS requirements, as compared to existing data-driven supervised deep learning method (UNet), data-driven probabilistic deep learning method (CVAE and CGAN), as well we moderate/high resolution numerical method (CFS and WRF).

There remain several challenges for our approach to stochastic parameterization. Till now, our model does not provide feedback to the resolved dynamics. It remains to be examined if the learned subgrid-scale noise can trigger circulation regime transitions, and support reliable probabilistic forecast. Also, the ensemble mean estimate from DiP fails to match the performance of CVAE, suggesting room for progress. Finally, to generate large ensemble estimates using DiP takes hundreds runs of the deep nets, which brings considerable computation burden in long term simulations. Future works may explore diffusion model distillation techniques to accelerate the generation process (Salimans & Ho, 2022; Song et al., 2023).

Acknowledgments

This research is supported by National Key R&D Program of China (2021YFA0718000), National Natural Science Foundation of China (42275174, 42288101) and Chinese Academy of Science Light of the West Interdisciplinary Research Grant (xbzg-zdsys-202104). We thank Dr. Juanjuan Liu, Dr. Li Dong, Dr. Guiwan Chen, Mr. Jie Chao, and Mr. Yucheng Zi for supporting the Human eYe Perceptual Evaluation. The Multi-Source Weighted-Ensemble Precipitation data are available from https://www.gloh2o.org/mswep/. The Climate Forecast System Reanalysis data are available from https://climatedataguide.ucar.edu/climate-data/climate-forecast-system-reanalysis-cfsr.

References

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I., ... Adler, R. F. (2019). Mswep v2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment. *Bulletin of the American Meteorological Society*, *100*(3), 473–500.

Berner, J., Achatz, U., Batte, L., Bengtsson, L., De La Camara, A., Christensen, H. M., ... others (2017). Stochastic parameterization: Toward a new view of weather and climate models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588.

Chen, G., & Wang, W.-C. (2022). Short-term precipitation prediction for contiguous united states using deep learning. *Geophysical Research Letters*, *49*(8), e2022GL097904.

Dorrestijn, J., Crommelin, D. T., Siebesma, A. P., & Jonker, H. J. (2013). Stochas-

tic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theoretical and Computational Fluid Dynamics*, *27*, 133–148.

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, *102*(477), 359–378.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

Hardiman, S. C., Dunstone, N. J., Scaife, A. A., Smith, D. M., Comer, R., Nie, Y., & Ren, H.-L. (2022). Missing eddy feedback may explain weak signal-to-noise ratios in climate predictions. *npj Climate and Atmospheric Science*, *5*(1), 57.

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

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

Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Holmes, C. C., & Walker, S. G. (2017). Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, *104*(2), 497–503.

Kingma, D., Salimans, T., Poole, B., & Ho, J. (2021). Variational diffusion models. *Advances in neural information processing systems*, *34*, 21696–21707.

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Palmer, T. (2019). Stochastic weather and climate models. *Nature Reviews Physics*, *1*(7), 463–471.

Palmer, T. N., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G. J., ... Weisheimer, A. (2009). Stochastic parametrization and model uncertainty.

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

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J., Lee, J., ... Ma, H.-Y. (2021). Learning to correct climate projection biases. *Journal of Advances in Modeling Earth Systems*, *13*(10), e2021MS002509.

Pan, B., Hsu, K., AghaKouchak, A., & Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*(3), 2301–2321.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., & Higgins, W. (2019). Precipitation prediction skill for the west coast united states: From short to extended range. *Journal of Climate*, *32*(1), 161–182.

Plant, R., & Craig, G. C. (2008). A stochastic parameterization for deep convection based on equilibrium statistics. *Journal of the Atmospheric Sciences*, *65*(1), 87–105.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., ... others (2021). Skilful precipitation nowcasting using deep generative models of radar. *Nature*, *597*(7878), 672–677.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–miccai 2015: 18th international conference, munich, germany, october 5-9, 2015, proceedings, part iii 18* (pp. 234–241).

Ruthotto, L., & Haber, E. (2021). An introduction to deep generative modeling. *GAMM-Mitteilungen*, *44*(2), e202100008.

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., ... others (2006). The ncep climate forecast system. *Journal of Climate*, *19*(15), 3483–3517.

Sakradzija, M., Seifert, A., & Dipankar, A. (2016). A stochastic scale-aware parameterization of shallow cumulus convection across the convective gray zone. *Journal of Advances in Modeling Earth Systems*, *8*(2), 786–812.

Salimans, T., & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Liu, Z., Berner, J., ... others (2019). A description of the advanced research wrf version 4. *NCAR tech. note ncar/tn-556+ str*, *145*.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning* (pp. 2256–2265).

Song, Y., Dhariwal, P., Chen, M., & Sutskever, I. (2023). Consistency models.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

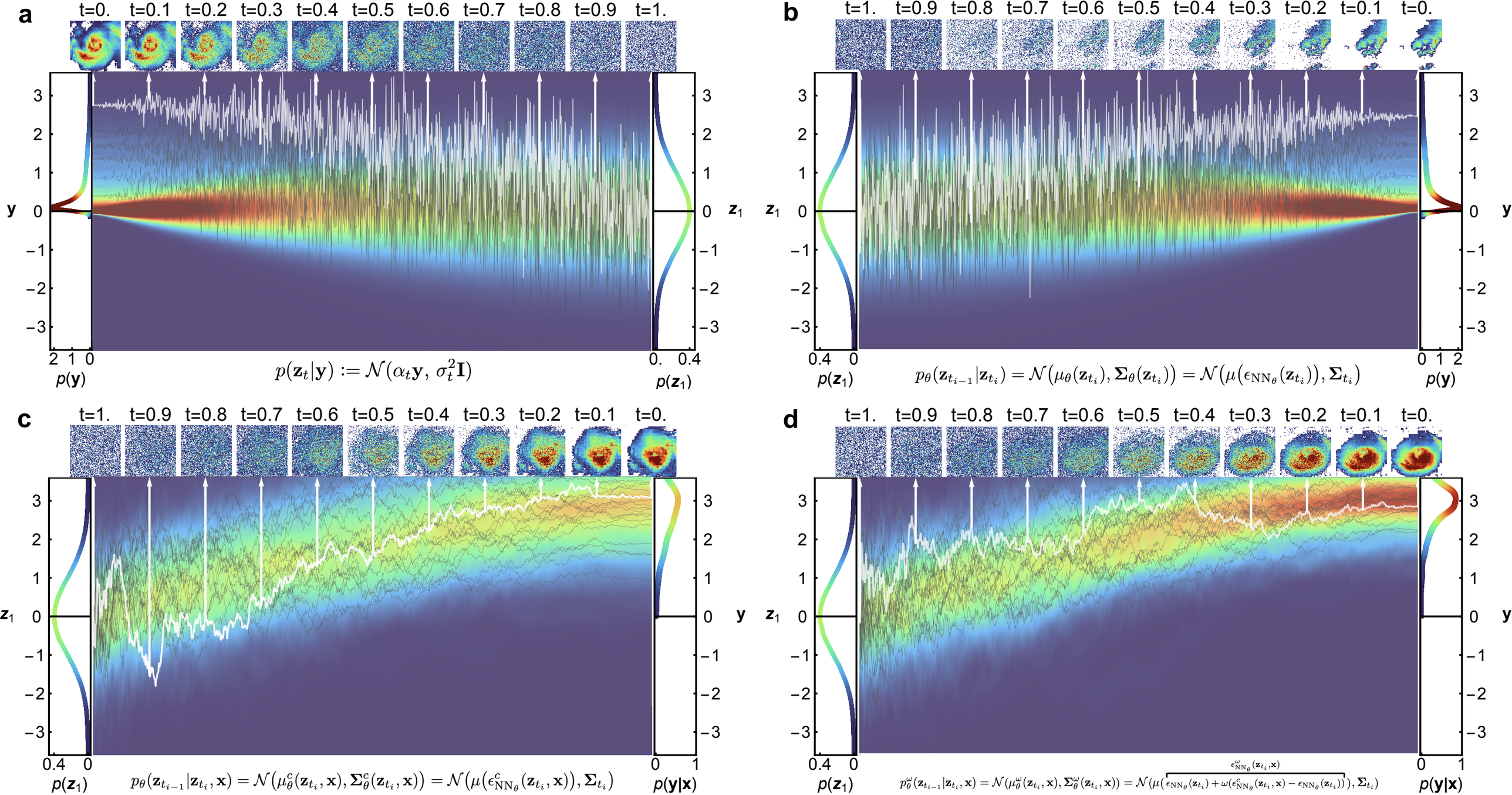
Stensrud, D. J. (2009). *Parameterization schemes: keys to understanding numerical weather prediction models*. Cambridge University Press.

Tapiador, F. J., Roca, R., Del Genio, A., Dewitte, B., Petersen, W., & Zhang, F. (2019). Is precipitation a good metric for model performance? *Bulletin of the American Meteorological Society*, *100*(2), 223–233.

Wang, L.-Y., & Tan, Z.-M. (2023). Deep learning parameterization of the tropical cyclone boundary layer. *Journal of Advances in Modeling Earth Systems*, *15*(1), e2022MS003034.

Yu, S., Hannah, W. M., Peng, L., Bhouri, M. A., Gupta, R., Lin, J., ... others (2023). Climsim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. *arXiv preprint arXiv:2306.08754*.

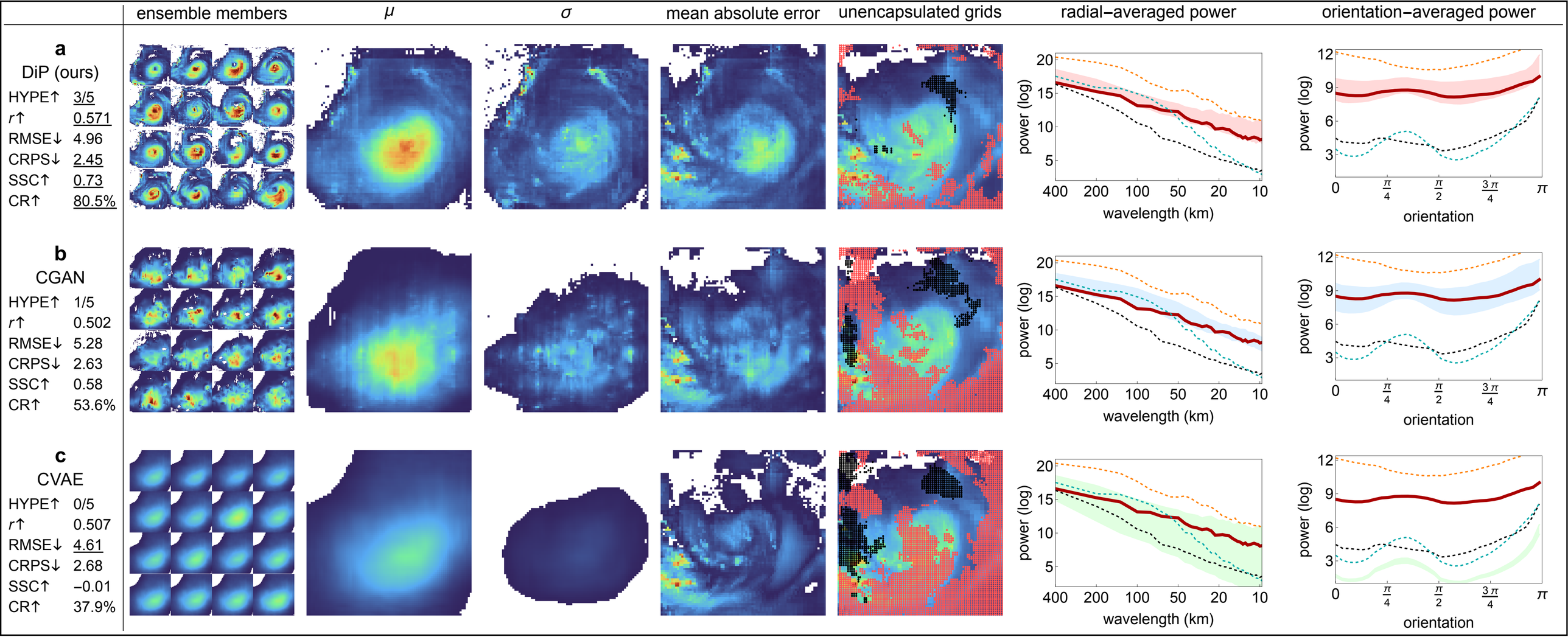
Zhou, S., Gordon, M., Krishna, R., Narcomey, A., Morina, D., & Bernstein, M. S. (2019). Hype: human-eye perceptual evaluation of generative models.

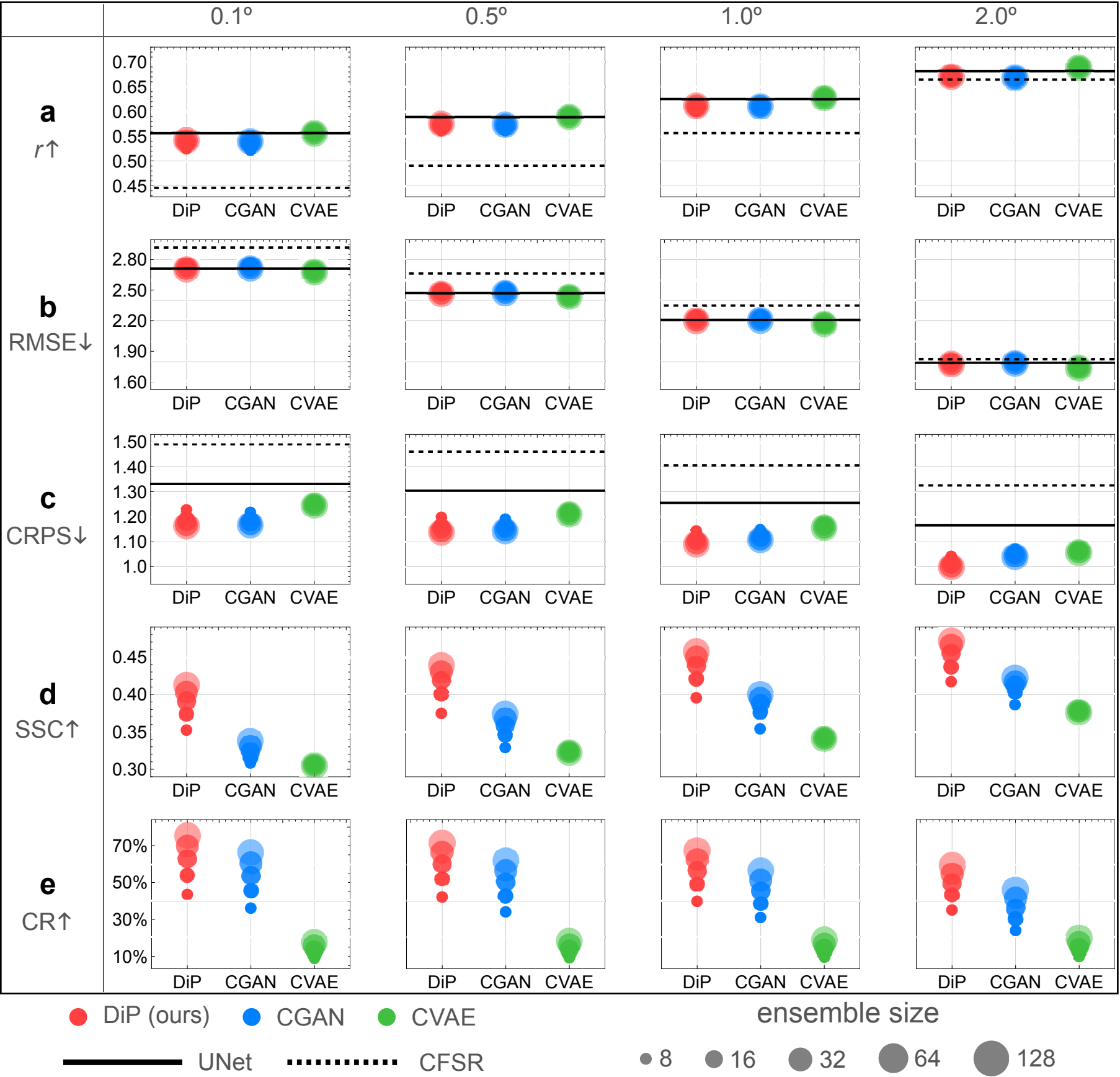

a
t=0.
t=0.1
t=0.2
t=0.3
t=0.4
t=0.5
t=0.6
t=0.7
t=0.8
t=0.9
t=1.
y
z_1
p(y)
$p(z_1)$
$p(\mathbf{z}_t|\mathbf{y}) := \mathcal{N}(\alpha_t \mathbf{y},\, \sigma_t^2 \mathbf{I})$
b
t=1.
t=0.9
t=0.8
t=0.7
t=0.6
t=0.5
t=0.4
t=0.3
t=0.2
t=0.1
t=0.
z_1
y
$p(z_1)$
p(y)
$p_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}) = \mathcal{N}\big(\mu_\theta(\mathbf{z}_{t_i}), \mathbf{\Sigma}_\theta(\mathbf{z}_{t_i})\big) = \mathcal{N}\big(\mu\big(\epsilon_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i})\big), \mathbf{\Sigma}_{t_i}\big)$
c
t=1.
t=0.9
t=0.8
t=0.7
t=0.6
t=0.5
t=0.4
t=0.3
t=0.2
t=0.1
t=0.
z_1
y
$p(z_1)$
p(y|x)
$p_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}, \mathbf{x}) = \mathcal{N}\big(\mu^c_\theta(\mathbf{z}_{t_i}, \mathbf{x}), \mathbf{\Sigma}^c_\theta(\mathbf{z}_{t_i}, \mathbf{x})\big) = \mathcal{N}\big(\mu\big(\epsilon^c_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}, \mathbf{x})\big), \mathbf{\Sigma}_{t_i}\big)$
d
t=1.
t=0.9
t=0.8
t=0.7
t=0.6
t=0.5
t=0.4
t=0.3
t=0.2
t=0.1
t=0.
z_1
y
$p(z_1)$
p(y|x)
$p^\omega_\theta(\mathbf{z}_{t_{i-1}}|\mathbf{z}_{t_i}, \mathbf{x}) = \mathcal{N}(\mu^\omega_\theta(\mathbf{z}_{t_i}, \mathbf{x}), \mathbf{\Sigma}^\omega_\theta(\mathbf{z}_{t_i}, \mathbf{x})) = \mathcal{N}(\mu(\overbrace{\epsilon_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}) + \omega(\epsilon^c_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}, \mathbf{x}) - \epsilon_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}))}^{\epsilon^\omega_{\mathrm{NN}_\theta}(\mathbf{z}_{t_i}, \mathbf{x})}), \mathbf{\Sigma}_{t_i})$

	0000 UTC 04 Aug 2019 16.3ºN 130.8ºE	0000 UTC 05 Aug 2019 17.9ºN 130.1ºE	0000 UTC 06 Aug 2019 18.8ºN 129.2ºE	0000 UTC 07 Aug 2019 20.4ºN 128.1ºE	0000 UTC 08 Aug 2019 22.7ºN 125.9ºE	0000 UTC 09 Aug 2019 26.5ºN 123.4ºE	0000 UTC 10 Aug 2019 28.9ºN 120.8ºE	0000 UTC 11 Aug 2019 33.6ºN 120.2ºE	0000 UTC 12 Aug 2019 37.3ºN 119.1ºE
a obser									
b DiP (ours)									
c CGAN									
d CVAE									
e UNet									
f CFSR									
g WRF									

precipitation rate (mm/h)

0 5 10 15 20 25 30 35 40 45 >50


ensemble members
μ
σ
mean absolute error
unencapsulated grids
radial−averaged power
orientation−averaged power
a
DiP (ours)
HYPE↑ 3/5
r↑ 0.571
RMSE↓ 4.96
CRPS↓ 2.45
SSC↑ 0.73
CR↑ 80.5%
b
CGAN
HYPE↑ 1/5
r↑ 0.502
RMSE↓ 5.28
CRPS↓ 2.63
SSC↑ 0.58
CR↑ 53.6%
c
CVAE
HYPE↑ 0/5
r↑ 0.507
RMSE↓ 4.61
CRPS↓ 2.68
SSC↑ −0.01
CR↑ 37.9%
power (log)
wavelength (km)
orientation
precipitation rate (mm/h)
0 5 10 15 20 25 30 35 40 45 >50
DiP
CGAN
CVAE
Observation
UNet
CFSR
WRF


0.1°
0.5°
1.0°
2.0°
a
r↑
b
RMSE↓
c
CRPS↓
d
SSC↑
e
CR↑
DiP
CGAN
CVAE
DiP (ours)
UNet
CFSR
ensemble size
8
16
32
64
128