

Diffusion model-based ensemble generation for highly nonlinear systems with limited ensemble size

Seoleun Shin ^{*}*Korea Research Institute of Standards and Science, Daejeon 34113, Republic of Korea*

(Received 28 July 2025; accepted 23 July 2026; published 21 August 2026)

We propose a computational strategy that incorporates time-lagged sampling into the reverse process of diffusion-based machine learning models. This approach aims to maintain ensemble diversity and prevent ensemble collapse related to underestimation of uncertainties, particularly in situations with limited ensemble sizes. We demonstrate the method's effectiveness using the Lorenz '96 system as a theoretical testbed, focusing on its highly nonlinear regime without explicitly resolving multiscale interactions. Our approach facilitates stable forecast-analysis cycles at small ensemble sizes without requiring covariance inflation or localization by enhancing the ensemble representation of uncertainties inherent in diffusion models and stochastic dynamics. The proposed models maintain superior temporal coherence. While the probability density function is narrowed under severely undersampled ensembles, stochastic variants recover distributional accuracy as ensemble size increases, whereas deterministic variants show persistent limitations but provide substantial computational efficiency. The method generalizes successfully to systems with increased multiscale complexity and maintains stable long-term performance without drift. Results indicate that active diversity preservation in ensemble generation can enable robust data assimilation in undersampled situations, with potential applications in geophysical flows, plasma systems, and turbulence modeling.

DOI: [10.1103/xq7z-9cg5](https://doi.org/10.1103/xq7z-9cg5)

I. INTRODUCTION

Recent studies suggest that state-of-the-art machine learning techniques can produce forecasts that are comparable to, or in some cases exceed, those from traditional numerical weather prediction (NWP) systems, when evaluated using specific performance metrics (e.g., Refs. [1,2]). The gain of the data-driven approach might be also highlighted in the context of its speed-up for a high-resolution prediction system requiring enormous degrees of freedom to represent a dynamical system properly. However, recent studies also discuss actively on the limitation of machine learning techniques especially regarding to its ability to capture extreme events such as high-impact weather (e.g., Refs. [3–6]). Also, concerns have been raised about the blurry ensemble mean and the corresponding power loss at high-frequency scales [1,2,7]. Models may systematically under- or overpredict certain variable states, leading to misrepresented uncertainty estimates. These challenges pose significant difficulties not only for physical models but also for data-driven approaches, and they affect both ensemble prediction and the data assimilation process. Moreover, the performance of the ensemble data assimilation depends on the structure of the ensemble as well as the dynamic characteristics of targeting systems [8]. This study investigates strategies to address these challenges for

a highly nonlinear system, by employing a diffusion-based machine learning model within a simplified ensemble data assimilation platform. To isolate model performance, we deliberately omit ensemble inflation and covariance localization, allowing the analysis to reveal intrinsic model characteristics unmasked by common data assimilation tuning techniques.

Time-series prediction models for physical values in various fields can be applied across a wide range of domains, such as traffic, transportation, air quality, energy consumption, financial markets, and weather and climate, provided they are trained on the relevant data. Depending on the nature of data, the focus can shift to spatiotemporal correlations, the coupling of different components, and functional dependencies, and so on. In this study, we concentrate on the prediction of highly nonlinear dynamic systems and the application of data-driven models for the purpose related to the spatiotemporal correlations. One of the most prominent examples of such systems is the Lorenz '96 model [9], a simplified theoretical emulator of a global weather or climate system. The Lorenz '96 model is widely used as a testbed to examine how well new algorithms handle dynamics that exhibit nonlinear and chaotic behaviors. The system is regarded as a simplified framework for real-world atmosphere in which multiscale motions interact in a complex manner. For example, a single-time scale model has been developed by parametrizing the fast time-scale motions into a machine learning model that solves only slow motions of the Lorenz '96 system [10]. Also, a fully connected neural network model was implemented [11] to explore whether machine learning-based models could rival conventional physical models in global weather forecasting, without needing to account for the interaction between slow and fast motions.

^{*}Contact author: seoleuns@kriss.re.kr

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Meanwhile, the use of machine learning models as predictive tools within an ensemble data assimilation framework has been explored, and their forecasting performance has been evaluated [12,13]. However, it remains questionable whether machine learning models can capture extreme local events or dynamical instabilities with rapid transitions with limited size of ensembles in this framework. It could be challenging for data-driven models to generate outputs that include states with low probability of occurrence. Ensemble-based data assimilation methods such as the ensemble Kalman filter (EnKF) are known to suffer from sampling errors and rank-deficient background error covariance estimates when the ensemble size is small relative to the state dimension (e.g., Ref. [14]). These limitations are conventionally mitigated through covariance inflation and localization, whose configuration typically requires extensive case-by-case tuning. Meanwhile, the impact of model dynamics on performance is considered significant, and such a study is essential for the development of accurate and robust EnKF [8]. Therefore, investigating how machine learning models affect the performance of ensemble data assimilation in simplified setups like the Lorenz '96 system could provide valuable insights for more complex, real-world applications such as weather prediction, climate modeling, fluid dynamics, and other chaotic systems. It would also be important to explore how machine learning models can be designed to optimize the ensemble by effectively capturing uncertainties in the generated predictions.

We have chosen a generative diffusion model that is known to be adept at reconstructing detailed information for targeting systems, based on its specific implementations in diverse fields (e.g., Refs. [5,15,16]). Especially, the use of diffusion models for time-series prediction has garnered significant attention across various fields [17]. Among the generative diffusion models for time-series prediction models, we adopt the denoising diffusion-based probabilistic spatiotemporal graph forecasting method (DiffSTG) [18] as the basis for our approach, taking advantage of its ability to capture complex spatiotemporal dependencies within the data. Since the predictions of diffusion models are heavily dependent on the reverse sampling process, we concentrate on optimizing the process for better representations of uncertainties in ensemble predictions. These challenges can affect the stability and accuracy of ensemble data assimilation, especially in situations with small ensembles. Our focus is particularly directed toward the potential issues of exposure bias due to the discrepancy between training and inference [19]. To increase the possibility of capturing samples in the distribution of probable outcomes, we transfer the concept of the lagged average forecasting (LAF) suggested by Refs. [20,21] aiming for stable estimation with sparse data [22]. More information on the customization of the model for the purpose of this study can be found in Sec. II. Additionally, detailed explanations of the multilevel Lorenz '96 systems and the parameters chosen for the tests are provided in Sec. II. The data assimilation process is also introduced in this section. In Sec. III, improvements by the approach in the sampling process will be described. In Sec. IV, we offer a discussion on the generation of ensemble for diffusion-based machine learning models with limited ensemble size.

II. METHODS

A. Modifications to DiffSTG

We have customized the DiffSTG [18] to reproduce the Lorenz '96 dynamics in an ensemble framework. Major customizations include (i) random shifting of the denoising step positions from which states are sampled during the reverse processes of denoising diffusion implicit models (DDIM) [23] and denoising diffusion probabilistic models (DDPM) [24], (ii) addition of attention mechanism [25] for the learning of temporal patterns, and (iii) the use of data and predictions in unnormalized form. Neither the input features nor the model outputs undergo normalization; all values are retained in their original scale. Here, we focus on describing the change by the (i) random shifting, which is implemented to enable more diverse state sampling since generative outputs from a diffusion model is highly dependent of the scheduling of reverse processes. This strategy is also important for capturing both short- and long-term dependencies in prediction by sequential sampling with temporal correlation. First, we describe the sampling process applied in Ref. [18] and explain the approach (i) that we have implemented for this study in the following section. For reference, a summary of the notation used throughout this study is provided in Table II of Appendix A.

B. Sampling process

The DiffSTG framework provides options for both DDPM and DDIM reverse processes [18]. In our framework, the reverse (denoising) process is used to generate forecast samples of the future system state. The sampling of the reverse process is explained here in a general description as given by [23]:

$$\mathbf{x}_{t_{i-1}} = \sqrt{\alpha_{t_{i-1}}} \left(\frac{\mathbf{x}_{t_i} - \sqrt{1 - \alpha_{t_i}} \epsilon_{\theta}^{(t_i)}(\mathbf{x}_{t_i})}{\sqrt{\alpha_{t_i}}} \right) + \sqrt{1 - \alpha_{t_{i-1}} - \sigma_{t_i}^2} \cdot \epsilon_{\theta}^{(t_i)}(\mathbf{x}_{t_i}) + \sigma_{t_i} \epsilon_{t_i}. \quad (1)$$

Here, $\mathbf{x}_{t_{i-1}}$ is the sample state generated from $\mathbf{x}_{t_i}$ by the reverse process for $i = \{1, \dots, T\}$. We emphasize that the index i , and the associated step variable t_i , labels the discrete steps of the diffusion denoising process only; it bears no correspondence to the physical time t of the Lorenz '96 system, which evolves continuously and whose temporal evolution is described separately in Sec. II C. The reverse process thus transforms noise into a forecast sample at a fixed physical time, rather than advancing the system forward in physical time. In the forward process, Gaussian noise β_{t_i} is added at each step, and the scaling parameter is defined as $\alpha_{t_i} := 1 - \beta_{t_i}$ in such a way that $\alpha_0 = 1$ and $\alpha_T = 0$. To predict x_0 , we use $\epsilon_{\theta}^{(t_i)}$, a set of T functions with trainable parameters θ . Random noises can be given by the product of σ_{t_i} and the standard Gaussian noise $\epsilon_{t_i} \sim \mathcal{N}(0, 1)$ to make the generative process probabilistic in DDPM. Meanwhile, setting $\sigma_{t_i} = 0$ can increase sampling efficiency by generating samples from latent variables with a fixed procedure [23]. In this special case, the generative process becomes deterministic, and thus it is called denoising diffusion implicit models (DDIM). Furthermore, if a subset sequence $s = \{1, \dots, \tau\}$, where $\tau < T$ is used instead of the full sequence with T iterations, the generative process can

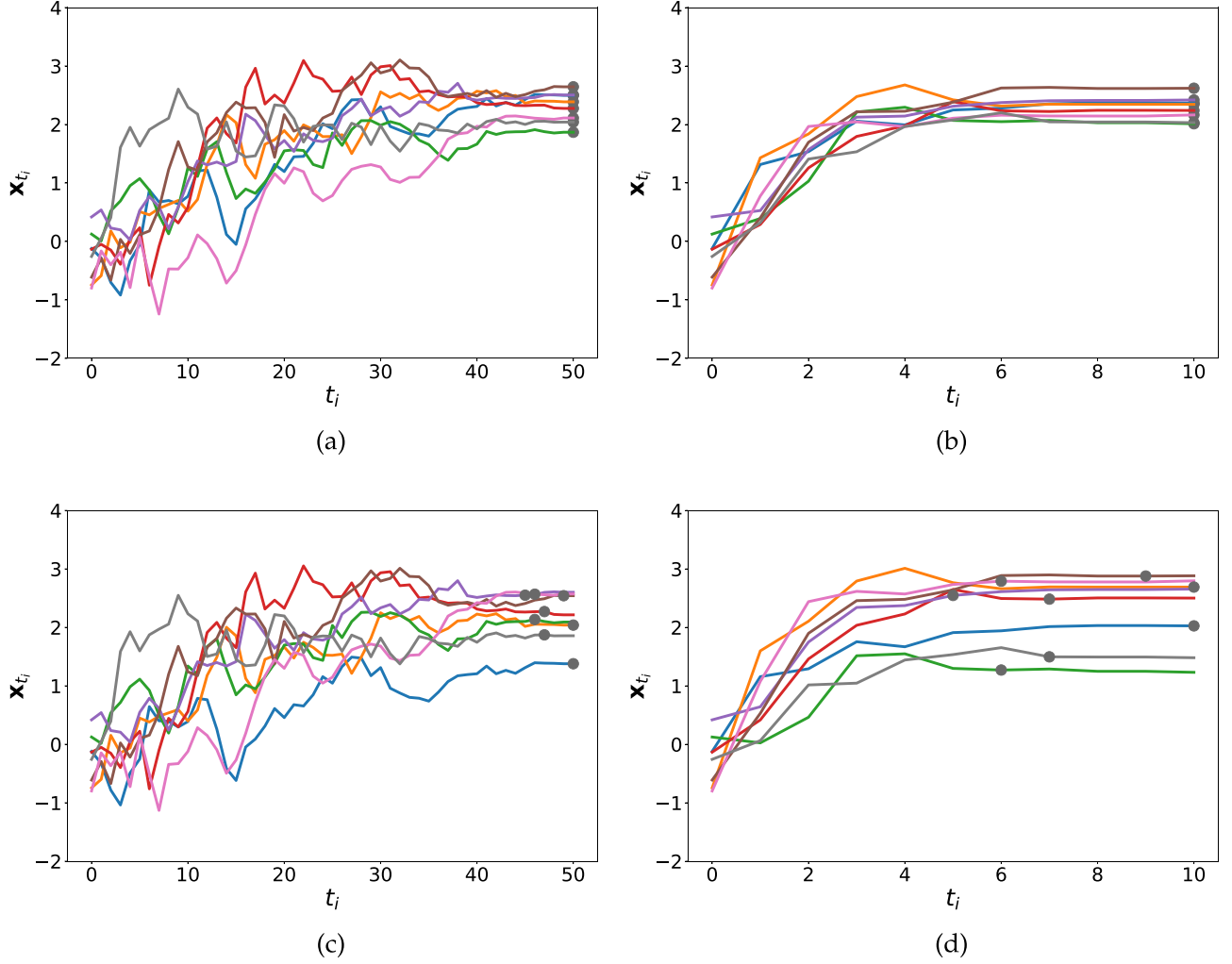


FIG. 1. Example of the proposed sampling strategy. Each panel shows trajectories of eight ensemble members during the reverse denoising process (x axis: reverse process step t_i , y axis: state value of $\mathbf{x}_{t_i}$ at grid point $k = 20$) at integration time step 100. Standard approaches [(a) DDPM, (b) DDIM] sample all ensemble members from a single fixed position (circles at rightmost step), while the proposed methods [(c) LAFDDPM, (d) LAFDDIM] sample from randomly distributed positions (circles at various steps), thereby enhancing ensemble diversity.

be further accelerated. By this non-Markovian way, sampling proceeds only at a smaller number of steps than those used in the forward process to approximate a Gaussian distribution.

The performance of diffusion-based models can be substantially affected by the design of the reverse process [16,26]. In particular, diffusion models tend to have higher sensitivity to the later stages of the reverse process, where minor prediction errors can significantly affect overall accuracy [19]. This sensitivity has been attributed to exposure bias, which can be alleviated by using iterative sampling strategies such as the “time-shift sampler” [19].

Building on these insights and motivated by the lagged average forecasting (LAF) framework, we propose a sampling strategy that enhances ensemble diversity while maintaining computational efficiency. In the ensemble data assimilation context, where multiple realizations are needed to represent prediction uncertainties, it is natural to leverage the stochastic nature of the diffusion process by sampling from different stages of the reverse process. Our approach randomly selects a reverse process time step to sample from at each prediction

time [Figs. 1(c) and 1(d)]. Specifically, instead of sampling from the same fixed reverse process time step at each prediction [Figs. 1(a) and 1(b)], we randomly vary the sampling time step at every prediction. This design addresses two key considerations: First, it mimics the LAF principle of collecting samples from different denoising lead times to account for flow-dependent error growth in dynamical systems. Second, it increases ensemble diversity by exposing the model to different stages of the denoising process, thereby reducing the impact of stage-specific biases.

The proposed sampling strategy requires negligible computational overhead. Through a simple yet effective redesign of the sampling procedure, it can be applied to both DDIM and DDPM frameworks. For DDIM in this study, we use ten sequences ($\tau = 10 < T = 50$), chosen to reduce computational cost while retaining sufficient reverse steps for accurate prediction. By organizing ensemble predictions through samples drawn from distinct steps of the reverse process at each cycle, we effectively simulate the LAF approach within the diffusion model framework, ensuring that

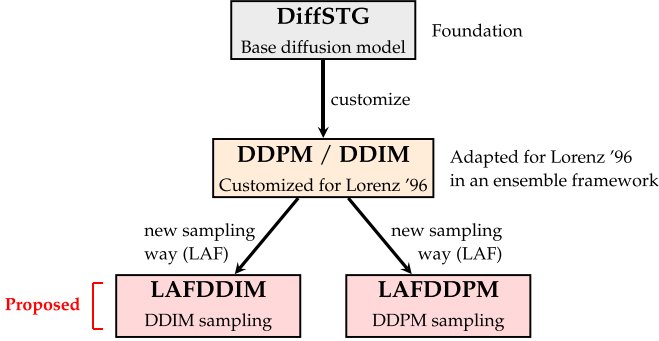


FIG. 2. ML model development hierarchy. Starting from the base DiffSTG model, we customize it for the Lorenz '96 system (DDPM and DDIM), then develop two variants with new sampling strategies: LAFDDIM (DDIM-based) and LAFDDPM (DDPM-based). The latter two constitute our main contributions.

the ensemble captures a reasonable distribution of prediction uncertainties.

In summary, our proposed models are developed through a hierarchical approach, as illustrated in Fig. 2. We begin with DiffSTG, a foundation diffusion model from the literature, and customize it for the Lorenz '96 system (Sec. II C). From this intermediate model, we develop two variants with sampling strategies: LAFDDIM, which employs DDIM-based sampling, and LAFDDPM, which uses a modified DDPM approach. These two models constitute the main contributions of this work.

C. Dynamic system

The two-level system of Lorenz '96 model [9] consists of a governing equation for slow ($\mathbf{x}$) and fast motions ($\mathbf{y}$), respectively.

$$\begin{aligned} \frac{dx_k}{dt} &= x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F - \frac{hc}{b} \sum_{j=1}^J y_{j,k}, \\ \frac{dy_{j,k}}{dt} &= -cb y_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - c y_{j,k} + \frac{hc}{b} x_k. \end{aligned} \quad (2)$$

Spatial positions are indexed by k and j , ranging from 1 to $K = 40$ and $J = 10$ for the slow and fast motions, respectively, in this study. The numerical parameters chosen for this study are listed in Table I. Among this set, F is the large-scale forcing term that modulates the characteristics of the dynamical system, and it is set to a value of 20.0 to develop highly nonlinear behavior in the slow-scale motions. For small values of the external forcing, the Lorenz '96 system exhibits diurnal or seasonal cycles with quasiperiodicity [27], whereas with $F = 20.0$ it demonstrates highly complex behavior (see Fig. 1 in Ref. [13]). The time series data for training is obtained from the numerical integration of Eq. (2) with the initial condition $x_1(t = 0) = 8.0$ and $x_k(t = 0) = 0.0$ for $k = 2, \dots, 40$. Equation (2) is temporally discretized by the fourth-order Runge-Kutta method [28] with a time-step size of $\Delta t = 0.001$ in model time units (MTUs) [9] (see Table I). The numerical solutions are provided as datasets for training and validation of a machine-learning model [11]. A time-series of predictions

TABLE I. List of parameters of the two- and three-level Lorenz '96 systems.

Notation	Definition	Value
K	Grid numbers for $\mathbf{x}$	40
J	Grid numbers for $\mathbf{y}$	10
I	Grid numbers for $\mathbf{z}$	10
h	Strength of the coupling between spatial scales	1.0
b	Relative magnitudes of the coupled variables $\mathbf{x}$ and $\mathbf{y}$	10.0
c	Relative evolution speed of the coupled variables $\mathbf{x}$ and $\mathbf{y}$	10.0
e	Relative evolution speed of the coupled variables $\mathbf{y}$ and $\mathbf{z}$	10.0
F	Large-scale forcing applied to the system	20.0
g_z	Damping parameter	1.0
$x_1(t = 0)$	Initial condition for the training simulation	8.0
$x_{20}(t = 0)$	Initial condition for the test simulation (TRUTH)	8.0
Δt	Time step size	0.001 MTUs

at each spatial and temporal grid is prepared for training [29], producing a dataset consisting of 1.6×10^6 time sequences, each containing 40 grid-point states. Only the large-scale states x_k is used for training and the machine-learning model predicts $x_k(t > 0)$ without explicitly solving the smaller scale states $y_{j,k}$.

To evaluate the model's performance, a test dataset is generated by numerically integrating Eq. (2) with a spatially shifted initial condition, serving as ground truth (TRUTH) for comparison. Specifically, the training dataset uses $x_1(t = 0) = 8.0$ with all other grid points set to zero, while the test dataset (TRUTH) uses $x_{20}(t = 0) = 8.0$ with all other grid points set to zero. This spatially shifted initial condition tests the model's ability to generalize beyond the training distribution.

The introduction of an additional scale of motion increases the system's complexity, and extending it to a three-level system can be formulated as [30]:

$$\begin{aligned} \frac{dx_k}{dt} &= x_{k-1}(x_{k+1} - x_{k-2}) - x_k + F - \frac{hc}{b} \sum_{j=1}^J y_{j,k}, \\ \frac{dy_{j,k}}{dt} &= -cb y_{j+1,k}(y_{j+2,k} - y_{j-1,k}) - c y_{j,k} + \frac{hc}{b} x_k - \frac{he}{d} \sum_{i=1}^I z_{i,j,k}, \\ \frac{dz_{i,j,k}}{dt} &= -ed z_{i-1,j,k}(z_{i+1,j,k} - z_{i-2,j,k}) - g_z e z_{i,j,k} + \frac{he}{d} y_{j,k}. \end{aligned} \quad (3)$$

The three-tier extension was originally developed to investigate the effects of reducing numerical precision to different degrees on different spatial scales, as part of research into more efficient computational methods for weather and climate modeling [30]. It has since been proposed as a useful testbed for novel methods, including machine learning approaches [31].

Building on this framework, we address a fundamental challenge in machine learning for chaotic systems: generalization across different levels of multiscale complexity. We train a model exclusively on time series of the slow $\mathbf{x}$ variables from the two-level system [Eq. (2)], learning their effective dynamics without explicit knowledge of the fast-scale forcing from $\mathbf{y}$. This trained model is then evaluated also on the three-level system [Eq. (3)], where $\mathbf{x}$ experiences forcing from a two-tier cascade ($\mathbf{y}$ and $\mathbf{z}$) rather than a single fast scale.

Although both systems produce chaotic $\mathbf{x}$ dynamics with similar amplitudes and timescales, the spectral and statistical characteristics of the effective forcing, which represents the net influence of smaller-scale variables on the largest-scale dynamics, differ due to the intermediate-scale interactions. This setup tests whether the model captures robust dynamical features that generalize across different forcing structures, or remains limited to the specific forcing statistics of its training environment.

D. Ensemble square root filter (EnSRF)

The use of machine learning models as forecast models for data assimilation has been explored [12]. Building on this, we aim to extend the investigation to assess the predictive skill and performance of a diffusion model within an ensemble data assimilation framework. In this study, serial ensemble square root filter (EnSRF) [32] is used for the data assimilation process. Each forecast ensemble $\mathbf{x}_n^f$ is updated using the model dynamics:

$$\mathbf{x}_n^f = \mathcal{M}(\mathbf{x}_n^a), \quad (4)$$

where $\mathbf{x}_n^a$ is the n th assimilated member state from the previous cycle and $\mathcal{M}$ represents a dynamical model, which may either be numerical or based on machine learning techniques. In this study, $\mathcal{M}$ is realized either by a numerical integration of the Lorenz '96 system (PHYS) or by the diffusion-based forecast model described in Sec. II B and Fig. 2 (DDPM, DDIM, LAFDDPM, and LAFDDIM).

The ensemble forecast mean $\bar{\mathbf{x}}^f$ and perturbations $\mathbf{X}^f$ are calculated as:

$$\bar{\mathbf{x}}^f = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n^f, \quad (5)$$

and $\mathbf{X}^f = [\mathbf{x}_1^f - \bar{\mathbf{x}}^f, \dots, \mathbf{x}_N^f - \bar{\mathbf{x}}^f]$ when N is the ensemble size. The analysis step updates the ensemble using the innovation, i.e., the difference between an observation and the forecast projected into the observation space, $\mathbf{y}^o - \mathbf{H}\mathbf{x}^f$, where $\mathbf{y}^o$ represents the observation vector and $\mathbf{H}$ the observation operator, which maps the model state ($\mathbf{x}^f$) space to the observation space. The EnSRF is a deterministic square root filter: unlike the stochastic ensemble Kalman filter, it updates the ensemble without adding random perturbations to the observations. In this study, a single observation is available at every grid point, and the serial EnSRF processes these observations one at a time in the order of the grid index, while the framework itself admits an arbitrary number and configuration of observations. The k th step of the serial update therefore corresponds to processing the observation at the k th grid point, and we use the subscript k consistently for the grid point, its observation, and the associated serial update step.

The observation error variance is $R_k = 0.1$ in this study. For the k th observation, the Kalman gain $\mathbf{K}_k$ is given by:

$$\mathbf{K}_k = \mathbf{P}_{k-1}^a \mathbf{H}_k^T (\mathbf{H}_k \mathbf{P}_{k-1}^a \mathbf{H}_k^T + R_k)^{-1}, \quad (6)$$

where the analysis error covariance is $\mathbf{P}_{k-1}^a = \frac{1}{N-1} \mathbf{X}_{k-1}^a (\mathbf{X}_{k-1}^a)^T$, evaluated from the perturbations available after the previous observation has been processed. The ensemble mean is then updated with the scalar innovation $y_k^o - \mathbf{H}_k \bar{\mathbf{x}}_{k-1}^a$, given by:

$$\bar{\mathbf{x}}_k^a = \bar{\mathbf{x}}_{k-1}^a + \mathbf{K}_k (y_k^o - \mathbf{H}_k \bar{\mathbf{x}}_{k-1}^a). \quad (7)$$

The serial EnSRF updates the ensemble perturbations deterministically, without perturbing the k th observation y_k^o , as given by:

$$\mathbf{X}_k^a = \mathbf{X}_{k-1}^a (\mathbf{I} - \alpha_k \mathbf{K}_k \mathbf{H}_k), \quad (8)$$

where α_k is derived from the square root update given by:

$$\alpha_k = \left(1 + \sqrt{R_k / (\mathbf{H}_k \mathbf{P}_{k-1}^a \mathbf{H}_k^T + R_k)} \right)^{-1}. \quad (9)$$

The recursion is initialized with $\bar{\mathbf{x}}_0^a = \bar{\mathbf{x}}^f$ and $\mathbf{X}_0^a = \mathbf{X}^f$, where $\bar{\mathbf{x}}^f$ and $\mathbf{X}^f$ are the forecast ensemble mean and perturbations. After all K observations have been processed, the final analysis ensemble mean $\bar{\mathbf{x}}^a$ and perturbations $\mathbf{X}^a$ are obtained. We conduct observing system simulation experiments to evaluate the performance of machine learning models in data assimilation. Figure 3 illustrates our overall experimental framework. The TRUTH, obtained from a single realization of the numerical integration of the full-level Lorenz '96 dynamics is used to generate synthetic observations of slow variables at every grid point. These observations are assimilated into ensemble forecasts using the ensemble square root filter (EnSRF).

Ensemble data assimilation utilizes the Bayesian inference framework, where observations take parts as likelihood and ensemble predictions as the prior. At each temporal interval, the analyzed states of the ensemble members are updated based on their relative weights, which are determined by the uncertainties in both the forecasts and observations. These updated analyzed states are then used as initial conditions for the ensemble members to generate the next set of predictions. This iterative process, known as the “forecast-analysis cycle,” progressively reduces the estimated uncertainties of the ensemble predictions. When the ensemble size is small, however, sampling errors can lead to underestimation of uncertainty and ultimately to filter divergence, in which the ensemble loses its ability to track the true state. In this study, we do not use any inflation or localization methods to deal with such sampling errors, since we focus on the better representation of ensemble variability in the context of diffusion-based models. The analysis is produced at every 50th time step, which corresponds to a six-hour interval.

III. RESULTS

We first examine the case with ensemble size $N = 32$ ($\approx K = 40$). At this configuration, all tested models produce accurate analyses through ensemble data assimilation. We then reduce the ensemble size to $N = 20$ ($< K = 40$), where the numerical model PHYS may experience divergence [8]

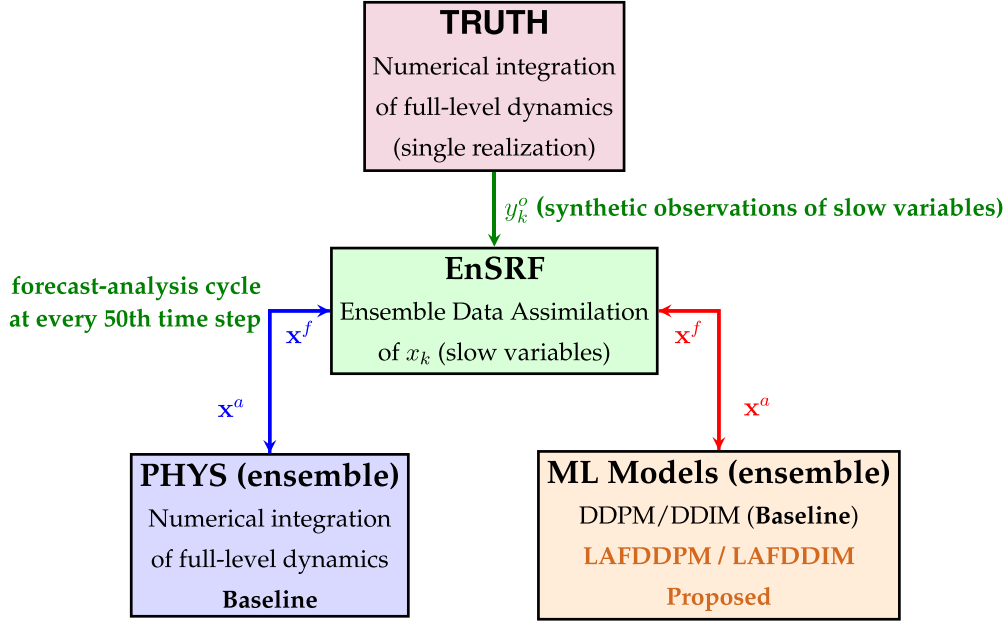


FIG. 3. Data assimilation framework for observing system simulation experiments. We compare against TRUTH is a single realization from numerical integration of the full-level Lorenz '96 dynamics, generating observations (y_k^o). EnSRF assimilates observations to update ensemble forecasts. The baseline approach (PHYS) uses numerical integration of the full dynamics, while the proposed ML models (LAFDDIM and LAFDDPM) directly learn to forecast slow variables from data (see Fig. 2). Both approaches operate through forecast-analysis cycles, where $\mathbf{x}^f$ and $\mathbf{x}^a$ denote forecast and analysis states, respectively.

(see their Fig. 2). Indeed, both PHYS and DDIM fail to reproduce the dynamics properly, while DDPM maintains stable forecast-analysis cycling (see Fig. 2 for model notations). When the ensemble size is further reduced to $N = 16$, even DDPM sampling struggles to control error growth, as shown in Fig. 4. For this small ensemble size, we also test the proposed models LAFDDPM and LAFDDIM. Figure 4 compares the time series of x_{20} from all tested models against the true state (TRUTH).

At the start, solutions from all tested models deviate greatly from TRUTH due to perturbed initial conditions [Fig. 4(a)]. However, LAFDDPM and LAFDDIM quickly capture the dynamics as forecast-analysis cycling continues, while PHYS and DDPM fail to converge. This performance difference is quantified by continuous ranked probability score (CRPS) [33] values. While root mean-squared difference provides pointwise error estimation, CRPS offers additional insight into how well models represent prediction uncertainties. Although CRPS values from all models start high due to deliberately perturbed initial conditions, LAFDDPM and LAFDDIM show rapid error reduction as data assimilation proceeds. In contrast, CRPS of PHYS and DDPM fluctuates significantly and remains high throughout [Fig. 4(c)].

When the ensemble size is further reduced to $N = 8$ ($\ll K = 40$), LAFDDPM and LAFDDIM continue to perform well with moderate degradation [Figs. 4(b) and 4(d)]. Both models keep their solutions in phase with TRUTH and produce stable forecast-analysis cycles. Although CRPS values increase compared to the $N = 16$ case, the error growth remains controlled and substantially lower than that of PHYS and DDPM, which diverge from TRUTH at this small ensemble size. This indicates that the proposed models (LAFDDPM

and LAFDDIM) maintain reasonable performance even in severely undersampled regimes.

This result indicates that the ensemble spread from the proposed models better represents prediction uncertainties compared to the numerical approach (PHYS) and the baseline DDPM. This is consistent with the ensemble samples generated during the reverse process in Figs. 1(c) and 1(d), which show enhanced ensemble spread representing better estimated uncertainties. The proposed sampling strategy prevents excessive clustering [see Figs. 1(a) and 1(b)] and enhances variability while maintaining a focused distribution of samples [see Figs. 1(c) and 1(d)]. This improves the ability to capture the true state even with small ensembles.

The coherence in the dynamical structure produced by the proposed method is further investigated through the evolution of ensemble mean $\bar{x}$ in space and time, illustrated in Hovmöller diagrams, which help identify wave propagation patterns. Using these diagrams, we assess the overall similarity between the true state and model predictions, providing clear visualization of their temporal and spatial agreement. Figure 5 shows that the ensemble means $\bar{x}$ in LAFDDPM and LAFDDIM reasonably reproduce the wave propagation pattern of TRUTH, capturing key structural features such as the location and shape of local extrema, in contrast to the numerical simulation PHYS and the conventional machine-learning approach (DDPM). However, the patterns generated by LAFDDIM tend to be slightly attenuated relative to TRUTH, suggesting some limitations in reproducing the full amplitude range. To quantify this behavior, we examine the marginal probability density functions (PDF) in Fig. 6.

Following the autocorrelation analysis, we subsample the time series every 274 time steps (the decorrelation timescale)

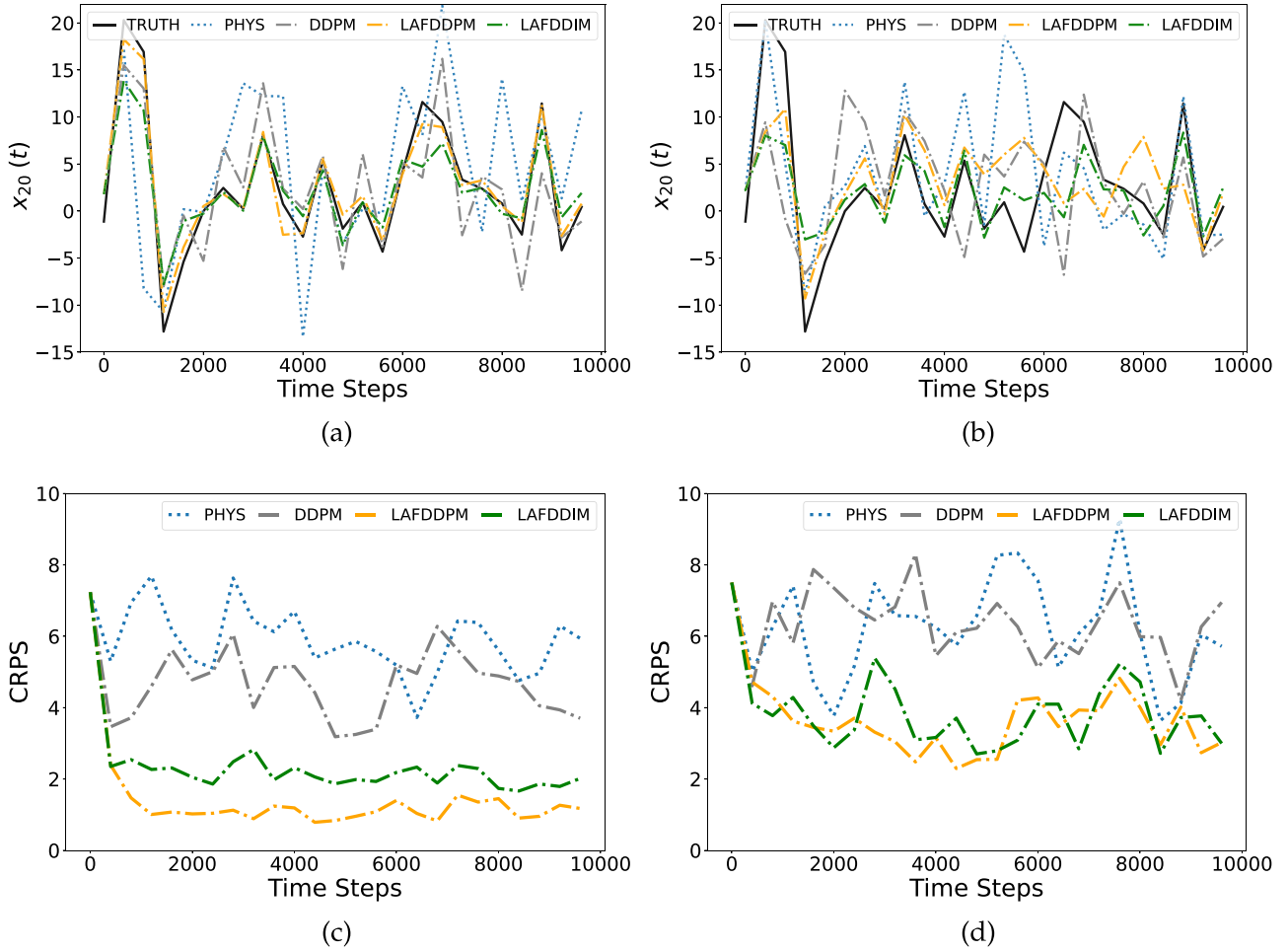


FIG. 4. Time series comparison for ensemble size $N = 16$ (left column) and $N = 8$ (right column). (a), (b) Time series of the true state x_{20} (TRUTH, black solid) and ensemble mean from PHYS (blue dotted), DDPM (gray dash-dotted), LAFDDPM (orange dash-dotted), and LAFDDIM (green dash-dotted). (c), (d) Time evolution of continuous ranked probability score (CRPS).

to ensure statistical independence, yielding approximately 1820 effective samples from the extended cycling run of 500 000 time steps for each model. While the proposed models LAFDDPM and LAFDDIM maintain reasonable ensemble spread and represent the wave propagation of TRUTH, they exhibit a narrowed PDF, with extreme values appearing less frequently than in the truth. The PDF analysis at $N = 16$ and $N = 8$ (Fig. 6) quantifies this behavior and reveals whether it is mitigated with larger ensembles.

Figure 6 compares the PDFs of x_{20} from TRUTH and all tested models. At $N = 16$ [Fig. 6(a)], LAFDDPM closely reproduces the overall shape of the TRUTH distribution, with the peak location well matched but showing slightly elevated probability density at the center and reduced density in both tails, indicating mild tail attenuation. LAFDDIM exhibits more pronounced compression with a noticeably higher and narrower peak and substantially reduced tail probabilities. In contrast, PHYS captures the tail regions reasonably well, particularly on the right side, but with its peak shifted to lower values. DDPM shows a broader, flatter distribution with a leftward-shifted peak and overestimated probabilities in both tails.

At the more constrained $N = 8$ [Fig. 6(b)], PDF narrowing intensifies across all models. LAFDDIM shows the

most extreme case, with a tall, high peak and severely underestimated tails. LAFDDPM exhibits a moderately compressed distribution. PHYS and DDPM maintain relatively good tail agreement despite leftward-shifted peaks. Notably, LAFDDPM demonstrates substantial improvement from $N = 8$ to $N = 16$, suggesting that PDF narrowing can be significantly mitigated with moderately larger ensemble sizes.

Notably, LAFDDIM's particularly strong PDF narrowing reflects its deterministic sampling framework. Unlike LAFDDPM's stochastic diffusion process, which introduces randomness at each reverse step, LAFDDIM generates deterministic samples that systematically underrepresent fluctuations. This results in an attenuated dynamic range that persists even at larger ensemble sizes ($N = 16$), where LAFDDPM substantially recovers distributional accuracy. The deterministic nature of LAFDDIM thus presents a fundamental tradeoff: while maintaining temporal coherence and computational efficiency, it inherently limits the representation of extreme values and the width of the long-term distribution.

The PDF analysis may also reflect an intrinsic characteristic of ensemble data assimilation with small ensemble sizes. When the ensemble is severely undersampled relative to the

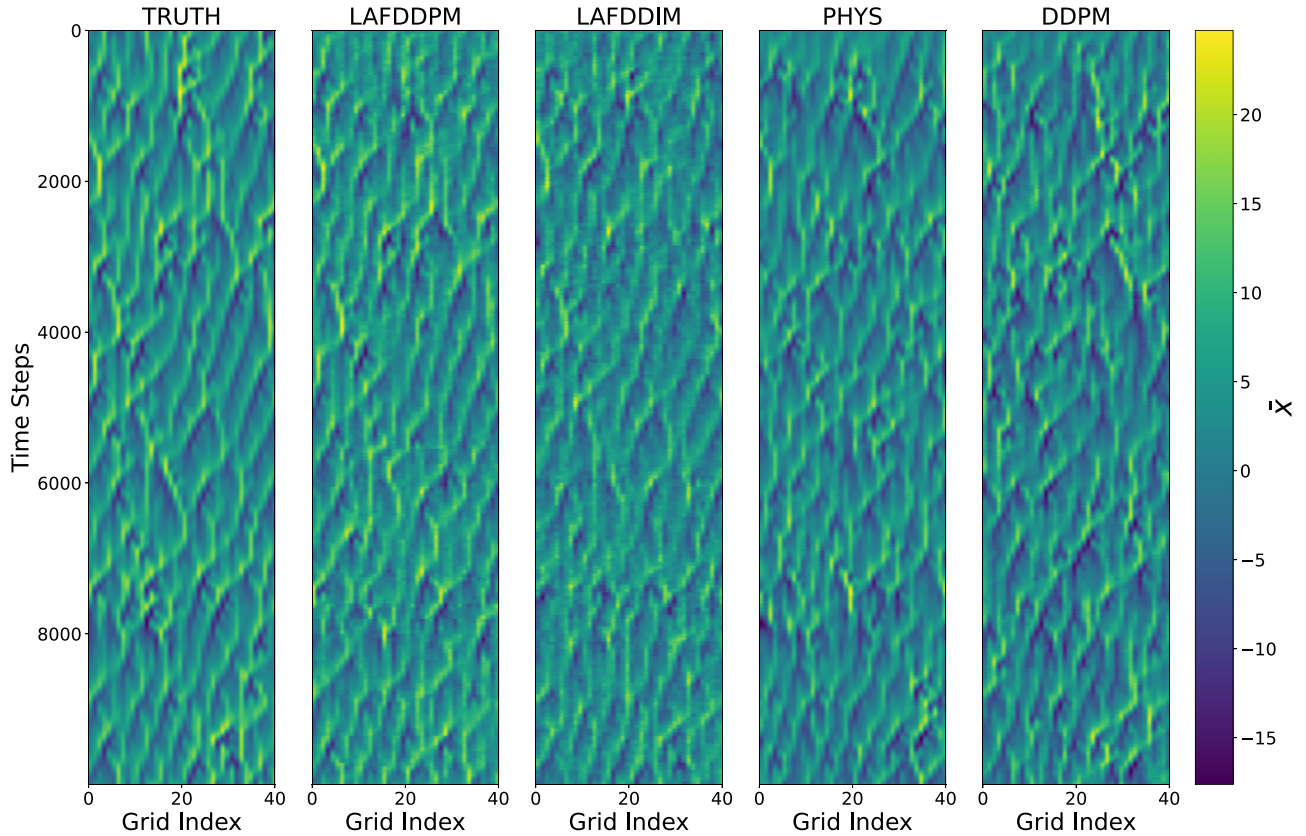


FIG. 5. Hovmöller diagrams from the TRUTH and ensemble ($N = 8$) mean $\bar{x}$ from LAFDDPM, LAFDDIM, PHYS, and DDPM (left to right).

system's degrees of freedom, forecast uncertainty can be systematically underestimated and the Kalman gain subsequently becomes negligible. In PHYS and DDPM, this underestimation is severe enough that the analysis update fails to correct the forecast, leaving the ensemble unable to track the true state, while the long-term PDF retains tails comparable to those of TRUTH. In contrast, in LAFDDPM and LAFDDIM, the Kalman gain, although smaller than for a larger ensemble, remains non-negligible. The correction through the analysis update is smaller than it would be for a larger ensemble in the early cycles. Consequently, the innovation $\mathbf{y}^o - \mathbf{H}\mathbf{x}^f$ remains non-negligible across cycles, and the analysis update applies corrections persistently. These sustained corrections throughout the forecast-analysis cycle reduce the frequency of extreme values, concentrating probability near the center of the distribution. In other words, the analysis update is effective enough to keep the proposed models tracking the true state, yet unable to fully restore its dynamic range. This limitation arises because the rank-deficient background error covariance, inherent to severely undersampled ensembles, constrains both the magnitude and the structure of each correction. The result is accurate tracking accompanied by a probability density function that is paradoxically narrower than for the poorly performing models.

We note, however, that this narrowing of the probability density function does not undermine the predictive skill of the proposed models. LAFDDPM and LAFDDIM closely track the true state, as shown by the time-series agreement, the

CRPS, and the Hovmöller diagrams in Figs. 4 and 5. This skill is sustained over the long term, as will be shown in Figs. 7 and 8.

Dynamical performance metrics (Fig. 7) further characterize model differences. Point-wise error statistics [Fig. 7(a)] align with the distributional patterns in Fig. 6: PHYS (Bias = +0.07, RMSE = 8.37) and DDPM (Bias = -0.10, RMSE = 7.80) exhibit substantially larger RMSE than the proposed models, while LAFDDPM (Bias = +0.10, RMSE = 4.47) and LAFDDIM (Bias = -0.29, RMSE = 4.28) show substantially lower errors. Lagged correlation analysis [Fig. 7(b)] reveals the same contrast in a different form: LAFDDPM and LAFDDIM maintain high temporal correlation with TRUTH (≈ 0.8 at zero lag), whereas PHYS and DDPM exhibit substantially lower correlation (≈ 0.2 and ≈ 0.35 , respectively) despite their better tail agreement. This illustrates a fundamental distinction: PHYS closely reproduces the attractor's statistical structure but shows limited skill in maintaining phase synchronization, whereas LAFDDPM and LAFDDIM achieve strong temporal coherence while producing narrowed PDFs.

To evaluate long-term stability, we examine CRPS evolution over the extended integration used for statistical analysis (500 000 time steps) at the challenging $N = 8$ ensemble size (Fig. 8). For the two-level system [Fig. 8(a)], both LAFDDPM and LAFDDIM maintain CRPS values consistently below PHYS, with no evidence of drift or degradation throughout the integration. This sustained performance demonstrates stable

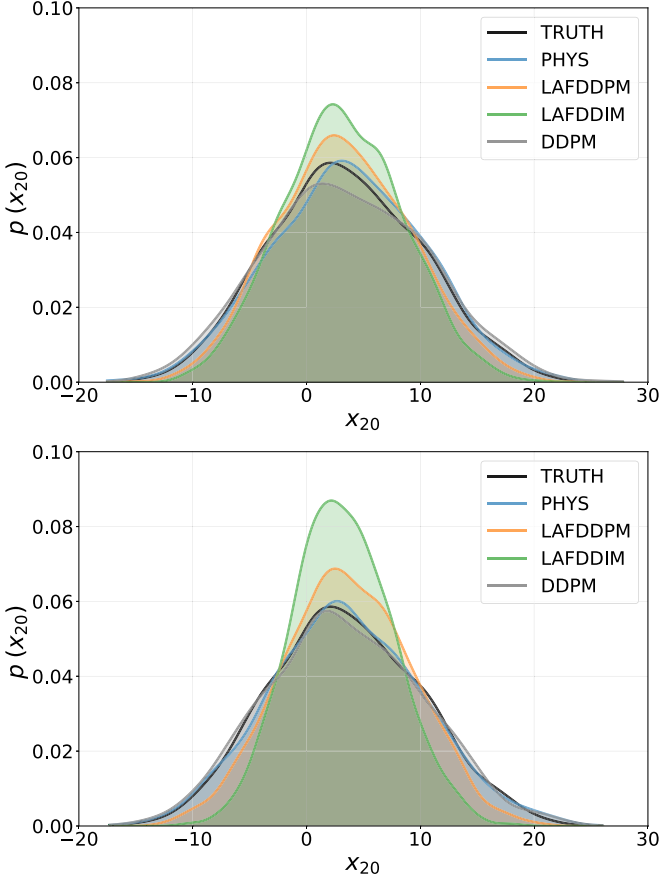


FIG. 6. Marginal probability density functions of x_{20} for (a) $N = 16$ and (b) $N = 8$. Colors indicate TRUTH (black), PHYS (blue), DDPM (gray), LAFDDPM (orange), and LAFDDIM (green). PDFs are estimated from 500 000 time steps (approximately 1820 independent samples based on decorrelation time of 0.274 time units).

forecast-analysis cycling without requiring covariance inflation or localization adjustments.

We further assess generalization capability by testing on the three-level system [Eq. (3)], where $\mathbf{x}$ experiences forcing

from a two-tier subgrid cascade ($\mathbf{y}$ and $\mathbf{z}$) rather than the single fast scale of the two-level system. Figure 8(b) demonstrates that the proposed models successfully handle this increased multiscale complexity, maintaining stable CRPS values comparable to the two-level case despite never having been directly trained on the three-level configuration. This successful generalization indicates that the models capture robust dynamical features that extend beyond the specific multi-scale forcing structure of their training environment.

IV. DISCUSSIONS

We have investigated diffusion model-based machine learning algorithms for ensemble data assimilation in highly nonlinear systems, with particular focus on performance under varying ensemble sizes. Accurately predicting states that align with observations presents challenges for data-driven models, especially in systems where extreme events occur with low probability due to nonlinear dynamics. These challenges intensify with small ensembles, where traditional approaches often exhibit degraded performance.

Our evaluation reveals that the proposed machine-learning models (LAFDDPM, LAFDDIM) achieve robust temporal correlation and phase tracking even at small ensemble sizes ($N = 8$, $N = 16$), significantly outperforming both the physical model (PHYS) and the conventional machine-learning approach (DDPM) in dynamical skill. This advantage stems from the time-lagged sampling strategy, which preserves ensemble diversity by varying the temporal positions from which states are sampled during reverse diffusion, thereby preventing filter divergence. The strategy sustains ensemble spread without requiring covariance inflation or localization, enabling stable forecast-analysis cycles even with severe undersampling ($N = 8$ for a state of $K = 40$ variables).

From a methodological perspective, these findings reveal several important insights for ensemble-based prediction in strongly nonlinear, multiscale systems. When computational constraints limit ensemble size far below the system's degrees of freedom, a common situation in complex geophysical, plasma, or turbulent systems, ensemble diversity preservation

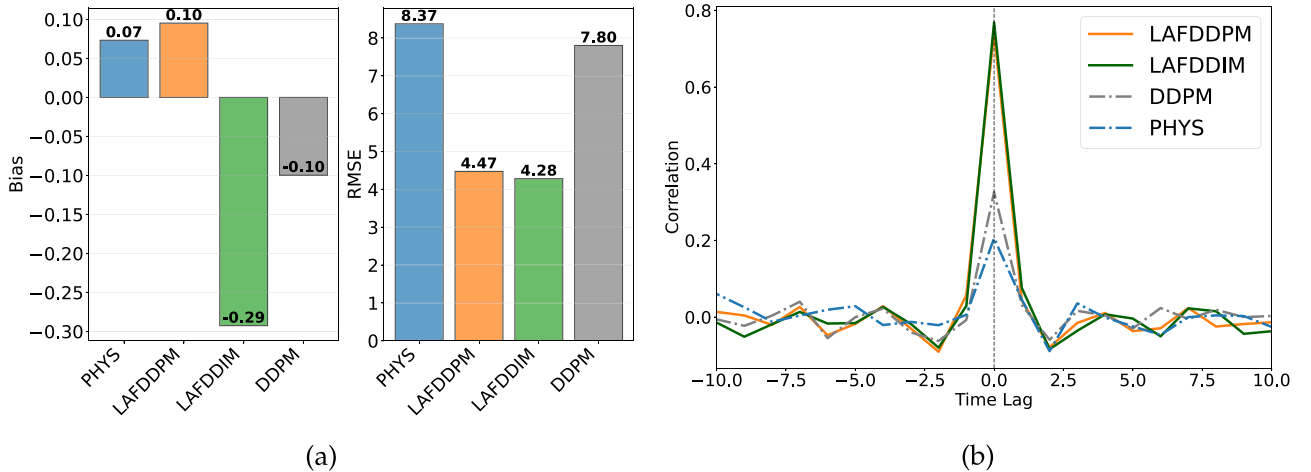


FIG. 7. Statistical analysis from 500 000 time step integration for $N = 8$. (a) Pointwise error metrics (bias and RMSE) computed from pooled ensemble members. (b) Lagged correlation computed from ensemble-mean time series.

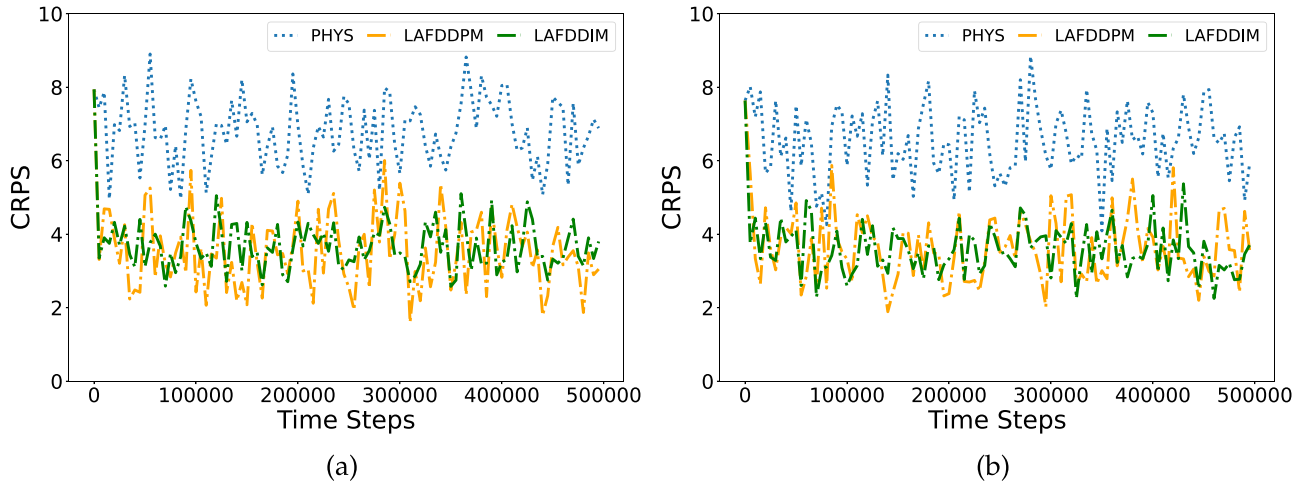


FIG. 8. Long-term CRPS evolution demonstrating sustained performance and generalization capability. (a) Two-level Lorenz '96 system at $N = 8$ over 500 000 time steps. (b) Three-level system with two-tier subgrid cascade, testing generalization to increased multiscale complexity beyond the training configuration. Line styles: PHYS (blue dotted), LAFDDPM (orange dashed), LAFDDIM (green dashed).

may be as critical as the accuracy of the underlying dynamical model. Our time-lagged sampling strategy successfully preserves this diversity by varying temporal positions during reverse diffusion, addressing the sensitivity of diffusion model predictions to the reverse sampling process. This approach enhances the model's ability to track the true state, even when ensemble size severely limits sampling capacity.

However, our analysis also reveals inherent tradeoffs in undersampled regimes. Comprehensive statistical assessment across extended simulations (500 000 time steps, effective sample size ≈ 1820) shows that while the proposed models maintain superior phase tracking, they exhibit attenuated PDF tails, with extreme values appearing less frequently than in the truth. This narrowing arises because the sustained analysis update in the proposed models, while effective enough to keep the ensemble tracking the true state, cannot fully restore its dynamic range under severe undersampling, where the rank-deficient background error covariance constrains both the magnitude and the structure of each correction. While LAFDDPM largely recovers distributional accuracy at $N = 16$, LAFDDIM shows persistent narrowing even with increased ensembles. The deterministic nature of LAFDDIM thus presents a fundamental tradeoff: while maintaining temporal coherence and computational efficiency, it inherently

limits the representation of extreme values and the width of the long-term distribution.

Future work will address two main directions. First, given the computational efficiency of deterministic sampling variants, exploring ways to improve extreme value representation while maintaining their speed advantages remains valuable, despite the fundamental constraints in severely undersampled regimes. Second, extending these methods to higher-dimensional systems and testing across different types of nonlinear dynamics will help establish their broader applicability.

ACKNOWLEDGMENTS

This research was supported by Characterization platform for advanced materials (GP2025-0005-15) and Quantum Scopes for Quantum Materials Property Evaluation (KRISS-2026-KP2026-0004) funded by Korea Research Institute of Standards and Science.

DATA AVAILABILITY

The data that support the findings of this article are openly available [34].

APPENDIX A: NOTATION SUMMARY

This appendix summarizes the notation used throughout this study in Table II.

TABLE II. Summary of notation used in this study.

Symbol	Description
Diffusion model	
$\mathbf{x}_{t_i}$	Sample state at reverse step t_i
t_i	Reverse process step index
β_{t_i}	Noise variance at step t_i
α_{t_i}	Scaling parameter at step t_i
$\epsilon_{\theta}^{(t_i)}$	Learned denoising function
σ_{t_i}	Noise standard deviation at step t_i (zero for DDIM)
T	Total number of diffusion steps ($T = 50$)
τ	Subset sequence length ($\tau = 10 < T = 50$)
Dynamical system (Lorenz '96)	
$\mathbf{x}$	Slow-scale state vector ($K = 40$)
x_k	Slow-scale variable at grid point k
$\mathbf{y}$	Fast-scale state vector ($J = 10$)
$\mathbf{z}$	Sub-fast-scale state vector ($I = 10$)
t	Physical time
Data assimilation	
N	Ensemble size
$\mathcal{M}$	Dynamical model (numerical or machine learning)
$\mathbf{x}^f$	Forecast state (generic notation)
$\mathbf{x}^a$	Analysis state (generic notation)
$\mathbf{x}_n^f$	Forecast state of the n -th ensemble member
$\mathbf{x}_n^a$	Analysis state of the n -th ensemble member
$\bar{\mathbf{x}}$	Slow variable (true state or mean over ensemble members)
$\bar{\mathbf{x}}^f$	Forecast ensemble mean
$\mathbf{X}^f$	Forecast ensemble perturbations
$\mathbf{y}^o$	Observation vector
y_k^o	Individual observation at grid point k
$\mathbf{H}_k$	Observation operator for the k th observation
R_k	Observation error variance for the k th observation
$\bar{\mathbf{x}}_k^a$	Analysis ensemble mean after processing the k th observation
$\mathbf{X}_k^a$	Analysis ensemble perturbations after processing the k th observation
$\bar{\mathbf{x}}^a$	Final analysis ensemble mean after all observations are processed
$\mathbf{X}^a$	Final analysis ensemble perturbations after all observations are processed
$\mathbf{P}_{k-1}^a$	Analysis error covariance before the k th update
$\mathbf{K}_k$	Kalman gain for the k th observation
α_k	Square root update factor for the k th observation

- [1] R. Lam, A. Sanchez-Gonzalez, M. Willson, P. Wirnsberger, M. Fortunato, F. Alet, S. Ravuri, T. Ewalds, Z. Eaton-Rosen, W. Hu, *et al.*, Learning skillful medium-range global weather forecasting, *Science* **382**, 1416 (2023).
- [2] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, *et al.*, Probabilistic weather forecasting with machine learning, *Nature (London)* **637**, 84 (2025).
- [3] A. J. Charlton-Perez, H. F. Dacre, S. Driscoll, S. L. Gray, B. Harvey, N. J. Harvey, K. M. Hunt, R. W. Lee, R. Swaminathan, R. Vandaale, *et al.*, Do AI models produce better weather forecasts than physics-based models? A quantitative evaluation case study of Storm Ciarán, *npj Clim. Atmos. Sci.* **7**, 93 (2024).
- [4] O. C. Pasche, J. Wider, Z. Zhang, J. Zscheischler, and S. Engelke, Validating deep learning weather forecast models on recent high-impact extreme events, *Artificial Intel. Earth Syst.* **4**, e240033 (2025).
- [5] S. Materia, L. P. García, C. van Straaten, S. O. A. Mamalakis, L. Cavicchia, D. Coumou, P. de Luca, M. Kretschmer, and M. Donat, Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives, *WIREs Climate Change* **15**, e914 (2024).
- [6] G. Camps-Valls, M.-Á. Fernández-Torres, K.-H. Cohrs, A. Höhl, A. Castelletti, A. Pacal, C. Robin, F. Martinuzzi, I. Papoutsis, I. Prapas, *et al.*, Artificial intelligence for modeling

- and understanding extreme weather and climate events, *Nature Commun.* **16**, 1919 (2025).
- [7] D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben, *et al.*, Neural general circulation models for weather and climate, *Nature (London)* **632**, 1060 (2024).
- [8] G.-H. Crystalng, D. McLaughlin, D. Entekhabi, and A. Ahanin, The role of model dynamics in ensemble Kalman filter performance for chaotic systems, *Tellus A: Dyn. Meteorol. Oceanogr.* **63**, 958 (2011).
- [9] E. N. Lorenz, in *Proceedings Seminar on Predictability* (ECMWF, 1996), Vol. 1.
- [10] The Lorenz-96 and its GCM analog, 2026, https://m2lines.github.io/L96_demo/notebooks/gcm-analogue.html.
- [11] P. D. Dueben and P. Bauer, Challenges and design choices for global weather and climate models based on machine learning, *Geosci. Model Dev.* **11**, 3999 (2018).
- [12] S. Hatfield, Evaluating neural network-based surrogates for the forecast model in an ensemble Kalman filter, https://github.com/samhatfield/ml_da, 2019, repository no longer accessible
- [13] S. Shin, The predictive skill of neural network models for the large-scale dynamics of the multi-level Lorenz '96 systems, ESS Open Archive (2021).
- [14] E. Calvellido, S. Reich, and A. M. Stuart, Ensemble Kalman methods: A mean-field perspective, *Acta Numer.* **34**, 123 (2025).
- [15] L. Li, R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson, Generative emulation of weather forecast ensembles with diffusion models, *Sci. Adv.* **10**, eadk4489 (2024).
- [16] S. Shin, I.-H. Lee, B. C. Park, and J.-H. Song, Applications of deep learning-based denoising methodologies for scanning electron microscope images, *Meas. Sci. Technol.* **36**, 015406 (2025).
- [17] C. Meijer and L. Y. Chen, The rise of diffusion models in time-series forecasting, [arXiv:2401.03006](https://arxiv.org/abs/2401.03006).
- [18] H. Wen, Y. Lin, Y. Xia, H. Wan, Q. Wen, R. Zimmermann, and Y. Liang, DiffSTG Probabilistic spatio-temporal graph forecasting with denoising diffusion models, in *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems* (ACM Press, New York, NY, 2023), pp. 1–12.
- [19] M. Li, T. Qu, R. Yao, W. Sun, and M.-F. Moens, Alleviating exposure bias in diffusion models through sampling with shifted time steps, in *International Conference on Learning Representations* (2024).
- [20] R. N. Hoffman and E. Kalnay, Lagged average forecasting, an alternative to Monte Carlo forecasting, *Tellus A: Dyn. Meteorol. Oceanogr.* **35**, 100 (1983).
- [21] C. Lu, H. Yuan, B. E. Schwartz, and S. G. Benjamin, Short-range numerical weather prediction using time-lagged ensembles, *Weather and forecasting* **22**, 580 (2007).
- [22] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning* (Springer, Berlin, 2006), Vol. 4.
- [23] J. Song, C. Meng, and S. Ermon, Denoising diffusion implicit models, in *International Conference on Learning Representations* (2021).
- [24] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *Adv. Neural Info. Proc. Syst.* **33**, 6840 (2020).
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, *Adv. Neural Info. Proc. Syst.* **30**, 5998 (2017).
- [26] A. Sabour, S. Fidler, and K. Kreis, Align your steps: Optimizing sampling schedules in diffusion models, in *Proceedings of the 41st International Conference on Machine Learning* (2024).
- [27] A. E. Sterk and D. L. van Kekem, Predictability of extreme waves in the Lorenz-96 model near intermittency and quasi-periodicity, *Complexity* **2017**, 1 (2017).
- [28] E. Ott, B. R. Hunt, I. Szunyogh, A. V. Zimin, E. J. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J. A. Yorke, A local ensemble Kalman filter for atmospheric data assimilation, *Tellus A: Dyn. Meteorol. Oceanogr.* **56**, 415 (2004).
- [29] S. Prabhakaran, ARIMA Model Complete Guide to Time Series Forecasting in Python, Machine Learning Plus (2021), <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>.
- [30] T. Thornes, P. Düben, and T. Palmer, On the use of scale-dependent precision in Earth system modelling, *Q. J. R. Meteorol. Soc.* **143**, 897 (2017).
- [31] A. Chattopadhyay, P. Hassanzadeh, and D. Subramanian, Data-driven predictions of a multiscale Lorenz 96 chaotic system using machine-learning methods: reservoir computing, artificial neural network, and long short-term memory network, *Nonlin. Proc. Geophys.* **27**, 373 (2020).
- [32] J. S. Whitaker and T. M. Hamill, Ensemble data assimilation without perturbed observations, *Mon. Weather Rev.* **130**, 1913 (2002).
- [33] A. Jordan, F. Krüger, and S. Lerch, Evaluating probabilistic forecasts with scoringRules, *J. Stat. Software* **90**, 1 (2019).
- [34] <https://github.com/seoleuns/MLforLorenz96>.