Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

DRL-EnVar: an adaptive hybrid ensemble–variational data assimilation method based on deep reinforcement learning*#

Lilan HUANG†§[1,2], Hongze LENG†‡§[2], Junqiang SONG†‡[2],
Dongzi WANG[1], Wuxin WANG[1], Ruisheng HU[2], Hang CAO[2]

[1]*College of Computer Science and Technology, National University of Defense Technology, Changsha 410073, China*

[2]*College of Meteorology and Oceanography, National University of Defense Technology, Changsha 410073, China*

†E-mail: huanglilan18@nudt.edu.cn; hzleng@nudt.edu.cn; junqiang@nudt.edu.cn

Received Dec. 14, 2024; Revision accepted Aug. 19, 2025; Crosschecked Sept. 3, 2025; Published online Nov. 26, 2025

Abstract: Accurate estimation of the background error covariance matrix denoted as $\boldsymbol{B}$ remains a critical challenge in numerical weather prediction (NWP), directly influencing data assimilation (DA) performance and forecast accuracy. Although hybrid ensemble–variational (EnVar) methods combine static and flow-dependent matrices to improve assimilation, their effectiveness is constrained by empirically fixed weights. To address this limitation, we propose DRL-EnVar, an adaptive hybrid EnVar DA method enhanced with deep reinforcement learning. DRL-EnVar integrates deep learning (DL) components, including a novel cyclic convolution module to extract abstract features from data, and employs reinforcement learning (RL) to dynamically optimize hybrid weighting strategies. The system adaptively combines multiple ensemble-based flow-dependent matrices with one or more static matrices to construct a time-varying hybrid matrix $\boldsymbol{B}$ that better reflects real-time background errors. Experimental results demonstrate that DRL-EnVar performs better than the traditional ensemble Kalman filter (EnKF) and hybrid covariance DA (HCDA) methods, especially under sparse observations or transitional changes in state variables. It achieves competitive or superior assimilation accuracy with lower computational cost, and can be flexibly integrated into both three-dimensional variational assimilation (3DVar) and four-dimensional variational assimilation (4DVar) frameworks. Overall, DRL-EnVar offers a novel and efficient approach to adaptive DA, particularly valuable for improving forecast skill during transitional weather regimes.

Key words: Adaptive data assimilation; Hybrid ensemble–variational method; Background error covariance; Deep reinforcement learning

https://doi.org/10.1631/FITEE.2401063 **CLC number:** TP391

1 Introduction

Data assimilation (DA) is vital in numerical weather prediction (NWP), climate monitoring, and environmental prediction (Sanz-Alonso et al., 2023). It improves the initial state by combining observations with background information from numerical models, improving the accuracy and reliability of predictions. The background error covariance matrix denoted as $\boldsymbol{B}$ plays a central role in DA, quantifying the uncertainty in the background state, balancing observations and model priors, and directly influencing the performance of DA (Kalman, 1960). In

‡ Corresponding authors
§ These two authors contributed equally to this work
* Project supported by the National Key R&D Program of China (No. 2022YFB3207304), the National Natural Science Foundation of China (No. 42205161), and the Natural Science Foundation of Hunan Province, China (No. 2023JJ30630)
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2401063) contains supplementary materials, which are available to authorized users
ORCID: Lilan HUANG, https://orcid.org/0000-0002-6101-0574; Hongze LENG, https://orcid.org/0009-0007-9992-3823; Junqiang SONG, https://orcid.org/0009-0003-2686-566X

scenarios with sparse observations and transitional weather regimes, accurately estimating $\boldsymbol{B}$ to ensure timely responses and precise evolution of background error information remains a key challenge in high-frequency DA research (James et al., 2022).

Among classical DA methods, three-dimensional variational assimilation (3DVar) is widely used in high-frequency assimilation due to its timeliness (Yokota et al., 2024). In 3DVar, $\boldsymbol{B}$ is typically estimated using the national meteorological center (NMC) method (Parrish and Derber, 1992). However, the NMC-derived $\boldsymbol{B}$ is static, climatological, and isotropic (hereafter denoted as $\boldsymbol{B}^{\mathrm{s}}$) and fails to capture the flow-dependent characteristics of the atmosphere (Bannister, 2008a, 2008b). To address this drawback, many operational DA systems have adopted the hybrid ensemble–variational (EnVar) assimilation method (Leng et al., 2013), which uses ensemble forecast statistics to derive a flow-dependent error covariance (denoted as $\boldsymbol{B}^{\mathrm{e}}$). A weighted average of $\boldsymbol{B}^{\mathrm{e}}$ and $\boldsymbol{B}^{\mathrm{s}}$ produces the hybrid background error covariance $\boldsymbol{B}^{\mathrm{h}}$, which is incorporated into the 3DVar cost function to improve adaptability to flow variability (Bannister, 2017).

The core of the EnVar method is to combine the strengths of $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$ and aims to improve assimilation accuracy while maintaining computational efficiency. However, it faces three main challenges: first, the quality of the flow-dependent $\boldsymbol{B}$ depends on the accuracy of the ensemble forecasts; second, the computational cost is limited by the cost of collecting ensemble samples; third, assimilation performance is sensitive to the choice of hybrid weights.

Ensemble samples can be obtained using various approaches. Common approaches include the ensemble Kalman filter (EnKF) (Buehner et al., 2005), time-lagged ensemble forecasting (Wang YB et al., 2017; Yokota et al., 2024), and the empirical orthogonal function (EOF) technique (Chen et al., 2020). The EnKF estimates $\boldsymbol{B}^{\mathrm{e}}$ by generating multiple ensemble members, effectively capturing dynamic and nonlinear flow features (Houtekamer et al., 1996). However, the number of ensemble members is limited by computational cost. Operational NWP centers typically use 30–100 members—far fewer than the dimensionality of the model state—resulting in a rank-deficient $\boldsymbol{B}^{\mathrm{e}}$ with substantial sampling errors and spurious long-range correlations. Valid-time-shifting (VTS) ensembles are economical approaches to increasing ensemble size (Lorenc, 2017). VTS includes two components: the valid-time-shifting method for ensemble members (VTSM), which accounts for sampling time and/or phase errors, and the valid-time-shifting method for ensemble perturbations (VTSP), which can be temporally smoothed to reduce pseudo-covariances (Huang B and Wang, 2018; Gasperoni et al., 2022). Another efficient alternative is time-lagged ensemble forecasting, which directly uses forecasts for the same valid time but with different initialization time from the DA system's historical forecast cycle. This significantly reduces computational and storage costs while capturing evolving flow-dependent forecast error covariances. However, long-range forecasts may diverge from truth, limiting the number of usable samples (Wang YB et al., 2017). To further supplement ensemble samples, the EOF technique selects historical forecasts that match the target assimilation period and region, termed optimal historical forecast samples. This method increases the ensemble size at low computational cost while better reflecting current flow-dependent characteristics (Chen et al., 2020).

Although extensive research has shown that EnVar outperforms either variational or ensemble methods, it still has notable limitations. Specifically, the combination of $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$ is typically achieved through linear weighting, with hybrid weights set as fixed empirical parameters ranging from 0 to 1 (Bannister, 2017). This approach presents two main issues: first, it does not fully extract features from the data; second, the choice of mixing parameters lacks scientific justification. A more reasonable alternative would be to derive hybrid weights based on the spatiotemporal characteristics of the data, enabling them to adapt to evolving weather conditions.

In recent years, the "AI for Science" paradigm based on deep learning (DL) has challenged traditional numerical simulation methods in the Earth sciences. Conventional approaches rely on mathematical–physical mechanisms, which are limited in expressiveness and susceptible to human bias (Reichstein et al., 2019). DL, through multilayer neural networks, enables automatic feature extraction and high-level abstraction, thereby improving prediction and classification performance (LeCun et al., 2015). Although DA methods, such as large-scale meteorological models, offer advantages

in medium-range forecasting and inference acceleration, they remain limited in predicting transitional weather and have reduced accuracy and poor interpretability (Lam et al., 2023). Although physics-informed neural networks (PINNs) attempt to incorporate physical constraints, they may introduce additional errors in complex, unresolved problems, potentially degrading the overall performance (Cuomo et al., 2022).

At the level of intelligent decision-making, traditional geoscience models rely on empirical parameters to ensure forecast usability, yet these parameters often lack a solid scientific basis. Reinforcement learning (RL) provides a promising alternative by learning an optimal policy network that maps environmental states to action choices (Kaelbling et al., 1996). Unlike heuristic rules, RL-based decisions are grounded in current state features, thereby improving interpretability over conventional black-box models. Through interaction with the environment and a defined reward mechanism, RL enables scientifically-grounded and adaptive decision-making.

In this paper, to improve assimilation performance under sparse observations and during transitional phases in weather evolution, we propose a deep reinforcement learning (DRL)-based ensemble–variational (DRL-EnVar) hybrid DA method. Built upon the traditional EnVar framework, this method directly incorporates $\boldsymbol{B}^{\mathrm{e}}$, estimated from ensemble samples, into the variational cost function for iterative assimilation. To support high-frequency assimilation, we primarily use the time-lagged ensemble method for its significant time efficiency, supplemented by other cost-effective ensemble expansion strategies. By integrating the powerful feature extraction capabilities of DL and the adaptive decision-making of RL, DRL-EnVar enhances assimilation performance and provides a physically consistent optimal initial state for NWP models, thereby improving forecast accuracy. Our main contributions are as follows:

1. Enhanced data processing and feature extraction. A multi-layer cyclic convolution module leverages neighboring spatial data to mitigate information loss in circular equatorial regions, ensuring consistent representations regardless of segmentation points. Meanwhile, DL modules enhance spatiotemporal feature extraction, improving feature completeness and the utilization of state variables and matrices.

2. Intelligent and adaptive parameter optimization. Hybrid parameter selection in EnVar is formulated as an RL-based decision-making problem, replacing empirical rules with real-time optimization and improving adaptability and assimilation performance.

3. Framework flexibility and consistency. The method preserves theoretical consistency by maintaining flow-dependent $\boldsymbol{B}$, supports both 3DVar and 4DVar frameworks with minimal adjustments, and avoids DL-induced smoothing, ensuring dynamic coherence.

4. Practicality and robustness. DRL-EnVar is simple to implement and embeds the RL policy network into the EnVar framework with low computational cost. Extensive experiments verify its demonstrated transferability, stability, and robustness.

2 Preliminaries

2.1 EnVar

EnVar extends the traditional variational system. For 3DVar, the cost function is expressed as

$$\begin{aligned} J(\boldsymbol{x}) = &\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}^{\mathrm{b}})^{\mathrm{T}}(\boldsymbol{B}^{\mathrm{h}})^{-1}(\boldsymbol{x}-\boldsymbol{x}^{\mathrm{b}}) \\ &+\frac{1}{2}(\boldsymbol{y}^{\mathrm{o}}-\mathcal{H}(\boldsymbol{x}))^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y}^{\mathrm{o}}-\mathcal{H}(\boldsymbol{x})), \end{aligned} \tag{1}$$

where $\boldsymbol{x}$ is the state vector to be analyzed and includes model variables such as temperature, wind, pressure, and humidity; $\boldsymbol{x}^{\mathrm{b}}$ denotes the background state, and $\boldsymbol{y}^{\mathrm{o}}$ represents the observations, in three dimensions. To focus on the study of matrix $\boldsymbol{B}$, the nonlinear observation operator $\mathcal{H}$ is assumed to be linear and fully known, simplified as $\boldsymbol{H}$ $(=\partial\mathcal{H}/\partial\boldsymbol{x})$; $\boldsymbol{R}$ is the observation error covariance matrix.

Notably, $\boldsymbol{B}^{\mathrm{h}}$, defined as a weighted average of $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$, effectively replaces $\boldsymbol{B}$ in the original 3DVar system:

$$\boldsymbol{B}^{\mathrm{h}} = (1-\beta)\boldsymbol{B}^{\mathrm{s}} + \beta\boldsymbol{B}^{\mathrm{e}}, \tag{2}$$

where $\beta \in [0,1]$ is a tunable factor.

2.2 DL

DL represents a potent and versatile approach within artificial intelligence (AI), leveraging multi-layer neural networks to model intricate patterns

and representations in large-scale datasets (LeCun et al., 2015). DL models typically consist of multiple layers of nonlinear computational units, where each layer takes the output of the previous one as its input. This structure enables the automatic learning of high-level abstract feature representations from vast amounts of training data, effectively capturing the distributed characteristics of the input. DL has demonstrated remarkable performance in a wide range of tasks, including image and speech recognition, as well as natural language processing. Among many architectures, fully-connected neural networks (FCNNs) (Sainath et al., 2015), convolutional neural networks (CNNs) (Ketkar and Moolayil, 2021), and gated recurrent units (GRUs) (Cho et al., 2014), i.e., a variant of recurrent neural networks (RNNs) (Gregor et al., 2015), have attracted considerable attention due to their effectiveness in handling specific types of data and tasks.

FCNNs are a fundamental category of artificial neural networks characterized by full connectivity between layers. Each neuron in one layer connects to every neuron in the next, enabling the FCNNs to model complex, nonlinear relationships. This makes them widely applicable in tasks such as classification, regression, and pattern recognition. The multilayer perceptron (MLP) is a widely used type of the FCNNs (Hornik et al., 1989).

CNNs are specifically engineered to handle grid-like data structures, such as images. CNNs, inspired by the visual cortex, use convolutional layers to adaptively learn spatial feature hierarchies from images. These layers detect local patterns like edges and textures, and deeper layers combine them to recognize complex structures and objects. This hierarchical feature extraction empowers the CNNs to attain outstanding performance in image recognition, object detection, and related tasks.

RNNs are designed for sequential data like time series or natural language, and use recurrent cells to process inputs and capture temporal dependencies. However, standard RNNs struggle with vanishing or exploding gradients, limiting long-term dependency learning. GRUs, a refined RNN variant, use update and reset gates to efficiently manage information flow, improving performance in tasks like time-series prediction, language modeling, and speech recognition by mitigating gradient issues while maintaining robust sequence modeling capabilities.

2.3 RL

RL focuses on training agents to make sequential decisions by interacting with environments to maximize cumulative rewards (Kaelbling et al., 1996). This framework relies on the Markov decision process (MDP), characterized by the tuple (S, A, P, R, γ), where S is the state space, A is the action space, P represents the state transition probability, R is the reward function, and γ is the discount factor (Bellman, 1957).

RL problems are primarily tackled using value- and policy-based methods. Value-based methods, like Q-learning and deep Q-network (Mnih et al., 2015), optimize value functions to derive optimal policies, suitable for discrete environments such as Go. Policy-based methods, such as policy gradients, iteratively improve policies, making them ideal for continuous action scenarios like robotic control. The actor–critic framework, a notable approach in RL, integrates both methods to address problems in continuous action spaces and high-dimensional state spaces. It employs two networks: the actor network, which learns parameterized policies for action generation, and the critic network, which evaluates state-action pair values. This dual structure combines value function approximation with direct policy optimization, enhancing learning efficiency and stability.

Proximal policy optimization (PPO) is an advanced RL algorithm that refines the actor–critic framework by introducing a clipped surrogate objective function for stable policy updates. The PPO addresses the high variance and instability issues of traditional policy gradients, offering a balance between simplicity and performance, making it a preferred method for complex RL tasks.

3 Method

This study aims to develop an efficient and cost-effective EnVar method. Unless stated otherwise, all methods are based on the 3DVar framework. The assimilation-forecast cycling system uses the Lorenz-96 model (Lorenz and Emanuel, 1998), a classic toy model for testing DA methods (see the Appendix).

3.1 Problem description

The EnVar method integrates $\boldsymbol{x}^{\mathrm{b}}$ and $\boldsymbol{y}^{\mathrm{o}}$, using their error covariance matrices $\boldsymbol{B}^{\mathrm{h}}$ and $\boldsymbol{R}$, to produce an optimal analysis state $\boldsymbol{x}_i^{\mathrm{a}}$ at the i^{th} time for the numerical forecast model. This enables the numerical forecast model to have the optimal prediction performance at the future time (denoted as the I^{th} time) after this analysis state is taken as the initial state. To directly assess the quality of the analysis state, this study assumes that the initialization process can be omitted for the Lorenz-96 model, proceeding directly with subsequent forecasts. So, the prediction results are close to the true state, denoted as

$$\left\|\boldsymbol{x}_I^{\mathrm{f}}-\boldsymbol{x}_I^{\mathrm{t}}\right\|, \tag{3}$$

where $\boldsymbol{x}^{\mathrm{t}}$ is the true state at the target time, $\boldsymbol{x}_I^{\mathrm{f}}=\mathcal{M}_{i\to I}(\boldsymbol{x}_i^{\mathrm{a}})$ is the forecast at time I, resulting from the model $\mathcal{M}$ taking the analyzed state $\boldsymbol{x}_i^{\mathrm{a}}$ as the initial state, and the prediction time is $I-i$. According to Eq. (1), $\boldsymbol{x}^{\mathrm{a}}$ at time step i is obtained by

$$\boldsymbol{x}_i^{\mathrm{a}}=\boldsymbol{x}_i^{\mathrm{b}}+\boldsymbol{W}\left(\boldsymbol{y}_i^{\mathrm{o}}-\boldsymbol{H}(\boldsymbol{x}_i^{\mathrm{b}})\right), \tag{4}$$

where the weighting matrix $\boldsymbol{W}$ is defined as

$$\boldsymbol{W}=\left[(\boldsymbol{B}^{\mathrm{h}})^{-1}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}. \tag{5}$$

Herein, $\boldsymbol{B}^{\mathrm{h}}$ consists of multiple components:

$$\boldsymbol{B}^{\mathrm{h}}=\alpha_1\boldsymbol{B}_1^{\mathrm{s}}+\cdots+\alpha_m\boldsymbol{B}_m^{\mathrm{s}}+\beta_1\boldsymbol{B}_1^{\mathrm{e}}+\cdots+\beta_n\boldsymbol{B}_n^{\mathrm{e}}, \tag{6}$$

where $\boldsymbol{B}_m^{\mathrm{s}}$ and $\boldsymbol{B}_n^{\mathrm{e}}$ are m static and n ensemble-based background error matrices, respectively, and the coefficients sum to one.

The study is conducted in a context of sparse and noisy observations, featuring year-long cycling assimilation experiments that include a 1-month mutation period during which state variables exceed climatological levels. In Eq. (5), because the statistical terms $\boldsymbol{R}$ and $\boldsymbol{H}$ are assumed to be known and fixed, the accuracy of $\boldsymbol{B}^{\mathrm{h}}$ in representing background errors and its ability to respond to flow changes directly determine the quality of the analysis. As shown in Eq. (6), this accuracy depends on the structure of the combined $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$ matrices, as well as the appropriate choice of blending parameters α and β.

To accurately reflect the real-time background error, this study leverages all available data samples, including historical data and time-lagged ensemble forecasts, without significantly increasing computational cost (Wang YB et al., 2017; Chen et al., 2020; Yokota et al., 2024). The focus is on the intelligent selection of blending parameters from different error sources, emphasizing real-time adaptability and responsiveness to improve assimilation performance.

Based on the aforementioned description, we formalize the task of real-time selection of hybrid parameters in the hybrid assimilation-forecast cycle as a decision-making problem that maps the current state $\boldsymbol{s}_i$ (analysis state) to an action $\boldsymbol{a}_i$ (given a set of parameters), that is,

$$f(\boldsymbol{s}_i)=\boldsymbol{a}_i=[\alpha_1,\alpha_2,\ldots,\alpha_m,\beta_1,\beta_2,\ldots,\beta_n], \tag{7}$$

where i is the current time, and m and n are the parameter counts of $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$.

This can be formulated as an MDP:

1. State $\boldsymbol{s}_i=\boldsymbol{x}_i^{\mathrm{a}}$ is a vector given by the EnVar system at time i.

2. Action $\boldsymbol{a}_i$ comprises weighting factors constrained to sum to one.

3. Transition $\boldsymbol{s}_{i+1}$ is the numerical solution that minimizes the cost function (1) in our task and denoted as

$$\begin{aligned}&\boldsymbol{s}_{i+1}\\ =&\mathcal{M}_{i\to i+1}(\boldsymbol{s}_i)+\boldsymbol{W}\left(\boldsymbol{y}_i^{\mathrm{o}}-\boldsymbol{H}(\mathcal{M}_{i\to i+1}(\boldsymbol{s}_i))\right).\end{aligned} \tag{8}$$

4. Reward $R(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{s}_{i+1})$ is the direct reward of taking action $\boldsymbol{a}_i$ at state $\boldsymbol{s}_i$ and arriving at the new state $\boldsymbol{s}_{i+1}$:

$$R=\left\|\boldsymbol{s}_i-\boldsymbol{x}_i^{\mathrm{t}}\right\|. \tag{9}$$

5. The discount factor γ is set to 0.99 because our task involves long-term continuous planning, where future rewards resulting from continuous actions are crucial.

This study retains the EnVar framework and leverages the DRL to refine $\boldsymbol{B}^{\mathrm{h}}$, enabling effective transmission of flow-dependent information and accurate detection of transitions, avoiding over-smoothing. DRL-EnVar extracts data features by DL to guide RL in training adaptive hybrid weights, dynamically adjusting $\boldsymbol{B}^{\mathrm{h}}$ based on real-time states, substantially improving assimilation performance. Notably, the application of DRL in the DA remains rare, unlike its extensive use in robotics and autonomous driving. This work introduces DRL-EnVar, an innovative method that adaptively selects

hybrid weights according to the specific characteristics of the DA. Rather than simple module integration, the approach combines novel and effective components to reduce the computational cost, shorten training time, and extract meaningful information from limited data, ensuring stable and efficient assimilation-forecast cycling. In the following subsections, we first outline the overall pipeline of the model, and then provide a detailed description of its unique components.

3.2 Overall pipeline

For clarity, the DRL-EnVar framework is divided into two interdependent stages, illustrated in Figs. 1 and 2, representing the DRL training and EnVar assimilation processes respectively; these stages form a closed feedback loop, with the output of each stage cyclically informing the other, jointly enabling a hybrid assimilation system.

In the DRL phase (Fig. 1), a DRL module is designed to optimize an adaptive hybrid strategy for constructing $\boldsymbol{B}^{\mathrm{h}}$. Within a task-specific simulation environment derived from the EnVar system, the agent observes the current state $\boldsymbol{s}_i$, defined as the analysis state $\boldsymbol{x}_i^{\mathrm{a}}$ (step 1), and selects the corresponding action $\boldsymbol{a}_i$ (step 3) using a policy network π_θ, which is iteratively updated via the PPO algorithm. The policy parameters θ are periodically refined by sampling mini-batches from a replay buffer that stores experience tuples $(\boldsymbol{s}_i, \boldsymbol{a}_i, r_i, \boldsymbol{s}_{i+1})$ collected during agent–environment rollouts (step 2). This process enables continuous learning from accumulated experiences and facilitates convergence toward a reward-maximizing policy. To enhance state representation, a multi-layer cyclic CNN (C-CNN) is employed to capture spatial continuity in circular data, while a GRU models temporal dependencies. A softmax layer is applied to enforce simplex constraints, ensuring that the learned hybrid weights are non-negative and sum to one.

In the EnVar phase (Fig. 2), the action $\boldsymbol{a}_i$ generated by the DRL phase (step 1) is used to construct $\boldsymbol{B}^{\mathrm{h}}$ (step 2). This hybrid involves three static $\boldsymbol{B}$ matrices derived from historical samples (denoted as $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F8M15}}$, $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F8}}$, and $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F15}}$), and two flow-dependent matrices computed from modified time-lagged ensemble samples (denoted as $\boldsymbol{B}^{\mathrm{e}}_{48}$ and $\boldsymbol{B}^{\mathrm{e}}_{24}$). Details of the statistical computations are provided in Section 3.3. The resulting $\boldsymbol{B}^{\mathrm{h}}$ is then applied in the EnVar system for assimilation (step 3), and the updated analysis state is fed back to the DRL module as the next environment state (step 4). Each

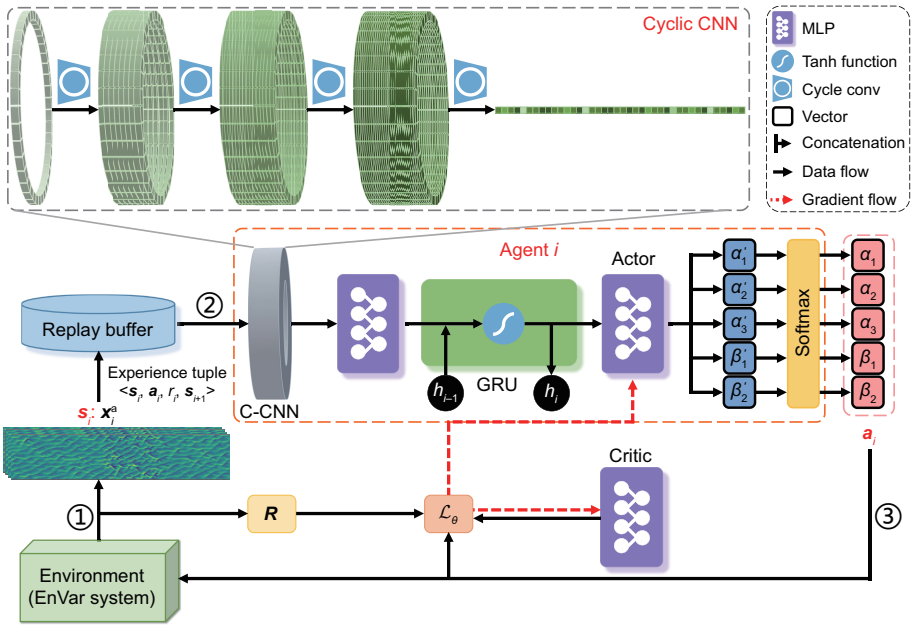

Fig. 1 The agent interacts with an EnVar-based simulation environment to learn a hybrid weighting policy, guided by a C-CNN encoder, a GRU module, and an actor–critic framework. Black solid and red dotted arrows represent data and gradient flows, respectively. References to color refer to the online version of this figure

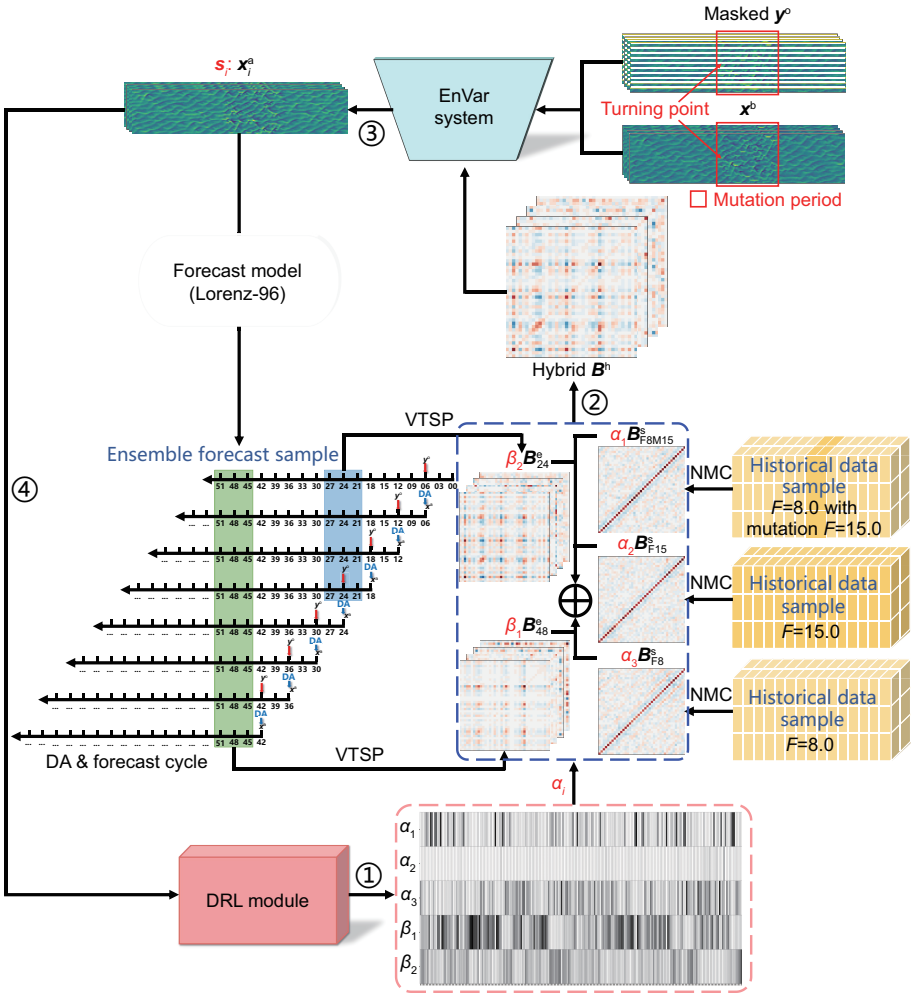

Fig. 2 EnVar assimilation phase. The DRL-output action (weights) is used to combine multiple B matrices. The resulting B^h is applied in the EnVar system for assimilation, and the resulting analysis state is fed back to the DRL module, forming a closed adaptive learning loop. Note that the black and white stripes correspond to the temporal change of weights across a span of 1 year. References to color refer to the online version of this figure

assimilation is followed by a 48-h forecast, with results saved every 3 h and iteratively stacked to generate ensemble samples required for $\boldsymbol{B}^e_{48}$ and $\boldsymbol{B}^e_{24}$.

3.3 Model architecture and algorithm description

3.3.1 C-CNN

C-CNN is a specialized convolutional structure designed for Earth system data, addressing modeling challenges of circular physical spaces. The Lorenz-96 model used in our experiments features variables on equidistant grid points along a latitude circle. Traditional 1D convolutions fail to capture these dependencies, causing information loss and reduced model performance. The C-CNN incorporates circular symmetry, ensuring consistent representations regardless of starting point and enabling high-level abstract feature extraction (Johnson and Zhang, 2017). It mitigates uneven error distribution in the analysis state by leveraging neighbor information, alleviating inconsistencies from sparse, noisy observations, and imperfect background error covariance estimates. The C-CNN enhances feature completeness and reduces disparities between variables with and without assimilated observations.

Each cyclic convolution is followed by a rectified linear unit (ReLU) activation to introduce nonlinearity (Glorot et al., 2011). The ReLU suppresses

negative values and adds sparsity by deactivating neurons while maintaining gradients for positives to mitigate vanishing gradients. Combined with BatchNorm, it stabilizes training and accelerates convergence. This approach enables the model to achieve robust performance and efficiency, even with limited data.

Fig. 3 illustrates a C-CNN with one input channel, eight convolution kernels (size 5), and eight output channels. The circular convolution starts from any point, moving counterclockwise. At each step, the kernel multiplies and sums values at corresponding positions. For example, the yellow dashed box in the input is convolved with eight kernels, producing eight output values forming the yellow dashed boxes in the output. The kernel then shifts to the blue, orange, and green positions, continuing until a full cycle completes. Kernel values are learned and updated during training with the policy network, and adapt to task-specific needs.

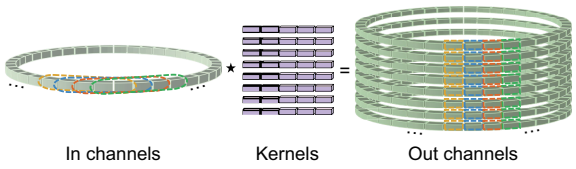

Fig. 3 Cyclic convolution neural network. References to color refer to the online version of this figure

3.3.2 Statistical method

$\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$ are computed statistically with distinct techniques. $\boldsymbol{B}^{\mathrm{s}}$ is typically calculated via the NMC method operationally (Bannister, 2008a). The equation is formulated as

$$\begin{cases} \boldsymbol{B} \approx \dfrac{1}{2}\left\langle (\boldsymbol{x}^{48}-\boldsymbol{x}^{24})(\boldsymbol{x}^{48}-\boldsymbol{x}^{24})^{\mathrm{T}} \right\rangle, \\ \boldsymbol{x}^{48} = \mathcal{M}_{0\to 48}(\boldsymbol{x}^{\mathrm{a}}), \\ \boldsymbol{x}^{24} = \mathcal{M}_{24\to 48}(\boldsymbol{x}^{\mathrm{a}}). \end{cases} \tag{10}$$

The key to the NMC method lies in the composition of historical data samples. Three samples are selected to represent different climate states. The first aligns with the study's background, featuring 1 month of transition (the Lorenz-96 model forcing term $F = 15.0$) and 11 months of stability ($F = 8.0$) per year. The second assumes continuous transitions ($F = 15.0$) throughout the year, effectively capturing historical data consistent with transition characteristics to provide a stronger error correlation for EnVar during transitions. The third assumes no transitions ($F = 8.0$) and reflects background error in stable conditions. Given the rarity of transitions and the prevalence of stability, this information is critical for the assimilation system.

To reduce computational cost and provide real-time flow-dependent information, this study develops an improved method for time-lagged ensemble sampling. First, inspired by the time-lagged ensemble in variational assimilation (Wang CC et al., 2022), the difference between forecast fields at the same time with different forecast lead time is used, such as the difference between the 24-h (blue highlight) and 48-h (green highlight) forecasts in Fig. 2. Second, motivated by the time-shifted ensemble used in some ensemble-based assimilation (Gasperoni et al., 2023), temporal phase errors in model forecasts are addressed by treating neighboring forecasts as valid for the target time, increasing ensemble size and reducing errors. For example, 21-h and 27-h forecasts are treated as 24-h forecasts, whereas 45-h and 51-h forecasts are treated as 48-h forecasts. Combining these two strategies, the ensemble samples used in this study are obtained. Specifically, assimilation is performed every 6 h with a 48-h forecast, saving results every 3 h. The ensemble size for $\boldsymbol{B}^{\mathrm{e}}_{24}$ is 12, and for $\boldsymbol{B}^{\mathrm{e}}_{48}$ it is 24, with both computed using the VTSP method (Huang B and Wang, 2018). These two samples, with distinct forecast lead time and accumulated errors, provide diverse flow-dependent error information for the EnVar system.

3.3.3 DRL-EnVar

DRL-EnVar is an intelligent hybrid assimilation method designed to address the selection of real-time adaptive mixing parameters in the EnVar cycle. By dynamically adjusting the weights of multiple $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$, it effectively responds to evolving observations and weather patterns. This task is formulated as a continuous decision-making problem in high-dimensional state spaces, aligning with classical RL frameworks. The EnVar-customized RL environment E is detailed in Algorithm 1, whereas the standard training process is provided in Algorithm S1 in the supplementary materials.

The initial background state $\boldsymbol{x}_0^{\mathrm{b}}$ of the Lorenz-96 model is set as $X_j = F$ ($j \neq 20$) and $X_j = 1.001F$ ($j = 20$) (Lorenz and Emanuel, 1998). It

is directly assigned to $\boldsymbol{x}_0^{\mathrm{a}}$ as the initial environment state without assimilation. $\boldsymbol{B}^{\mathrm{h}}$ consists of five components, and the agent selects five blending weights summing to one. The reward function combines two metrics: the root mean square error $\mathrm{RMSE_a}$, measuring the RMSE between the analysis state $\boldsymbol{x}^{\mathrm{a}}$ and the true state $\boldsymbol{x}^{\mathrm{t}}$, and $\mathrm{RMSE_f}$, assessing the 48-h forecast $\boldsymbol{x}_{48}^{\mathrm{f}}$ and $\boldsymbol{x}^t$, with $R=-\mathrm{RMSE_a}-\mathrm{RMSE_f}$.

The state transition rule T follows a year-long (360-day) assimilation cycle, excluding 90 transient days, totaling 450 days (Lorenz and Emanuel, 1998). Numerical integration uses the fourth-order Runge–Kutta (RK4) method with a 6-h time step (dt = 0.05), yielding 4 assimilations per day and 1800 iterations (Algorithm 1, line 7). A switch adjusts the forcing term to 15.0 during transitions or 8.0 otherwise.

Updating flow-dependent $\boldsymbol{B}_{48}^{\mathrm{e}}$ and $\boldsymbol{B}_{24}^{\mathrm{e}}$ requires eight assimilations (num_DA = 8). For num_DA < 8, experiments indicate that using climatological $\boldsymbol{B}_{\mathrm{F8M15}}^{\mathrm{s}}$ provides optimal performance. The environment returns states $\boldsymbol{s}_i$, action $\boldsymbol{a}_i$, and reward r_i for DRL-EnVar training.

Policy learning is conducted using PPO, chosen for its robustness in continuous action spaces (Arulkumaran et al., 2017). DRL-EnVar adopts an actor–critic architecture, where both actor and critic are MLPs. C-CNN is integrated to extract spatiotemporal features from the evolving states. This architecture enables the agent to adaptively adjust hybrid weights based on system dynamics and observation patterns. Detailed training steps are provided in Algorithm S1 in the supplementary materials.

Algorithm 1 Definition of EnVar custom environment E

1: **Input:** $\boldsymbol{x}_0^{\mathrm{b}}$, $\boldsymbol{y}_0^{\mathrm{o}}$, $\boldsymbol{R}$, $\boldsymbol{H}$, $\boldsymbol{B}_{\mathrm{F8M15}}^{\mathrm{s}}$, $\boldsymbol{B}_{\mathrm{F15}}^{\mathrm{s}}$, $\boldsymbol{B}_{\mathrm{F8}}^{\mathrm{s}}$, $\mathcal{M}^{\mathrm{F8}}$, $\mathcal{M}^{\mathrm{F15}}$
2: **Output:** states $\boldsymbol{s}_i$, actions $\boldsymbol{a}_i$, rewards r_i
3: Initialize environment state $\boldsymbol{s}_0 \leftarrow \boldsymbol{x}_0^{\mathrm{a}}$, $\boldsymbol{x}_0^{\mathrm{a}} \leftarrow \boldsymbol{x}_0^{\mathrm{b}}$
4: Define action space $\boldsymbol{A}$ as a vector containing 5 weights that sum to 1
5: Define reward function $R=-\mathrm{RMSE_a}-\mathrm{RMSE_f}$
6: Set the state transition rule according to Eq. (8)
7: **for** iteration = 1, 2, …, 1800 **do**
8: $\quad\boldsymbol{s}_{i-1} \leftarrow \boldsymbol{x}_{i-1}^{\mathrm{a}}$
9: $\quad$Execute strategy π_θ for one timestep, and take action $\boldsymbol{a}_i=[\alpha_1,\alpha_2,\alpha_3,\beta_1,\beta_2]_i$ // actor network
10: $\quad$**if** no turning point detected **then**
11: $\quad\quad\mathcal{M} \leftarrow \mathcal{M}^{\mathrm{F8}}$
12: $\quad$**else**
13: $\quad\quad\mathcal{M} \leftarrow \mathcal{M}^{\mathrm{F15}}$
14: $\quad$**end if**
15: $\quad$**if** num_DA < 8 **then**
16: $\quad\quad\boldsymbol{B}^{\mathrm{h}} \leftarrow \boldsymbol{B}_{\mathrm{F8M15}}^{\mathrm{s}}$
17: $\quad\quad$num_DA $\leftarrow$ num_DA + 1
18: $\quad$**else**
19: $\quad\quad$num_DA $\leftarrow$ 0
20: $\quad\quad$Calculate $\boldsymbol{B}_{48}^{\mathrm{e}}$ and $\boldsymbol{B}_{24}^{\mathrm{e}}$
21: $\quad\quad\boldsymbol{B}^{\mathrm{h}} \leftarrow \alpha_1\boldsymbol{B}_{\mathrm{F8M15}}^{\mathrm{s}}+\alpha_2\boldsymbol{B}_{\mathrm{F15}}^{\mathrm{s}}+\alpha_3\boldsymbol{B}_{\mathrm{F8}}^{\mathrm{s}}+\beta_1\boldsymbol{B}_{48}^{\mathrm{e}}+\beta_2\boldsymbol{B}_{24}^{\mathrm{e}}$
22: $\quad$**end if**
23: $\quad\boldsymbol{x}_i^{\mathrm{a}} \leftarrow \boldsymbol{x}_i^{\mathrm{b}}+\left[(\boldsymbol{B}^{\mathrm{h}})^{-1}+\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\boldsymbol{H}\right]^{-1}\boldsymbol{H}^{\mathrm{T}}\boldsymbol{R}^{-1}\cdot\left(\boldsymbol{y}_i^{\mathrm{o}}-\boldsymbol{H}(\boldsymbol{x}_i^{\mathrm{b}})\right)$
24: $\quad\boldsymbol{x}_{48}^{\mathrm{f}} \leftarrow \mathcal{M}_{i+48}(\boldsymbol{x}_i^{\mathrm{a}})$
25: $\quad$Calculate the reward function r_i
26: $\quad\boldsymbol{s}_i \leftarrow \boldsymbol{x}_i^{\mathrm{a}}$
27: **end for**
28: **Return:** $\boldsymbol{s}_i, \boldsymbol{a}_i, r_i$

4 Experimental setup

4.1 Basic configuration of the EnVar

The experiments use the widely adopted Lorenz-96 chaotic model as the numerical model for assimilation method studies (Lorenz and Emanuel, 1998; Kurosawa and Poterjoy, 2023). The model's initial state is a state vector corresponding to a specified number of variables N, and the true state is derived from the model's evolution over the integration time steps. The model's forcing term is set to $F=8.0$ or $F=15.0$. The model integration similarly generates the short-term forecast state; however, to simulate the inherent forecast errors, the forcing term of the Lorenz-96 model is adjusted to $F=8.4$ or $F=15.75$ ($F+\Delta F$, $\Delta F=0.05F$) during the forecasting process.

The assimilation framework is based on the 3DVar method, with simulated observational values assimilated every 6 h. The observations are generated by adding small Gaussian random perturbations to the true state, with the Gaussian noise having a mean of 0 and a covariance matrix specified for the observation error. The observation error is assumed to be spatially independent and uncorrelated, represented by a proportional relationship with the identity matrix $\boldsymbol{I}$, with the standard deviation of the observation error set to 1.0 by default ($\boldsymbol{R}=\sigma_{\boldsymbol{y}^{\mathrm{o}}}\boldsymbol{I}$). The hybrid background error covariance matrix $\boldsymbol{B}^{\mathrm{h}}$ employs a combination of various

mixed background error covariance schemes, with detailed configurations provided in the comparative experimental setup. The verification period for the cycling assimilation system is set to 1 year, preceded by a 90-day spin-up period.

4.2 Setup of the comparative experiments

Two comparative experiments are conducted to validate the method.

The first experiment evaluates EnVar schemes that combine $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$, with mixing weights adaptively selected via DRL. It tests whether matrices $\boldsymbol{B}^{\mathrm{e}}$, derived from modified time-lagged ensemble forecasts, effectively capture flow-dependent features and enhance assimilation performance when hybridized with $\boldsymbol{B}^{\mathrm{s}}$ using DRL-driven parameters.

The second experiment targets mutational periods under sparse observations. It evaluates whether hybrid covariance schemes, blending static and modified time-lagged ensemble covariances, can dynamically adjust mixing weights via DRL to capture real-time background error dynamics. The focus is on enhancing assimilation during mutational periods and across the annual cycle, with comparisons validating adaptability and responsiveness to evolving meteorological conditions.

Six comparative methods are implemented in both experimental settings:

1. Pure 3DVar;

2. CTL-EnVar (Wang YB et al., 2017; Yokota et al., 2024), a time-lagged hybrid assimilation method with a constant weight;

3. EnKF;

4. MLP-EnVar (Huang LL et al., 2025), a hybrid assimilation method employing DRL with MLPs for feature extraction;

5. DRL-EnVar, the proposed C-CNN-based DRL method;

6. Hybrid covariance DA (HCDA) (Yang and Wang, 2024), the classical hybrid covariance DA method.

To address sampling errors and spurious correlations from a limited ensemble size, all methods except pure 3DVar applied localization to the ensemble covariance using the Gaspari–Cohn function, a Gaussian-like fifth-order rational function with weights decaying from 1 to 0 within a defined radius (Gaspari and Cohn, 1999).

4.2.1 Experiment 1: comparison experimental setup under sparse observations without a mutational phase

In this experiment, observations are sparsified using a uniform mask that removes one value for every n state variables. In the Lorenz-96 model with 40 state variables, $n = 9$, 4, 3, and 1 correspond to observation coverage ratios of 10%, 20%, 25%, and 50%, respectively. The system remains steady ($F = 8.0$), without transitional weather processes. This setup assesses whether modified time-lagged ensembles, derived from the assimilation cycle, can capture flow-dependent structures and propagate observational information to unobserved variables under sparse conditions. It also investigates whether RL, combined with DL-based feature extraction, can optimize hybrid weights based on long-term rewards, thereby enhancing background error utilization and improving assimilation accuracy. The design integrates modified time-lagged covariance with DRL to boost performance under sparse observation scenarios.

4.2.2 Experiment 2: comparison experimental setup under sparse observation with a mutational period

This experiment builds on uniformly-masked observations by introducing a 30-day abrupt change period (June 15–July 15) to simulate the East Asian rainy season (Ding and Chan, 2005). During this phase, state variables deviate markedly, with the forcing term set to $F = 15.0$, compared to $F = 8.0$ during stable periods including spin-up. This mutational phase alters climatic characteristics, requiring background error covariance analysis across two contrasting climate states: $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F8M15}}$ and $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F15}}$.

4.3 Selection of empirical parameters for existing assimilation methods: pre-experiments

The fixed mixing parameters for the three methods were reproduced using the enumeration approach adopted by each method, aimed at identifying the optimal fixed parameters.

4.3.1 CTL-EnVar

$\boldsymbol{B}^{\mathrm{h}}$ in CTL-EnVar is computed as

$$\boldsymbol{B}^{\mathrm{h}} = (1-\alpha)\boldsymbol{B}^{\mathrm{s}}_{\mathrm{NMC}} + \alpha\boldsymbol{B}^{\mathrm{e}}_{\mathrm{TL}}, \tag{11}$$

in which $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{NMC}}$ and $\boldsymbol{B}^{\mathrm{e}}_{\mathrm{TL}}$ denote the static $\boldsymbol{B}$ derived from the NMC method and the ensemble-based $\boldsymbol{B}$ estimated from time-lagged (TL) ensemble members, respectively.

The proposed method generates time-lagged ensemble members during 3DVar assimilation using historical forecasts at the same time with varying lead time. Assimilation and 48-h forecasts are performed every 3 h, storing outputs at 3-h intervals to form 16 members and 120 forecast difference samples. For consistency with other methods, the assimilation interval was set to 6 h, and $\boldsymbol{B}^{\mathrm{s}}$ was computed using the NMC method. The original hybrid parameters (α=0.25, 0.5, 0.75, 1.0) were extended to 0–1 with a step size of 0.1, yielding 10 candidates, each tested 10 times with averages used. The initial results identified α=0.1 as the optimal. Further refinement narrowed α to 0.01–0.1 (step 0.01), 0.001–0.01 (step 0.001), and 0.0001–0.001 (step 0.0001), with the optimal range found between 0.0001 and 0.01, depending on observation sparsity and experimental conditions, for the Lorenz-96 model.

4.3.2 EnKF

EnKF serves as a benchmark to evaluate whether DRL-EnVar could match EnKF's performance at significantly lower computational costs. The goal is to determine the optimal ensemble size under various experimental setups. Ensemble sizes are evaluated from 5 to 40, in increments of 5, with 40 chosen as the upper limit, matching the number of state variables in the Lorenz-96 model. A 1:1 variable-to-ensemble ratio is impractical for operational EnKF applications due to high cost. Each configuration is tested 10 times, and the average result is reported.

4.3.3 HCDA

HCDA is benchmarked to evaluate whether EnVar could achieve comparable performance with lower computational cost. Key parameters include ensemble size (5–40, step 5) and the weight parameter between static and ensemble $\boldsymbol{B}$ (0–1, step 0.1). Results (Fig. 4) reveal the joint effects of these parameters on HCDA performance. Increasing the weight initially improves the performance but later causes a decline, whereas increasing the ensemble size had minimal impact under low observation masking. Based on these findings, the ensemble size range was refined to 2–5 (step 1), with the weight parameter range unchanged. Each configuration is tested 10 times, and averages are reported.

4.4 Model performance evaluation metrics

Two metrics are used for performance evaluation. One is $\mathrm{RMSE_a}$, which measures the RMSE between $\boldsymbol{x}^{\mathrm{a}}$ and $\boldsymbol{x}^{\mathrm{t}}$. Another is $\mathrm{RMSE_f}$, which assesses the forecast $\boldsymbol{x}^{\mathrm{f}}$ against $\boldsymbol{x}^{\mathrm{t}}$. Forecasts are generated every 3 h up to 48 h, with the average performance being evaluated at each interval. All experiments are repeated 50 times, and the average results are reported. Further evaluation during the mutation period (1-h forecast performance) for Experiment 2 is included in Fig. S1 in the supplementary materials.

5 Experimental results

This study integrates DL for feature extraction and RL for decision-making to enhance the EnVar. A novel adaptive hybrid assimilation method is proposed to avoid overly smooth results from purely data-driven approaches while efficiently capturing flow-dependent information with minimal computational cost. This approach replaces traditional enumeration with intelligent hybrid strategies. Focusing on state transitions, the method's performance is validated using the Lorenz-96 model. Preliminary results demonstrate its potential to improve assimilation performance. In the future, the method can be extended to operational DA systems.

Before evaluating the proposed method quantitatively, we first present results from two benchmark experiments. Fig. 5 presents EnKF assimilation results with the ensemble size from 5 to 40. Fig. 5a corresponds to Experiment 1 (no mutation period), while Fig. 5b shows Experiment 2 (mutation period). Colored lines indicate results for varying observation ratios. In both subplots, solid lines represent the annual mean (AM) RMSE between $\boldsymbol{x}^{\mathrm{a}}$ and $\boldsymbol{x}^{\mathrm{t}}$, while the dashed line in Fig. 5b indicates the 1-month mean RMSE during the mutation period (MPm). Values in Fig. 5a serve as references for subsequent comparisons. Light gray bars in both subplots show computational costs for different ensemble sizes across the assimilation cycle.

Fig. 4 displays heatmaps of HCDA annual assimilation results under two experimental setups.

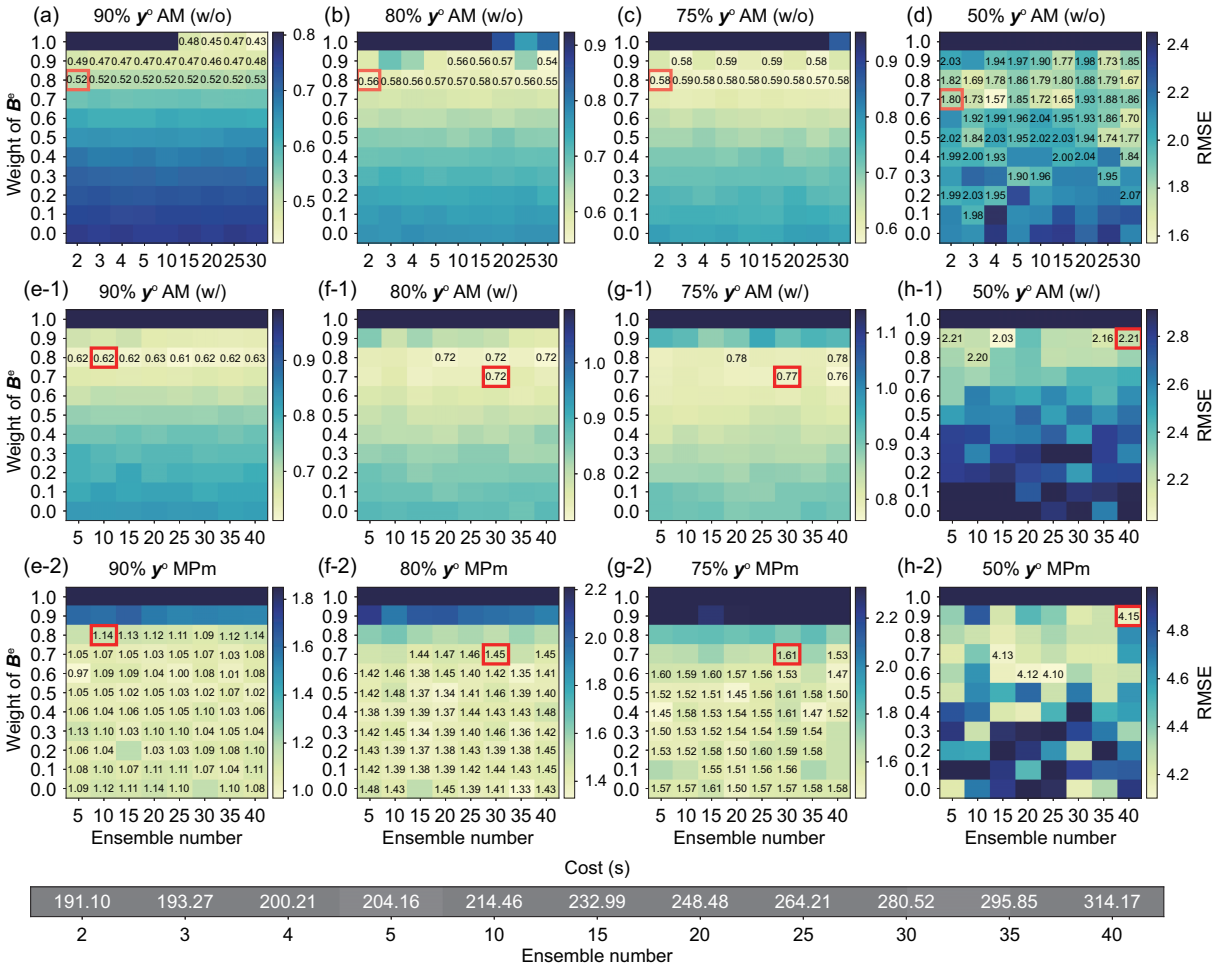

Fig. 4 HCDA assimilation performance heatmaps. Results are shown from left to right for 90% y^o, 80% y^o, 75% y^o, and 50% y^o. The gray heatmap represents the computational cost. Figs. 4a–4d correspond to Experiment 1 (without mutation), and Figs. 4e–4h correspond to Experiment 2 (with mutation). Index 1 represents the AM RMSE (line 2), and index 2 represents the RMSE during the mutation period (MPm) (line 3). References to color refer to the online version of this figure

The x axis represents the ensemble size, while the y axis shows the weight of $\boldsymbol{B}^{\mathrm{e}}$ in $\boldsymbol{B}^{\mathrm{h}}$. Color shading indicates the average RMSE between $\boldsymbol{x}^{\mathrm{a}}$ and $\boldsymbol{x}^{\mathrm{t}}$, with lighter colors representing better performance (lower RMSE) and darker colors representing worse performance (higher RMSE). Columns indicate different observation coverage ratios: 90%, 80%, 75%, and 50%. Values within red boxes are highlighted for detailed discussion in subsequent comparisons. At the bottom, computational costs for a single HCDA run are shown, averaging over 50 repetitions per ensemble size. Darker gray indicates higher cost, with all results representing 50-trial averages.

5.1 Quantitative comparison of assimilation performance under sparse observations without a mutation period

Table 1 summarizes the average assimilation performance of each method over a 1-year cycle under varying observation coverage ratios. Key findings are listed as follows:

1. $\boldsymbol{B}^{\mathrm{e}}$ matrices driven by modified lagged ensemble forecasts effectively capture flow-dependent information. The three methods using time-lagged ensemble forecasts for $\boldsymbol{B}^{\mathrm{h}}$ outperform pure 3DVar, demonstrating their ability to capture flow-dependent information and enhance assimilation.

2. Observation sparsity impacts assimilation performance. As observation sparsity increases, the

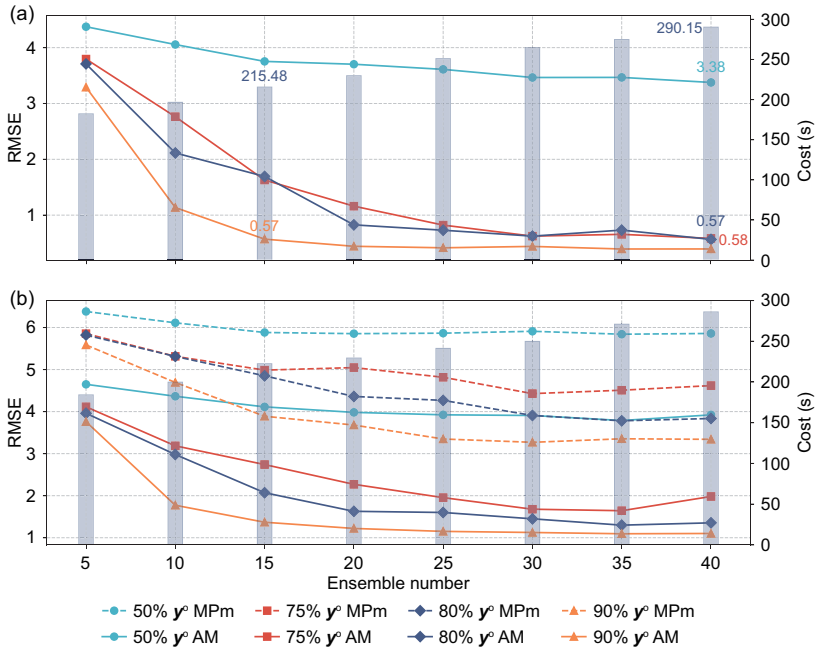

Fig. 5 EnKF annual cycle assimilation results: (a) without mutation period; (b) with mutation period. Results are as follows: light blue represents 50% y^o, orange-red represents 75% y^o, dark blue represents 80% y^o, and orange represents 90% y^o. The solid line represents the AM RMSE between x^a and x^t, while the dashed line represents the mean RMSE during the MPm. All results are the averages of 50 repeated experiments. References to color refer to the online version of this figure

Table 1 AM and the standard deviation for different methods, along with the percentage improvement in the performance of the three methods relative to the pure 3DVar method

Observation ratio	AM±standard deviation			
	Pure 3DVar	CTL-EnVar	MLP-EnVar	DRL-EnVar (ours)
90%	0.5792 ± 0.0085	0.5611 ± 0.0052 (3.13%)	$\underline{0.5505} \pm 0.0050$ (4.96%)	$\mathbf{0.5480} \pm 0.0054$ (5.39%)
80%	0.6439 ± 0.0130	0.6088 ± 0.0084 (5.45%)	$\underline{0.5850} \pm 0.0067$ (9.15%)	$\mathbf{0.5831} \pm 0.0079$ (9.44%)
75%	0.6695 ± 0.0131	0.6239 ± 0.0085 (6.81%)	$\underline{0.6022} \pm 0.0106$ (10.05%)	$\mathbf{0.5995} \pm 0.0089$ (10.46%)
50%	2.8550 ± 0.0992	2.4770 ± 0.1099 (13.24%)	$\underline{2.1890} \pm 0.1019$ (23.33%)	$\mathbf{2.0698} \pm 0.0833$ (27.50%)
Cost (s)	**24.87**	36.82 (1.48)	35.91 (1.44)	$\underline{31.17}$ (1.25)

The final row lists the computational cost for each method, as well as the computational cost multiples relative to the pure 3DVar in the bracket. The best results are in bold and the sub-optimal results are underlined

performance of all three methods improves significantly, emphasizing the critical role of $\boldsymbol{B}$ in delivering flow-dependent information.

3. DRL-EnVar outperforms other methods in assimilation performance improvement. DRL-EnVar is superior to others across all observation sparsity levels (bolded in Table 1). Higher observation coverage ratios yield greater performance gains with minimal computational cost increases.

4. DRL-EnVar achieves a comparable or superior performance to EnKF while requiring lower computational cost. As shown in Fig. 5a, under 90% observation coverage, DRL-EnVar matches the performance of EnKF with 15 ensemble members (EnKF-15, orange triangle) but at a fraction of the computational cost (1.25× vs. 8.66× of pure 3DVar). At 80% and 75% coverage, it performs comparably to EnKF-40, with significantly lower costs (1.25× vs. 11.67×). Under 50% coverage, DRL-EnVar outperforms EnKF, even with sufficient ensemble members. These results show that DRL-EnVar performance matches or exceeds EnKF performance under sparse observations with significantly lower computational cost.

5. HCDA attains a performance comparable to that of DRL-EnVar but with significantly higher

computational expense. As shown in Figs. 4a–4d, the marked values in black are the RMSE values for HCDA and are equal to or lower than those of DRL-EnVar. However, increasing ensemble members under sparse observations does not necessarily improve HCDA performance. The hybrid weight between $\boldsymbol{B}^{\mathrm{s}}$ and $\boldsymbol{B}^{\mathrm{e}}$ is critical, highlighting the importance of intelligent weight selection in this study. Red boxes in the figure indicate configurations achieving comparable performance to DRL-EnVar with minimal computational cost. Notably, HCDA with only two ensemble members and a suitable hybrid weight matches DRL-EnVar performance but still incurs significantly higher computational costs—17.69 times that of 3DVar.

Although Table 1 reports the computational cost during the inference stage, we recognize the importance of disclosing the training cost for transparency and reproducibility. The MLP-EnVar and DRL-EnVar methods involve training compact policy networks with approximately 66 432 and 58 572 trainable parameters, respectively. Both models are trained for 10 million steps (i.e., interactions with the Lorenz-96 simulation environment) using standard hyperparameters (see the supplementary materials for details). Training is conducted on a workstation equipped with an NVIDIA GeForce RTX 3090GPU (approximately 35.7 TFLOPS FP32) and an AMD EPYC 7H12 CPU, requiring approximately 15 h and 27 min for MLP-EnVar and 13 h and 55 min for DRL-EnVar. DRL-based methods require substantial computational resources during training due to extensive environment interaction and policy optimization—a well-recognized characteristic of DRL (Silver et al., 2017). However, this cost constitutes a one-time investment during model development. Once training is complete, the inference phase involves only a single forward pass through a compact neural network, as reflected in Table 1. This training-inference decoupling principle has been widely adopted in large-scale DRL frameworks (Espeholt et al., 2018), enabling real-time deployment with relatively low and stable runtime overhead, which demonstrates the practical feasibility of our approach as validated in the Lorenz-96 model. To further ensure transparency and reproducibility, we provide a comprehensive description of the computational environment in the supplementary materials, including hardware specifications, software versions, and detailed hyperparameter settings.

The primary goal of DA is to provide high-precision initial states for numerical forecasts. In this study, the analysis states that four time points (0, 6, 12, 18 h) on the 1$^{\mathrm{st}}$ and 15$^{\mathrm{th}}$ of each month over a 1-year cycle are used as the initial states for the Lorenz-96 model. Given the model's simplicity, these states require no additional initialization. A 48-h forecast is performed, with forecasts saved every 3 h, generating 96 samples repeated 50 times per experiment. Mean RMSE (line 1) and ACC (line 2) are calculated at 3-h intervals and plotted in Figs. 6a–6d. Index 1 represents the RMSE between $\boldsymbol{x}^{\mathrm{f}}$ and $\boldsymbol{x}^{\mathrm{t}}$, and index 2 shows the ACC. Results indicate that DRL-EnVar consistently delivers the best performance (red solid circular markers), particularly under 50% observation sparsity, where its initial states lead to significantly superior forecasts.

In summary, based on the quantitative results of Experiment 1, our method achieves superior assimilation performance with reduced computational cost under sparse observations, eliminating the need for parameter tuning by enumeration. Unlike CTL-EnVar and HCDA, which require re-enumeration and empirical optimization of hybrid weights when system settings change (e.g., varying observation sparsity), our approach adapts seamlessly.

5.2 Quantitative comparison of assimilation performance under sparse observations and with a mutation period

Fig. 7 presents the AM (yellow-orange solid line, AM) and mutation period average (dark blue dashed line, MPm) RMSE between $\boldsymbol{x}^{\mathrm{a}}$ and $\boldsymbol{x}^{\mathrm{t}}$ for various DA methods under different observation sparsity ratios, with a 1-month mutation during the assimilation cycle. The error bar for each value is marked. Key findings are as follows:

1. DRL-EnVar outperforms all sparsity levels. Consistent with Experiment 1, DRL-EnVar achieves the smallest AM and MPm across all sparsity settings (bold) in Experiment 2.

2. DRL-EnVar demonstrates an enhanced performance during mutation periods. Compared to Experiment 1, DRL-EnVar shows superior performance, particularly during mutation periods. Although MLP-EnVar is generally the second-best, CTL-EnVar outperforms it at 90% and 75% sparsity levels but suffers from poor stability, as indicated

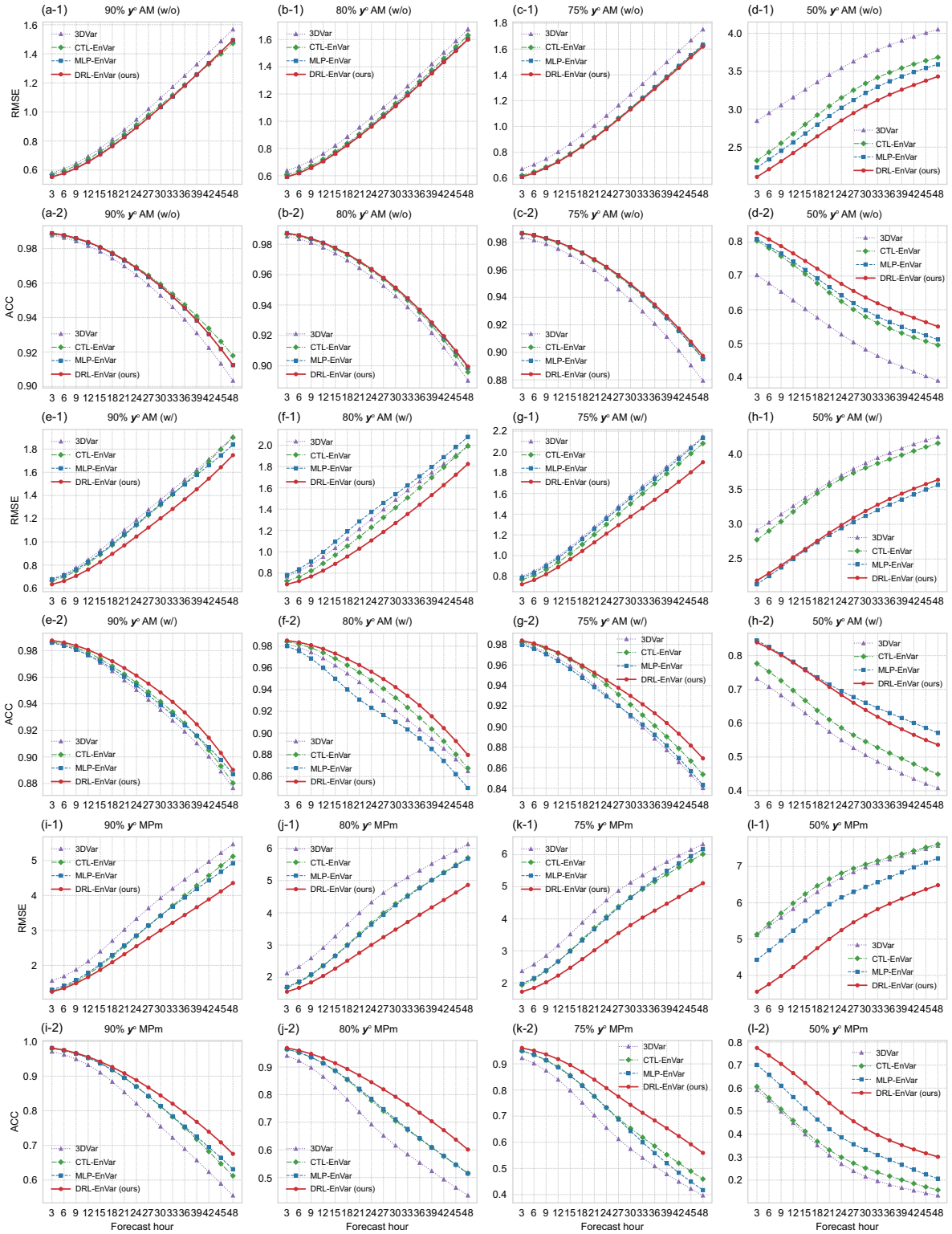

Fig. 6 Comparison of 3-h forecasts for averaged RMSE (line 1) and ACC (line 2) curves for different models under two experimental setups. Results are shown from left to right for 90% y^o, 80% y^o, 75% y^o, and 50% y^o. DRL-EnVar is marked with red circles, MLP-EnVar with blue squares, CTL-EnVar with green diamonds, and pure 3DVar with purple triangles. References to color refer to the online version of this figure

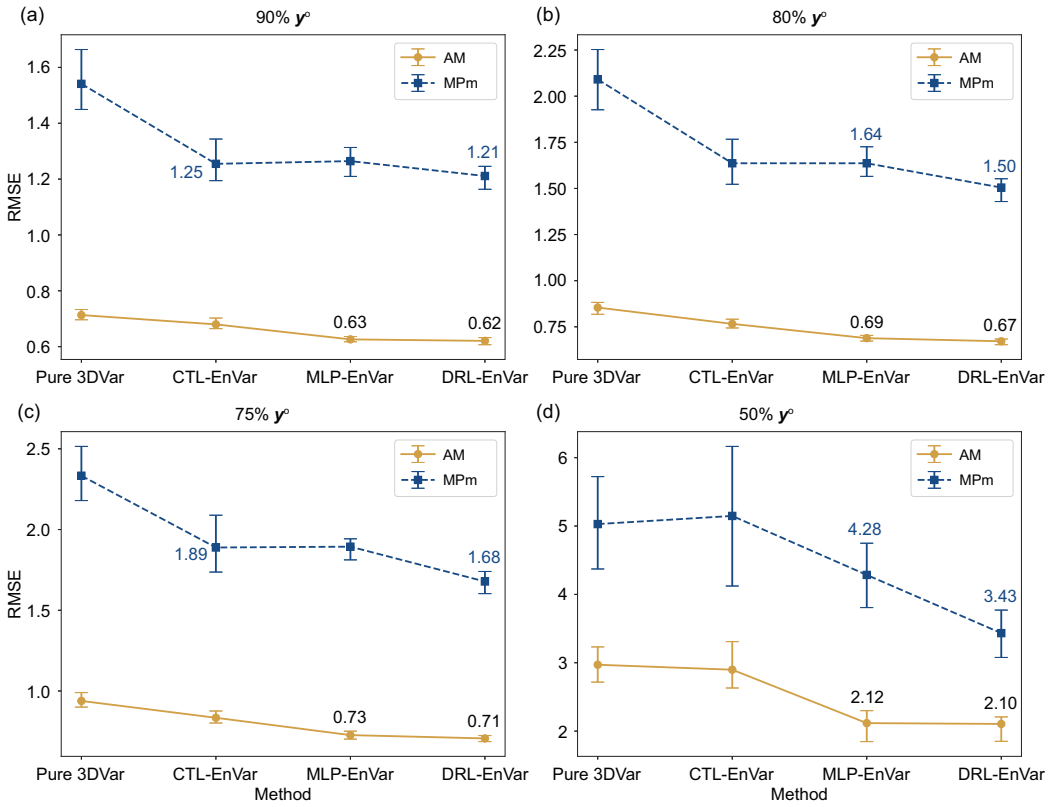

Fig. 7 RMSE of assimilation methods under varying observation sparsity, with an anomaly occurring in the state variable during 1 month of the annual assimilation cycle: (a) 90% y^o; (b) 80% y^o; (c) 75% y^o; (d) 50% y^o. The solid yellow-orange line represents the AM, while the dashed dark blue line represents the MPm. Error bars are shown for each value. References to color refer to the online version of this figure

by larger error bars. The comparison highlights the effectiveness of the C-CNN module in DRL-EnVar, with its advantages becoming more pronounced at higher sparsity levels.

3. EnKF underperforms DRL-EnVar in sparse and mutation conditions. As shown in Fig. 5b, EnKF fails to achieve ideal performance under sparse observations and mutation conditions, even with many ensemble members. EnKF lags significantly behind DRL-EnVar in AM and MPm results, indicating that increasing ensemble size alone cannot resolve issues like filter divergence, even with localization. Addressing these limitations in EnKF warrants further investigation.

4. HCDA relies on weight enumeration and large ensembles, increasing computational cost. Figs. 4e–4h show HCDA AM (index 1, line 2) and MPm (index 2, line 3) results across ensemble sizes (x axis) and $\boldsymbol{B}^{\mathrm{e}}$ weights (y axis). Values matching or outperforming those of DRL-EnVar are highlighted, with red boxes marking the minimal ensemble size meeting this criterion. DRL-EnVar matches HCDA performance with 10 or more ensemble members (e.g., at 90% observation sparsity, it approximates HCDA-E10W0.8; at 80% sparsity, HCDA-E30W0.7; at 75% and 50%, HCDA-E30W0.7 and HCDA-E40W0.9, respectively). However, HCDA's computational cost exceeds 200 s, far higher than that of DRL-EnVar. Furthermore, HCDA's reliance on enumerated weights, sensitive to factors like model type, ensemble method, and observation sparsity, necessitates re-enumeration for the optimal performance under changing conditions, lacking robust theoretical support.

Based on 3-h forecast results, DRL-EnVar consistently achieves a near-optimal performance in the RMSE and ACC between $\boldsymbol{x}^{\mathrm{f}}$ and $\boldsymbol{x}^{\mathrm{t}}$, as shown by comparisons for both AM (Fig. 6, lines 3–4) and MPm (Fig. 6, lines 5–6), with pronounced advantages during mutation periods.

In conclusion, our proposed DRL-EnVar method leverages the feature extraction capabilities of a

hybrid C-CNN and MLP architecture to provide a robust representation of spatiotemporal data. Combined with RL-driven decision-making, it enables optimal hybrid weight selection and ensures consistently high assimilation performance across both stable and mutation periods.

5.3 Qualitative analysis of DRL-EnVar

Fig. 8 depicts the evolution of background error covariance weights over 360 days under 50% observation sparsity, with a mutation period. The results, following model convergence on the test set, reveal time-dependent changes in weight selection. As shown in the method model diagram, the weights for matrix $\boldsymbol{B}$ exhibit temporal variation throughout the assimilation forecast cycle, with all matrix weights non-zero, indicating continuous contribution to $\boldsymbol{B}^{\mathrm{h}}$ (α_1 corresponds to the weight of $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F8M15}}$, α_2 to $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F15}}$, α_3 to $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F8}}$, β_1 to $\boldsymbol{B}^{\mathrm{e}}_{24}$, and β_2 to $\boldsymbol{B}^{\mathrm{e}}_{48}$).

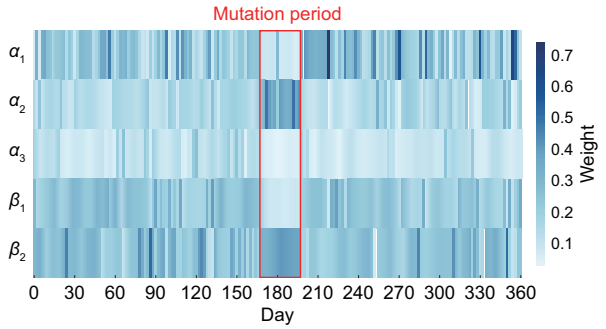

Fig. 8 Heatmap of variable weights at different time steps during the annual assimilation cycle. The gradient from light blue to dark blue represents the weight variation, with light colors indicating low weights and dark colors indicating high weights. The red rectangle highlights the weight variation patterns of variables during the mutation period. References to color refer to the online version of this figure

DRL-EnVar dynamically adjusts the weights of each covariance matrix after each assimilation, based on evolving state variables. Over the 360-d period, every 8 cycles yield 5 weight selections per action, totaling 225 weight updates. These selections exhibit significant fluctuations throughout the assimilation cycle, highlighting the method's capacity to adapt to variations in background error covariances.

The red rectangular box in the figure highlights weight changes during the mutation period, illustrating the method's idealized response. Notably, during this period, the weights of $\boldsymbol{B}^{\mathrm{s}}_{\mathrm{F15}}$ and $\boldsymbol{B}^{\mathrm{e}}_{48}$ increase significantly, dominating the total weight. This behavior confirms the DRL-EnVar method's ability to swiftly respond to mutations in state variables, adjusting mixed weights and updating $\boldsymbol{B}^{\mathrm{h}}$. Such adjustments allow the method to more accurately represent error correlations and effectively handle abrupt changes in state variables.

6 Discussion

6.1 Transferability of DRL-EnVar to 4DVar

To explore whether the DRL-EnVar framework can generalize beyond 3DVar, we conduct a preliminary test by embedding DRL-based hybrid weight selection into a 4DVar system (denoted DRL-En4DVar). Under similar experimental settings, DRL-En4DVar achieves the best or second-best assimilation performance across multiple observation sparsity ratios, particularly under high-sparsity conditions. These results suggest that the core DRL-based decision framework is compatible with 4DVar and retains its adaptability advantages. However, further optimization is needed to fully leverage 4DVar's temporal assimilation characteristics during regime transitions. Detailed results and discussion are provided in Fig. S2 in the supplementary materials.

6.2 Data sample selection and theoretical justification for constructing $\boldsymbol{B}$

In DA, the true state variables are inaccessible, making direct computation of background error information impossible. Thus, background error estimation relies on assumptions and approximations, such as the innovation covariance method (Buehner et al., 2005), NMC method, and ensemble methods. Due to sparse observations in this study, the innovation covariance method is not applicable. The NMC method simulates background errors using the differences between model predictions at different lead time, whereas the ensemble method averages the differences among ensemble members, offering a more realistic reflection of background errors.

Both NMC and ensemble methods depend on background error information within data samples. For this study, data samples were selected based on climate state and real-time modified time-lagged samples (Section 3.3), focusing on DL for feature

extraction and RL for decision-making, minimizing computational cost. Although the sample selection process was informed by literature and expert judgment, future research should focus on a more rigorous approach to sample selection and a quantitative evaluation of different $\boldsymbol{B}$'s impact on assimilation performance.

6.3 Observation masking uniformity

This study employs a uniform masking strategy tailored to the Lorenz-96 model, although real-world observational sparsity is typically non-uniform. This raises concerns about whether the DRL-EnVar method will maintain its performance when applied to real-world data. However, the core DRL approach in DA—extracting data features via DL and making decisions with RL—remains unchanged. The C-CNN module in DRL-EnVar extracts features based on the Lorenz-96 model's cyclic characteristics, which align with the spherical features of real-world data (e.g., latitude and longitude). Extending the 2D C-CNN to a 3D version is sufficient for real-world validation.

However, non-uniform observational sparsity may degrade C-CNN performance in real-world applications, necessitating further validation. Additionally, this sparsity may introduce other challenges, requiring adjustments based on the model's specific characteristics. Despite these issues, the framework of data feature-based intelligent decision-making remains robust, and the application of DRL in DA holds significant promise for future research.

6.4 Complexity of the mutation period

This study aims to assess whether DRL can enhance hybrid DA methods within computational limits and improve assimilation performance, with a focus on hybrid assimilation innovations. Experiments are conducted using the Lorenz-96 model. Although DRL-EnVar shows significant improvements in annual cycle assimilation, especially during mutation periods, the experimental setup is simpler than real-world meteorological phenomena.

In contrast to the Lorenz-96 model, which involves a single variable with limited complexity, real-world systems, such as the East Asian rainy season, exhibit high seasonal variability and complex covariance relationships among multiple variables. The variation patterns of real atmospheric systems are far more complex and nonlinear, especially during mutation periods in the rainy season.

Future work will apply DRL-powered DA methods to real-world scenarios, such as the rainy season, to verify their practical effectiveness. The goal is to optimize regional assimilation performance while ensuring timeliness and improving forecasts of atmospheric variables like humidity and precipitation.

6.5 Real-world adaptability

The DRL-EnVar has demonstrated effectiveness in the Lorenz-96 model, yet its applicability to operational NWP remains an open challenge. The core paradigm—DL for feature extraction, RL for decision-making, and the $\boldsymbol{B}$ matrix for transmitting flow-dependent information—provides a transferable foundation for extension to more complex meteorological settings. However, this transition from low-dimensional, uniformly sampled Lorenz-96 states to high-dimensional, multivariate real-world states (e.g., temperature, wind, and humidity) introduces challenges, particularly with heterogeneous data sources such as gridded reanalysis, remote sensing imagery, and point clouds. The C-CNN module, designed for 1D cyclic data, may be insufficient for such inputs, motivating the adoption of alternative architectures like Earthformer for spatiotemporal fields (Gao et al., 2022) and PointNet for non-Euclidean data structures (Qi et al., 2017).

Additionally, the action space, defined by hybrid weights, must align with the parameterization used in operational DA systems (e.g., WRFDA), where control variable transforms (CVT) represent $\boldsymbol{B}$. This necessitates redefining actions as spatially and temporally adaptive hybrid parameters for the CVT-based components of $\boldsymbol{B}$—including variances, eigenvectors, eigenvalues, and horizontal/vertical length scales—ensuring compatibility with operational DA formulations and improving assimilation under heterogeneous observational coverage and flow-dependent uncertainties. Building on the current model-based success, future work will aim to validate DRL-EnVar in realistic NWP environments by incorporating multimodal data and model characteristics, with sensitivity analysis guiding the evolution of the feature extraction pipeline and action design to ensure scalability and robustness in operational use.

7 Conclusions

Traditionally, the performance of variational DA has been enhanced by linearly combining static and ensemble background error covariance matrices. However, conventional EnVar hybrid assimilation methods are computationally expensive and lack transferability. To address these issues, we propose a novel DRL-based hybrid assimilation method, DRL-EnVar. This approach frames the selection of hybrid parameters as an intelligent decision-making task, using an innovative cyclic convolution module and an MLP to extract data features. RMSE is used as a reward function to optimize the mixing weights, generating a real-time background error covariance matrix $\boldsymbol{B}^{\mathrm{h}}$, which effectively improves assimilation performance under sparse observations and during rapid transitions in weather conditions.

Experimental validation shows that DRL-EnVar significantly outperforms traditional methods at a lower computational cost. Specifically, at 50% observation coverage, the computational cost is only 1.25 times that of pure 3DVar, yet the performance is comparable to that of EnKF-15. Furthermore, under sparse observation and during rapid state transitions, DRL-EnVar intelligently selects mixing weights using DRL, optimizes the spatiotemporal adaptability of the background error covariance, and significantly enhances assimilation stability and efficiency. Compared to traditional methods that rely on large ensemble members, DRL-EnVar reduces computational costs while improving assimilation performance, offering a more competitive solution for handling complex state transitions.

Overall, this work empowers traditional variational assimilation with DRL, alleviating smoothing issues in large-scale weather models by enhancing the flow-dependent characteristics of background errors. The successful validation of this method in toy models lays the foundation for its application in real-world weather assimilation, indicating a promising future for research and application in this field.

Contributors

Lilan HUANG, Hongze LENG, and Junqiang SONG designed the research. Dongzi WANG conducted the verification. Lilan HUANG and Hongze LENG drafted the paper. Wuxin WANG helped organize the paper. Ruisheng HU and Hang CAO revised the paper. Lilan HUANG and Hongze LENG finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

References

Arulkumaran K, Deisenroth MP, Brundage M, et al., 2017. Deep reinforcement learning: a brief survey. *IEEE Signal Process Mag*, 34(6):26-38. https://doi.org/10.1109/MSP.2017.2743240

Bannister RN, 2008a. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. *Quart J Roy Meteor Soc*, 134(637):1951-1970. https://doi.org/10.1002/qj.339

Bannister RN, 2008b. A review of forecast error covariance statistics in atmospheric variational data assimilation. II: modelling the forecast error covariance statistics. *Quart J Roy Meteor Soc*, 134(637):1971-1996. https://doi.org/10.1002/qj.340

Bannister RN, 2017. A review of operational methods of variational and ensemble-variational data assimilation. *Quart J Roy Meteor Soc*, 143(703):607-633. https://doi.org/10.1002/qj.2982

Bellman R, 1957. A Markovian decision process. *J Math Mech*, 6(5):679-684. https://doi.org/10.1512/iumj.1957.6.56038

Buehner M, Gauthier P, Liu Z, 2005. Evaluation of new estimates of background- and observation-error covariances for variational assimilation. *Quart J Roy Meteor Soc*, 131(613):3373-3383. https://doi.org/10.1256/qj.05.101

Chen YD, Guo S, Meng DM, et al., 2020. The impact of optimal selected historical forecasting samples on hybrid ensemble-variational data assimilation. *Atmos Res*, 242:104980. https://doi.org/10.1016/j.atmosres.2020.104980

Cho K, van Merriënboer B, Gulcehre C, et al., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proc Conf on Empirical Methods in Natural Language Processing, p.1724-1734. https://doi.org/10.3115/v1/D14-1179

Cuomo S, Di Cola VS, Giampaolo F, et al., 2022. Scientific machine learning through physics-informed neural networks: where we are and what's next. *J Sci Comput*, 92(3):88. https://doi.org/10.1007/s10915-022-01939-z

Ding YH, Chan JCL, 2005. The East Asian summer monsoon: an overview. *Meteor Atmos Phys*, 89(1):117-142. https://doi.org/10.1007/s00703-005-0125-z

Espeholt L, Soyer H, Munos R, et al., 2018. IMPALA: scalable distributed deep-RL with importance weighted actor-learner architectures. Proc 35$^{\mathrm{th}}$ Int Conf on Machine Learning, p.1406-1415.

Gao ZH, Shi XJ, Wang H, et al., 2022. Earthformer: exploring space-time Transformers for Earth system forecasting. Proc 36$^{\mathrm{th}}$ Advances in Neural Information Processing Systems, p.25390-25403.

Gaspari G, Cohn SE, 1999. Construction of correlation functions in two and three dimensions. *Quart J Roy Meteor Soc*, 125(554):723-757. https://doi.org/10.1002/qj.49712555417

Gasperoni NA, Wang XG, Wang YM, 2022. Using a cost-effective approach to increase background ensemble member size within the GSI-based EnVar system for improved radar analyses and forecasts of convective systems. *Mon Wea Rev*, 150(3):667-689. https://doi.org/10.1175/MWR-D-21-0148.1

Gasperoni NA, Wang XG, Wang YM, 2023. Valid time shifting for an experimental RRFS convection-allowing EnVar data assimilation and forecast system: description and systematic evaluation in real time. *Mon Wea Rev*, 151(5):1229-1245. https://doi.org/10.1175/MWR-D-22-0089.1

Glorot X, Bordes A, Bengio Y, 2011. Deep sparse rectifier neural networks. Proc 14th Int Conf on Artificial Intelligence and Statistics, p.315-323.

Gregor K, Danihelka I, Graves A, et al., 2015. DRAW: a recurrent neural network for image generation. Proc 32nd Int Conf on Machine Learning, p.1462-1471.

Hornik K, Stinchcombe M, White H, 1989. Multilayer feedforward networks are universal approximators. *Neur Netw*, 2(5):359-366. https://doi.org/10.1016/0893-6080(89)90020-8

Houtekamer PL, Lefaivre L, Derome J, et al., 1996. A system simulation approach to ensemble prediction. *Mon Wea Rev*, 124(6):1225-1242. https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2

Huang B, Wang XG, 2018. On the use of cost-effective valid-time-shifting (VTS) method to increase ensemble size in the GFS hybrid 4DEnVar system. *Mon Wea Rev*, 146(9):2973-2998. https://doi.org/10.1175/MWR-D-18-0009.1

Huang LL, Leng HZ, Song JQ, et al., 2025. An adaptive variance adjustment strategy for a static background error covariance matrix—part I: verification in the Lorenz-96 model. *Appl Sci*, 15(12):6399. https://doi.org/10.3390/app15126399

James EP, Alexander CR, Dowell DC, et al., 2022. The high-resolution rapid refresh (HRRR): an hourly updating convection-allowing forecast model. Part II: forecast performance. *Wea Forecast*, 37(8):1397-1417. https://doi.org/10.1175/WAF-D-21-0130.1

Johnson R, Zhang T, 2017. Deep pyramid convolutional neural networks for text categorization. Proc 55th Annual Meeting of the Association for Computational Linguistics, p.562-570. https://doi.org/10.18653/v1/P17-1052

Kaelbling LP, Littman ML, Moore AW, 1996. Reinforcement learning: a survey. *J Artif Intell Res*, 4:237-285. https://doi.org/10.1613/jair.301

Kalman RE, 1960. A new approach to linear filtering and prediction problems. *J Basic Eng*, 82(1):35-45. https://doi.org/10.1115/1.3662552

Ketkar N, Moolayil J, 2021. Convolutional neural networks. In: Ketkar N, Moolayil (Eds.), Deep Learning with Python: Learn Best Practices of Deep Learning Models with PyTorch. Apress, Berkeley, CA, USA. https://doi.org/10.1007/978-1-4842-5364-9_6

Kurosawa K, Poterjoy J, 2023. A statistical hypothesis testing strategy for adaptively blending particle filters and ensemble Kalman filters for data assimilation. *Mon Wea Rev*, 151(1):105-125. https://doi.org/10.1175/mwr-d-22-0108.1

Lam R, Sanchez-Gonzalez A, Willson M, et al., 2023. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416-1421. https://doi.org/10.1126/science.adi2336

LeCun Y, Bengio Y, Hinton G, 2015. Deep learning. *Nature*, 521(7553):436-444. https://doi.org/10.1038/nature14539

Leng HZ, Song JQ, Yin FK, et al., 2013. Notes and correspondence on ensemble-based three-dimensional variational filters. *J Zhejiang Univ SCIENCE C*, 14(8):634-641. https://doi.org/10.1631/jzus.C1300024

Lorenc AC, 2017. Improving ensemble covariances in hybrid variational data assimilation without increasing ensemble size. *Quart J Roy Meteor Soc*, 143(703):1062-1072. https://doi.org/10.1002/qj.2990

Lorenz EN, Emanuel KA, 1998. Optimal sites for supplementary weather observations: simulation with a small model. *J Atmos Sci*, 55(3):399-414.

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

Parrish DF, Derber JC, 1992. The national meteorological center's spectral statistical-interpolation analysis system. *Mon Wea Rev*, 120(8):1747-1763. https://doi.org/10.1175/1520-0493(1992)120<1747:tnmcss>2.0.co;2

Qi CR, Su H, Mo K, et al., 2017. PointNet: deep learning on point sets for 3D classification and segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.77-85. https://doi.org/10.1109/cvpr.2017.16

Reichstein M, Camps-Valls G, Stevens B, et al., 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743):195-204. https://doi.org/10.1038/s41586-019-0912-1

Sainath TN, Vinyals O, Senior A, et al., 2015. Convolutional, long short-term memory, fully connected deep neural networks. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.4580-4584. https://doi.org/10.1109/ICASSP.2015.7178838

Sanz-Alonso D, Stuart A, Taeb A, 2023. Inverse problems and data assimilation. Cambridge University Press, New York, USA. https://doi.org/10.1017/9781009414319

Silver D, Schrittwieser J, Simonyan K, et al., 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354-359. https://doi.org/10.1038/nature24270

Wang CC, Tsai CH, Jou BJD, et al., 2022. Time-lagged ensemble quantitative precipitation forecasts for three landfalling typhoons in the Philippines using the CReSS model, part II: verification using global precipitation measurement retrievals. *Remote Sens*, 14(20):5126. https://doi.org/10.3390/rs14205126

Wang YB, Min JZ, Chen YD, et al., 2017. Improving precipitation forecast with hybrid 3DVar and time-lagged ensembles in a heavy rainfall event. *Atmos Res*, 183:1-16. https://doi.org/10.1016/j.atmosres.2016.07.026

Yang Y, Wang XG, 2024. A comparison of 3DEnVar and 4DEnVar for convective-scale direct radar reflectivity data assimilation in the context of a filter and a smoother. *Mon Wea Rev*, 152(1):59-78. https://doi.org/10.1175/MWR-D-23-0082.1

Yokota S, Banno T, Oigawa M, et al., 2024. JMA operational hourly hybrid 3DVar with singular vector-based mesoscale ensemble prediction system. *J Meteor Soc Japan Ser II*, 102(2):129-150. https://doi.org/10.2151/jmsj.2024-006

Appendix: The Lorenz-96 model

The Lorenz-96 model is a typical nonlinear system, whose governing equation is

$$\frac{\mathrm{d}X_j}{\mathrm{d}t} = (X_{j+1} - X_{j-2})\,X_{j-1} - X_j + F,$$

where $j = 1, 2, \ldots, J$. $J = 40$ denotes the scalar state variables on 40 equally spaced grid points around a latitude circle. $F = 8.0$ and an RK4 scheme with a time step $\mathrm{d}t = 0.05$ (counting as 6 h) are used for solving the equations numerically.

List of supplementary materials

Algorithm S1 DRL-EnVar training algorithm

Fig. S1 Hourly forecasted average RMSE trends over forecast days for four methods at different observation data ratios

Fig. S2 Comparison of assimilation performance of four methods at different observation ratios

Table S1 Core training hyperparameters used in the DRL-EnVar method

Supplementary evaluation during mutation period (Experiment 2)

Supplementary explanation of computational environment and hyperparameter settings

Supplementary discussion: transferability of DRL-EnVar to 4DVar