SCHRÖDINGER BRIDGE DATA ASSIMILATION WITH A CONSISTENT PROBABILISTIC VIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Data assimilation (DA), which estimates the current atmospheric state by combining prior forecasts with sparse observations, is critical to the accuracy of modern weather prediction systems. The recently developed diffusion DA methods exhibit several superior aspects in comparison to the classical ones. Although DA targets a state distribution conditioned on both the background and observations, they learn a background-conditioned distribution during training and incorporate observations only during inference, separating observation conditioning from generative learning. To address this limitation, we propose Schrödinger Bridge Data Assimilation (SBDA), which integrates classical 3D-Var with diffusion Schrödinger bridges through an observation-enhanced reparameterization of DA conditioning, providing a consistent probabilistic treatment of observations across bridge learning and sampling. The resulting bridge generates analyses directly from observation-enhanced background states rather than Gaussian noise. We evaluate SBDA using background forecasts from a deep learning weather model, 1.5° ERA5 reference fields, and simulated observations at 1% of the ERA5 grid points. SBDA achieves lower single-step assimilation errors than the evaluated baselines on most variables, with a mean relative reduction of 12.9% across all variables. More importantly, SBDA consistently improves downstream forecast across all variables, yielding a forecast time gain of up to 13.3 h relative to the strongest evaluated baseline. Furthermore, SBDA retains substantial assimilation gains at unseen observation locations without retraining.

1 INTRODUCTION

Weather prediction plays a vital role in modern society, supporting applications such as disaster mitigation, transportation, agriculture, energy management, and climate research (Bauer, 2024). Numerical weather prediction (NWP) (Bauer et al., 2015), the dominant paradigm in traditional weather forecasting, generates predictions by numerically solving the complex partial differential equations governing atmospheric dynamics, resulting in substantial computational cost. Recently, deep learning weather models have emerged as a promising alternative to traditional NWP systems, achieving comparable forecasting skill while requiring only a fraction of the computational resources and offering orders-of-magnitude faster inference (Bi et al., 2023; Lam et al., 2023; Bodnar et al., 2025). Most deep learning weather models are trained on high-quality reanalysis datasets such as ERA5 (Hersbach et al., 2020). However, as ERA5 is generated through a computationally intensive retrospective data assimilation process that assimilates observations over an extended temporal window, it is not available in real time and therefore cannot be directly used as the initial condition in operational forecasting. Consequently, similar to traditional NWP systems, deep learning weather models also rely on data assimilation (DA) to provide an estimate of the current atmospheric state. The quality of this initial condition plays a crucial role in forecast accuracy, since initial errors can propagate and amplify during model integration, particularly in autoregressive forecasting systems (Lorenz, 1963; Bauer et al., 2015). DA combines real-time observations with prior short-range forecasts, termed the *background*, to produce an updated state estimate, termed the *analysis* (Lei et al., 2025).

Classical DA methods are mainly categorized into variational and sequential approaches (Asch et al., 2016). Variational methods formulate DA as an optimization problem that minimizes the mismatch among the analysis, background state, and observations, including 3D-Var (Lorenc, 1981)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

for single-step assimilation and 4D-Var (Courtier et al., 1994) for assimilation within a temporal window. Sequential methods, represented by the Ensemble Kalman Filter (EnKF) (Evensen, 1994), recursively update the model state using newly available observations and estimate flow-dependent background error statistics through forecast ensembles. Despite their success in NWP, classical DA methods suffer from several shortcomings: 1) the accurate specification of background error covariance structures in high-dimensional nonlinear systems is hard, and 2) they are computationally expensive because they require solving high-dimensional optimization problems or propagating ensembles through numerical models (Carrassi et al., 2018).

These limitations motivate the development of deep learning-based DA methods. Let $\mathbf{x}^b$ denote the background state, $\mathbf{y}$ the observations, and $\mathbf{x}^*$ the reference state. Existing deterministic learning-based DA methods are commonly formulated as end-to-end supervised regression models that learn a mapping $(\mathbf{x}^b, \mathbf{y}) \rightarrow \mathbf{x}^*$ (Arcucci et al., 2020; Chen et al., 2023; Xu et al., 2025). Although effective at producing a single analysis, these deterministic mappings do not explicitly represent the flow-dependent uncertainty that remains under imperfect backgrounds and sparse observations (Leutbecher & Palmer, 2008). From a probabilistic perspective, DA is more naturally formulated as estimating the conditional distribution $p(\mathbf{x}^* \mid \mathbf{x}^b, \mathbf{y})$. Motivated by this view, diffusion-based generative models (Ho et al., 2020; Song et al., 2021) have emerged as a powerful framework for DA. Existing diffusion-based DA methods generally train background-conditioned models for $p(\mathbf{x}^* \mid \mathbf{x}^b)$ and incorporate observations $\mathbf{y}$ only during inference through guided sampling (Huang et al., 2024; Sun et al., 2025). Nevertheless, these approaches exhibit two design limitations in the DA setting. First, the DA target is conditioned on observations, whereas these methods learn the reverse dynamics without explicitly incorporating them, creating a mismatch in observation conditioning between training and inference. Second, generating analyses from Gaussian noise does not fully exploit the structured background and may therefore limit analysis quality (Liu et al., 2023).

To address these limitations, we propose Schrödinger Bridge Data Assimilation (SBDA), which introduces an observation-enhanced reparameterization of DA conditioning to incorporate observations into both training and inference. Meanwhile, the diffusion Schrödinger bridge formulation (Liu et al., 2023) enables SBDA to generate analyses directly from structured states rather than Gaussian noise. Specifically, we integrate classical 3D-Var (Lorenc, 1981) into the development of the reparameterization and factorize the DA conditional distribution into two factors that both incorporate observational information. The first factor is modeled through bridge learning, while the second motivates observation refinement during sampling. During training, observations shape the terminal distribution of DSB. During inference, the available observations further guide analysis generation through per-step refinement of the predicted reference state. Our contributions are threefold:

1. We develop SBDA through an observation-enhanced reparameterization of DA conditioning implemented with classical 3D-Var, providing a consistent probabilistic treatment of observations across bridge learning and sampling.
2. We introduce a diffusion Schrödinger bridge formulation for weather data assimilation that generates analyses directly from observation-enhanced background states rather than Gaussian noise.
3. Experiments demonstrate that SBDA outperforms the evaluated baselines across most variables in single-step assimilation and consistently across all variables in downstream forecasting, while retaining substantial assimilation gains at unseen observation locations without retraining.

2 Related Work

Schrödinger bridges for field reconstruction. Diffusion Schrödinger bridges (Bortoli et al., 2021; Chen et al., 2022) model stochastic transport between two arbitrary endpoint distributions. $\mathrm{I}^2\mathrm{SB}$ (Liu et al., 2023) learns bridges between paired degraded and clean images, enabling reconstruction directly from structurally informative inputs. For physical fields, PalSB (Li et al., 2025) incorporates a loss based on the governing equations into bridge learning. PICSB (Bu et al., 2026) considers reconstruction from low-fidelity fields and sparse observations without complete high-fidelity supervision. It constructs iteratively updated surrogate endpoints under observation constraints and trains the reconstruction process using PDE residuals. The above methods provide a foundation for extending diffusion Schrödinger bridges to weather data assimilation.

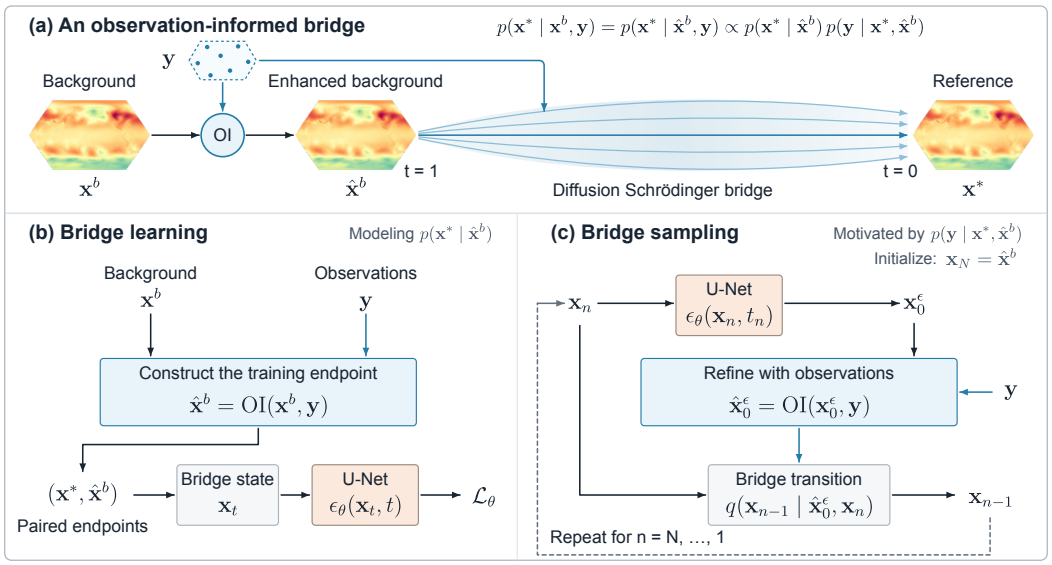


Figure 1: **Framework of SBDA. (a)** An observation-enhanced reparameterization provides a consistent probabilistic treatment of observations across bridge learning and sampling. **(b)** Optimal interpolation (OI) constructs the observation-enhanced terminal distribution, and paired endpoint training learns the first factor. **(c)** The second factor motivates per-step OI refinement of the predicted reference state before each bridge transition.

Diffusion-based weather data assimilation. The score-based filter (SF) and ensemble score filter (EnSF) (Bao et al., 2024b;a) obtain scores from predicted sample ensembles and assimilate observations through likelihood-guided diffusion sampling. Manshausen et al. (2025) instead learn an unconditional prior from historical reference states and use sparse weather-station observations to guide the generation of regional surface wind and precipitation fields. DiffDA (Huang et al., 2024) and LO-SDA (Sun et al., 2025) further incorporate forecast backgrounds with reference states into prior learning, while introducing observations during inference through inpainting or latent optimization. In these approaches, observation conditioning is introduced only through inference-time guidance. SBDA extends this conditioning to both training and inference.

3 Methods

Figure 1 presents an overview of SBDA, which provides a consistent probabilistic treatment of observations across bridge learning and sampling through an observation-enhanced reparameterization of DA. In this section, we first formulate DA probabilistically and introduce the DSB construction, before describing observation-enhanced bridge learning and observation-refined bridge sampling.

3.1 Problem Formulation

On a global grid consisting of n grid points, the reference state of a single atmospheric variable can be represented as a vector $\mathbf{x}^* \in \mathbb{R}^n$. Data assimilation aims to estimate the most probable state of it, denoted as $\mathbf{x}^a \in \mathbb{R}^n$, at a given time by combining a background state $\mathbf{x}^b \in \mathbb{R}^n$ with observations $\mathbf{y} \in \mathbb{R}^m$:

$$\mathbf{x}^a = \arg\max_{\mathbf{x}} \; p(\mathbf{x} \mid \mathbf{x}^b, \mathbf{y}). \tag{1}$$

The observations $\mathbf{y}$ are related to the reference state $\mathbf{x}^*$ through an observation operator $\mathcal{H}$, such that $\mathbf{y} = \mathcal{H}(\mathbf{x}^*)$. In this study, the observations are obtained from the reference state $\mathbf{x}^*$ at randomly selected grid points. Under this setting, $\mathcal{H}$ becomes a sparse matrix $\mathbf{H} \in \mathbb{R}^{m \times n}$.

Notation. Let $\mathbf{x}_t \in \mathbb{R}^n$ be an n-dimensional stochastic process indexed by $t \in [0, 1]$, and let $0 = t_0 < t_1 < \cdots < t_N = 1$ be a discretization of the time interval. Throughout the paper, we use the shorthand $\mathbf{x}_n := \mathbf{x}_{t_n}$.

3.2 DIFFUSION SCHRÖDINGER BRIDGE

Originating from optimal transport and stochastic processes, the Schrödinger bridge (Schrödinger, 1932; Léonard, 2012; Chen et al., 2016) offers a probabilistic framework for efficiently transporting an initial distribution p_0 to a terminal distribution p_1 through the following SDEs:

$$\mathrm{d}\mathbf{x}_t = [f_t + \beta_t \nabla \log \Psi(\mathbf{x}_t, t)]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t, \tag{2a}$$

$$\mathrm{d}\mathbf{x}_t = [f_t - \beta_t \nabla \log \widehat{\Psi}(\mathbf{x}_t, t)]\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}\overline{W}_t, \tag{2b}$$

where $\mathbf{x}_0 \sim p_0$, $\mathbf{x}_1 \sim p_1$, $t \in [0, 1]$ is time variable, f_t is drift term, β_t is diffusion term, W_t and $\overline{W}_t$ denote standard forward and backward Wiener processes. The functions $\Psi, \widehat{\Psi} \in C^{2,1}(\mathbb{R}^n \times [0, 1])$ are time-varying energy potentials governed by a coupled PDE system (Liu et al., 2023). Compared with conventional diffusion models (Ho et al., 2020; Song et al., 2021), the diffusion Schrödinger bridge (DSB) model generalizes the terminal distribution p_1 from a Gaussian distribution to an arbitrary distribution. To avoid directly solving the coupled PDE system of Ψ and $\widehat{\Psi}$, we adopt the paired-data DSB formulation proposed by I^2SB (Liu et al., 2023). Set $p_0 \coloneqq \delta_a$, where $a \in \mathbb{R}^n$ denotes a single data point, such that the backward SDE Equation (2b) can be interpreted as the reverse process of the following linear SDE,

$$\mathrm{d}\mathbf{x}_t = f_t(\mathbf{x}_t)\mathrm{d}t + \sqrt{\beta_t}\mathrm{d}W_t,\ \mathbf{x}_0 \sim \widehat{\Psi}(\cdot, 0), \tag{3}$$

and the initial condition is then given by $\widehat{\Psi}(\cdot, 0) = \delta_a(\cdot)$. For an endpoint pair $(\mathbf{x}_0, \mathbf{x}_1)$ with $\mathbf{x}_1 \sim p_1(\cdot \mid \mathbf{x}_0)$, provided $f_t = 0$, the posterior of Equation (2) admits an analytic form:

$$q(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_1) = \mathcal{N}(\mathbf{x}_t; \mu_t, \mathbf{\Sigma}_t), \tag{4}$$

$$\mu_t = \frac{\bar{\sigma}_t^2}{\sigma_t^2 + \bar{\sigma}_t^2}\mathbf{x}_0 + \frac{\sigma_t^2}{\sigma_t^2 + \bar{\sigma}_t^2}\mathbf{x}_1,\ \mathbf{\Sigma}_t = \frac{\sigma_t^2 \bar{\sigma}_t^2}{\sigma_t^2 + \bar{\sigma}_t^2} \cdot \mathbf{I},$$

where $\sigma_t^2 = \int_0^t \beta_\tau \mathrm{d}\tau$ and $\bar{\sigma}_t^2 = \int_t^1 \beta_\tau \mathrm{d}\tau$ are variance accumulated from either sides. The same formulation applies to an arbitrary terminal time t_n, with $\mathbf{x}_1$ replaced by $\mathbf{x}_n$ and $\bar{\sigma}_t^2 = \int_t^{t_n} \beta_\tau \, \mathrm{d}\tau$. The model is then trained using paired data $(\mathbf{x}_0, \mathbf{x}_1)$ by optimizing the score-matching objective associated with Equation (3):

$$\begin{aligned}\mathcal{L}_\theta &= \mathbb{E}_{\mathbf{x}_0 \sim p_0, \mathbf{x}_1 \sim p_1(\cdot|\mathbf{x}_0), t \sim \mathcal{U}(0,1)} \|\epsilon_\theta(\mathbf{x}_t, t) + \sigma_t \nabla \log p(\mathbf{x}_t, t \mid \mathbf{x}_0)\|_2^2, \\ &= \mathbb{E}_{\mathbf{x}_0 \sim p_0, \mathbf{x}_1 \sim p_1(\cdot|\mathbf{x}_0), t \sim \mathcal{U}(0,1)} \left\| \epsilon_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t} \right\|_2^2.\end{aligned} \tag{5}$$

3.3 OBSERVATION-ENHANCED BRIDGE LEARNING

Background-conditioned diffusion DA methods (Huang et al., 2024; Sun et al., 2025) can be understood through the Bayesian factorization

$$p(\mathbf{x}^* \mid \mathbf{x}^b, \mathbf{y}) \propto p(\mathbf{x}^* \mid \mathbf{x}^b)\, p(\mathbf{y} \mid \mathbf{x}^*), \tag{6}$$

assuming that observations and backgrounds are conditionally independent given the reference state $\mathbf{x}^*$. These methods learn the background-conditioned factor from data and impose observational constraints during inference. Consequently, the generative model is not directly trained to use the current observations to resolve ambiguities left by the background. To incorporate observations into both factors, we introduce an observation-enhanced representation $\hat{\mathbf{x}}^b = \mathcal{T}(\mathbf{x}^b, \mathbf{y})$, where $\mathcal{T}$ is chosen to be invertible in $\mathbf{x}^b$ for fixed $\mathbf{y}$. The DA target can therefore be expressed as

$$p(\mathbf{x}^* \mid \mathbf{x}^b, \mathbf{y}) = p(\mathbf{x}^* \mid \hat{\mathbf{x}}^b, \mathbf{y}) \propto p(\mathbf{x}^* \mid \hat{\mathbf{x}}^b)\, p(\mathbf{y} \mid \mathbf{x}^*, \hat{\mathbf{x}}^b). \tag{7}$$

This reparameterization preserves the DA target and incorporates observational information into both factors, providing a common probabilistic basis for bridge learning and observation-refined sampling. To model $p(\mathbf{x}^* \mid \hat{\mathbf{x}}^b)$, we set $p(\hat{\mathbf{x}}^b)$ as the terminal distribution and $p(\mathbf{x}^*)$ as the initial distribution of the DSB. Next, we specify the formulation of the invertible transformation $\mathcal{T}$.

3D-Var-inspired endpoint construction. Classical 3D-Var (Lorenc, 1981) provides a variational formulation for combining background and observational information under prescribed error statistics. Given a background state $\mathbf{x}^b$ and observations $\mathbf{y}$, 3D-Var attempts to produce the analysis state

$\mathbf{x}^a$ by solving the following optimization problem:

$$\mathbf{x}^a = \arg\min_{\mathbf{x}} \left[\frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^\top \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2}(\mathbf{y} - \mathcal{H}(\mathbf{x}))^\top \mathbf{R}^{-1}(\mathbf{y} - \mathcal{H}(\mathbf{x})) \right], \tag{8}$$

where $\mathbf{B} \in \mathbb{R}^{n \times n}$ denotes the background error covariance and $\mathbf{R} \in \mathbb{R}^{m \times m}$ the observation error covariance. In our sparse-observation setting, the observation operator $\mathcal{H}$ is set to a sparse matrix $\mathbf{H}$, under which this problem admits a closed-form solution, referred to as optimal interpolation (OI):

$$\mathbf{x}^a = \mathrm{OI}(\mathbf{x}^b, \mathbf{y}) \coloneqq \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b), \tag{9}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^\top (\mathbf{H}\mathbf{B}\mathbf{H}^\top + \mathbf{R})^{-1}. \tag{10}$$

This formulation meets all our requirements: it optimally combines background and observations under the prescribed error covariances while remaining invertible in $\mathbf{x}^b$ for fixed $\mathbf{y}, \mathbf{H}, \mathbf{B}, \mathbf{R}$ with $\mathbf{R} \succ 0$. Since the true background error covariance is generally unknown in practice, it remains to specify an appropriate $\mathbf{B}$.

Intuitively, observations should primarily influence regions of the background field surrounding the observation locations, with their influence gradually decaying as distance increases. To achieve this, we define $B_{ij} \coloneqq \sigma_b^2 \exp(-d_{ij}^2/L^2)$, where $\sigma_b \in \mathbb{R}$ denotes the background-error magnitude, d_{ij} denotes the physical distance between points i and j, and L denotes the correlation length scale. This covariance model imposes isotropic spatial correlations, with an exponential decay with respect to the squared distance. Although the observations coincide with the reference values at the observed locations in our setting, we prescribe $\mathbf{R} \coloneqq \sigma_r^2 \mathbf{I}$, with $\sigma_r = \alpha\sigma_b$ and $0 < \alpha \ll 1$, to regularize the inversion in Equation (10). This small observation uncertainty relative to the background-error scale also relaxes exact observation matching, partially mimicking the presence of observation errors in practical applications. In this manner, we can obtain the observation-enhanced background state $\hat{\mathbf{x}}^b = \mathrm{OI}(\mathbf{x}^b, \mathbf{y})$. The paired data $(\mathbf{x}^*, \hat{\mathbf{x}}^b)$ are then used to train the DSB model ϵ_θ by optimizing Equation (5), thereby learning a bridge from the observation-enhanced background distribution to the reference state distribution.

3.4 Observation-Refined Bridge Sampling

To incorporate the observation constraint represented by the second factor $p(\mathbf{y} \mid \mathbf{x}^*, \hat{\mathbf{x}}^b)$ in Equation (7), we first examine the vanilla DDPM-style sampling procedure of DSB (Ho et al., 2020; Liu et al., 2023). Initialize the reverse process at $\mathbf{x}_N = \hat{\mathbf{x}}^b$ and generate an analysis $\mathbf{x}_0$ through N reverse transitions. The reverse transition from $\mathbf{x}_n$ to $\mathbf{x}_{n-1}$ at each sampling step is formulated as follows:

$$\mathbf{x}_0^\epsilon = \mathbf{x}_n - \sigma_{t_n} \epsilon_\theta(\mathbf{x}_n, t_n), \quad \mathbf{x}_{n-1} \sim q(\mathbf{x}_{n-1} \mid \mathbf{x}_0^\epsilon, \mathbf{x}_n).$$

Here, $\mathbf{x}_0^\epsilon$ denotes the predicted reference state, and the transition distribution q is given by Equation (4) with terminal time t_n.

Per-step observation refinement. In our noiseless observation setting, the second factor $p(\mathbf{y} \mid \mathbf{x}^*, \hat{\mathbf{x}}^b)$ in Equation (7) restricts candidate reference states to those satisfying $\mathbf{H}\mathbf{x}^* = \mathbf{y}$. This motivates refining the predicted reference state $\mathbf{x}_0^\epsilon$ according to $\mathbf{y}$ at each reverse step before using it in the bridge transition. We implement this refinement through a local Gaussian surrogate centered at $\mathbf{x}_0^\epsilon$, with a soft constraint on observation consistency:

$$\widetilde{p}_n(\mathbf{x} \mid \mathbf{x}_0^\epsilon, \mathbf{y}) \propto \mathcal{N}(\mathbf{x}; \mathbf{x}_0^\epsilon, \mathbf{B}) \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{x}, \mathbf{R}). \tag{11}$$

Maximizing this surrogate density yields $\hat{\mathbf{x}}_0^\epsilon = \mathrm{OI}(\mathbf{x}_0^\epsilon, \mathbf{y})$, the same form as in bridge learning. We reuse the same background covariance $\mathbf{B}$ and observation covariance $\mathbf{R}$ from endpoint construction. The corrected sampling step is formulated as follows:

$$\mathbf{x}_0^\epsilon = \mathbf{x}_n - \sigma_{t_n} \epsilon_\theta(\mathbf{x}_n, t_n), \tag{12}$$

$$\hat{\mathbf{x}}_0^\epsilon = \mathrm{OI}(\mathbf{x}_0^\epsilon, \mathbf{y}), \tag{13}$$

$$\mathbf{x}_{n-1} \sim q(\mathbf{x}_{n-1} \mid \hat{\mathbf{x}}_0^\epsilon, \mathbf{x}_n), \tag{14}$$

The final line uses the Gaussian bridge transition in Equation (4), with $\hat{\mathbf{x}}_0^\epsilon$ replacing $\mathbf{x}_0^\epsilon$ as the reference endpoint and thereby modifying the transition mean. For deterministic inference, we replace each draw from $q(\mathbf{x}_{n-1} \mid \hat{\mathbf{x}}_0^\epsilon, \mathbf{x}_n)$ with its conditional mean, retaining the same per-step refinement. Observations thus shape the endpoint distribution during bridge learning and the reverse transition means during sampling, connecting the two stages within a consistent probabilistic view of DA. The detailed training and inference algorithms are available in Section A.2.

Table 1: **Single-step assimilation LW-RMSE on the 2018 evaluation set.** Best and second-best results are bold and underlined. Δ_{bg} and Δ_{best} denote SBDA's relative changes (%) versus the background and variable-wise best baseline, respectively. Negative values indicate improvements. The final row averages these percentage changes equally across all 69 variables.

	LW-RMSE ↓				SBDA change (%)	
Variable (unit)	Background	3D-Var	DiffDA	SBDA	Δ_{bg}	Δ_{best}
$z500$ ($\mathrm{m^2\,s^{-2}}$)	139.74	131.68	<u>128.48</u>	**84.01**	−39.88	−34.62
$u500$ ($\mathrm{m\,s^{-1}}$)	2.86	<u>2.77</u>	2.95	**2.40**	−16.12	−13.21
$v500$ ($\mathrm{m\,s^{-1}}$)	2.96	<u>2.86</u>	3.00	**2.45**	−17.04	−14.21
$t850$ (K)	1.13	<u>1.09</u>	1.18	**1.03**	−9.39	−6.06
$q700$ ($\mathrm{g\,kg^{-1}}$)	0.87	<u>0.86</u>	0.91	**0.83**	−5.03	−3.17
$t2m$ (K)	<u>1.08</u>	**1.04**	1.13	1.21	+12.26	+16.86
$u10$ ($\mathrm{m\,s^{-1}}$)	1.47	<u>1.42</u>	1.52	**1.26**	−14.17	−11.55
$v10$ ($\mathrm{m\,s^{-1}}$)	1.52	<u>1.48</u>	1.57	**1.30**	−14.40	−11.96
msl (Pa)	147.47	139.09	<u>136.18</u>	**99.09**	−32.81	−27.24
Mean (69 variables)	–	–	–	–	−16.63	−12.90

4 Experiments

4.1 Experimental Settings

To evaluate the effectiveness of SBDA for assimilating forecasts generated by deep learning weather foundation models, we use 48-hour forecast fields produced by Aurora (Bodnar et al., 2025) as the background states. The observations are taken directly from all 69 variables at 288 randomly selected grid points (1% of the global grid) of the ERA5 reference state. Unless otherwise specified, the observation locations are fixed across training and evaluation and shared by all methods; complete ERA5 fields serve as the reference states. Data from 1989 to 2017 are used for training, and data from 2018 are reserved for evaluation.

We compare SBDA against the background states without data assimilation, the classical DA method 3D-Var (Lorenc, 1981), and the diffusion-based approach DiffDA (Huang et al., 2024), which addresses the same assimilation task. Although LO-SDA (Sun et al., 2025) also addresses this task, we exclude it from the comparison because its source code is not publicly available. For a fair comparison, we implement DiffDA using the same simplified U-Net backbone as SBDA, while retaining the additional background-conditioning input required by its conditional diffusion formulation. We evaluate 69 variables: four surface variables and five atmospheric variables at 13 pressure levels, as detailed in the appendix. All fields are represented at a spatial resolution of 1.5°, corresponding to a global grid of size 120×240.

4.2 Single-step Assimilation

Given a 48-hour forecast background state and sparse observations at the target time, each method generates an analysis field, which is then evaluated against the ERA5 reference state. For the LW-RMSE comparison, we use deterministic SBDA inference and average the LW-RMSE computed at each evaluation time.

Aggregated Errors. Table 1 reports LW-RMSE for nine representative variables, while Figure 4 in the appendix shows relative error changes across all 69 variables. SBDA achieves the lowest LW-RMSE on 66 variables and improves upon the background on 68, with a mean relative reduction of 16.6%. The largest gains relative to the background occur for geopotential (z), averaging 34.9% across 13 pressure levels, and mean sea-level pressure (msl), with a reduction of 32.8%. However, LW-RMSE for two-meter temperature ($t2m$) increases by 12.3% relative to the background. This degradation may partly stem from imbalanced optimization across heterogeneous variables trained jointly under a shared mean-squared-error (MSE) objective.

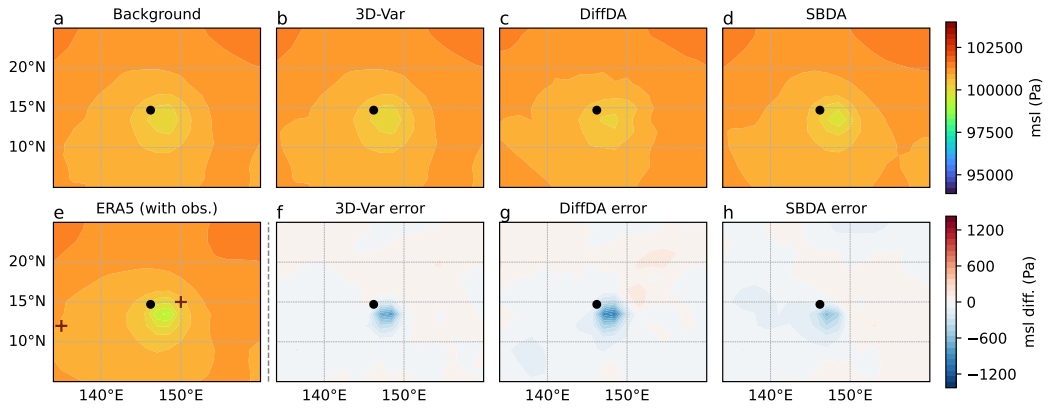


Figure 2: **Mean sea-level pressure near Typhoon Yutu at 00:00 UTC on 24 October 2018.** Background (**a**), analyses from 3D-Var (**b**), DiffDA (**c**), and SBDA (**d**), ERA5 with observation locations (**e**, dark red crosses), and analysis errors for 3D-Var (**f**), DiffDA (**g**), and SBDA (**h**). Black circles mark the cyclone center; errors are computed as $\mathrm{ERA5} - \mathrm{Field}$.

Case Study on Typhoon. We examine the reconstruction of the localized low-pressure structure around Typhoon Yutu using the mean sea-level pressure (msl) field at 00:00 UTC on 24 October 2018 (Figure 2). This case requires correcting an overly weak background pressure depression without an observation at the cyclone center, as neither of the two observation locations in the displayed region samples the center (Figure 2(a,e)). The pronounced negative-error cores retained by 3D-Var and DiffDA (Figure 2(f,g)) indicate that both analyses still underestimate the depth of the local pressure depression. SBDA weakens this error core (Figure 2(h)) while recovering a deeper pressure depression (Figure 2(d,e)), supporting partial reconstruction of the localized low under sparse observations, although its depth remains underestimated relative to ERA5.

Table 2: **20-member SBDA ensemble versus deterministic inference.** Coverage and width refer to central 80% ensemble intervals.

Variable (unit)	Det. CRPS	Ens. CRPS	CRPSS (%)	Coverage (%)	Width
$z500$ ($\mathrm{m^2\,s^{-2}}$)	56.36	44.36	21.30	58.29	109.28
$u500$ ($\mathrm{m\,s^{-1}}$)	1.77	1.39	21.26	46.46	2.60
$q700$ ($\mathrm{g\,kg^{-1}}$)	0.55	0.43	21.25	54.48	0.86
$t2m$ (K)	0.80	0.61	23.97	76.44	2.26
Mean (69 variables)	–	–	21.59	53.84	–

Probabilistic Analysis. At each analysis time, we evaluate a 20-member stochastic SBDA ensemble against the deterministic output of the same trained model, treating the latter as a point-mass distribution. The ensemble achieves positive CRPS skill scores (CRPSS) across all 69 variables, averaging 21.59% (Table 2). The central 80% prediction intervals have a mean empirical coverage of 53.84% across variables, suggesting room for improved calibration.

4.3 Impact on Forecast Skill

To evaluate the practical value of improved analyses, we further assess their impact on downstream forecasts. Specifically, we initialize Aurora forecasts from each analysis or the background and evaluate LW-RMSE at 6-h intervals from 6 to 48 h (Figure 3).

SBDA achieves the lowest forecast LW-RMSE across all 69 variables at every evaluated lead time from 6 to 48 h. Matching SBDA's 48-h LW-RMSE to the strongest baseline's error curves yields estimated forecast time gains of 10.8–13.3 h for the six representative variables, with an unweighted mean of 11.3 h across all 69 variables.

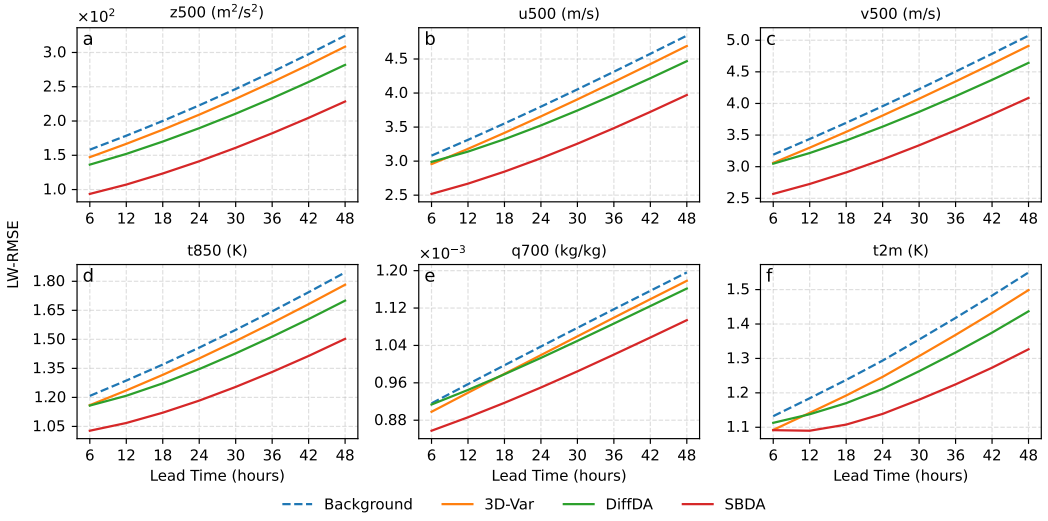


Figure 3: **Latitude-weighted RMSE (LW-RMSE) over forecast lead times up to 48 h.** Panels **a–f** show $z500$, $u500$, $v500$, $t850$, $q700$, and $t2m$, respectively. Forecasts are initialized from the background (blue dashed), 3D-Var (yellow), DiffDA (green), and SBDA (red) analyses.

Comparing Table 1 with Figure 3 reveals that single-step analysis rankings do not always carry over to downstream forecasts. DiffDA can outperform 3D-Var in forecasting despite higher single-step assimilation errors. For $t2m$, SBDA produces the highest single-step LW-RMSE among all evaluated methods, yet its analyses yield the lowest forecast LW-RMSE at every evaluated lead time from 6 to 48 h. These observations suggest that, for deep learning weather foundation models, *a variable's single-step analysis accuracy alone does not determine its downstream forecast skill*, possibly because improvements in other dynamical variables also benefit its subsequent evolution.

Table 3: **Zero-shot single-step assimilation at resampled observation locations.** Subscripts indicate sample standard deviations across five sets of observation locations. Δ_{bg} indicates the averaged relative LW-RMSE change (%) against the background. Units follow Table 1.

		Training locations			Resampled locations		
Variable	Background	3D-Var	DiffDA	SBDA	3D-Var	DiffDA	SBDA
$z500$	139.74	131.68	128.48	**84.01**	$131.76_{0.480}$	$128.45_{0.513}$	$\mathbf{96.34}_{0.984}$
$u500$	2.86	2.77	2.95	**2.40**	$2.77_{0.004}$	$2.95_{0.006}$	$\mathbf{2.51}_{0.009}$
$v500$	2.96	2.86	3.00	**2.45**	$2.86_{0.004}$	$3.00_{0.004}$	$\mathbf{2.57}_{0.008}$
$t850$	1.13	1.09	1.18	**1.03**	$1.09_{0.001}$	$1.18_{0.001}$	$\mathbf{1.05}_{0.002}$
$q700$	0.87	0.86	0.91	**0.83**	$0.86_{0.001}$	$0.91_{0.001}$	$\mathbf{0.84}_{0.001}$
$t2m$	1.08	**1.04**	1.13	1.21	$\mathbf{1.04}_{0.001}$	$1.13_{0.002}$	$1.23_{0.004}$
$u10$	1.47	1.42	1.52	**1.26**	$1.42_{0.003}$	$1.52_{0.004}$	$\mathbf{1.31}_{0.007}$
$v10$	1.52	1.48	1.57	**1.30**	$1.48_{0.003}$	$1.57_{0.004}$	$\mathbf{1.36}_{0.008}$
msl	147.47	139.09	136.18	**99.09**	$139.19_{0.518}$	$136.49_{0.743}$	$\mathbf{111.53}_{1.134}$
Δ_{bg} (%)	–	−3.74	−0.52	**−16.63**	$-3.69_{0.140}$	$-0.45_{0.195}$	$\mathbf{-13.29}_{0.296}$

4.4 Zero-shot Generalization to Different Observation Locations

To assess generalization to different observation locations, we evaluate the same SBDA checkpoint on five new sets of observation locations sampled using different random seeds, without retraining or fine-tuning. Each set contains 288 grid points (1%) and remains fixed throughout evaluation.

Across the five sets of resampled observation locations, SBDA retains a mean relative LW-RMSE reduction of 13.29% against the background and 10.05% against 3D-Var evaluated at the same locations (Table 3). For each set of resampled observation locations, SBDA improves upon the

Table 4: Single-step assimilation ablations and diffusion controls. Δ is the mean relative LW-RMSE change (%) against FULL across all 69 variables. Units follow Table 1.

Method	$z500$	$u500$	$v500$	$t850$	$q700$	$t2m$	$u10$	$v10$	msl	Δ (%)
FULL	**84.01**	**2.40**	**2.45**	**1.03**	**0.83**	1.21	**1.26**	**1.30**	**99.09**	0.00
w/o I	89.20	2.44	2.49	1.05	0.83	1.23	1.27	1.31	102.80	+2.84
w/o T	112.13	2.67	2.76	1.07	0.85	1.23	1.41	1.47	126.18	+11.14
w/o TI	128.21	2.81	2.90	1.13	0.87	1.28	1.46	1.51	140.05	+18.88
DiffDA	128.48	2.95	3.00	1.18	0.91	1.13	1.52	1.57	136.18	+21.00
DiffDA-OI	128.17	2.95	3.01	1.18	0.91	**1.13**	1.52	1.57	135.97	+21.00

background on 67 of 69 variables and upon 3D-Var on 64–65 variables. Compared with evaluation at the training locations, SBDA's mean relative LW-RMSE increase is 4.50%. These results demonstrate that SBDA retains substantial assimilation benefits at unseen observation locations, although accuracy decreases relative to evaluation at the training locations.

4.5 Ablation Study

We compare the full model (FULL) with three variants: w/o T uses raw background states as training endpoints; w/o I removes per-step observation refinement during sampling; and w/o TI removes both components. We also include DiffDA and DiffDA-OI, with DiffDA-OI conditioned on the same OI-enhanced background used by SBDA. All methods are evaluated using single-step LW-RMSE (Table 4).

Observations during training and inference. FULL achieves lower LW-RMSE than each of the three ablated variants across all 69 variables. The larger error increase for w/o T than for w/o I ($\Delta = 11.1\%$ versus 2.8%) suggests that *incorporating observations during training contributes more than per-step observation refinement during inference*. The largest Δ among the ablated variants occurs for w/o TI (18.9%), supporting the joint use of observations during training and sampling.

Background-to-analysis transport. DiffDA-OI yields little overall change from DiffDA, with a mean relative LW-RMSE increase of 0.005% across all 69 variables. SBDA outperforms DiffDA-OI on 68 of 69 variables, with a mean relative LW-RMSE reduction of 16.41%. The 14.14% mean relative reduction retained by w/o I against DiffDA-OI shows that this advantage persists without per-step observation refinement. These results provide empirical support for direct background-to-analysis transport through DSB, beyond supplying the OI-enhanced background as a conditioning input to a conventional diffusion model.

5 Discussion

We introduce SBDA, a generative data assimilation framework that integrates diffusion Schrödinger bridges with classical 3D-Var under a consistent probabilistic view. SBDA learns to generate analyses directly from observation-enhanced backgrounds, with observations informing both bridge learning and per-step refinement during sampling. Experiments demonstrate improved analysis accuracy across most variables and better downstream forecast skill relative to the evaluated baselines, with substantial assimilation gains retained at unseen observation locations. Together, these findings support observation-informed diffusion bridges as a promising approach to providing initial conditions for deep learning weather models.

Although SBDA retains substantial assimilation gains at unseen locations when trained with a fixed observation layout, varying observation locations during training remains an important direction for improving generalization. Future work should also extend SBDA from single-time to 4D assimilation, incorporating observations across a temporal window.

AI USE STATEMENT

Large language models were used to assist with writing and to polish the manuscript for clarity and readability. The authors take responsibility for the final content of this work.

REFERENCES

Rossella Arcucci, Lamya Moutiq, and Yi-Ke Guo. Neural assimilation. In *Computational Science – ICCS 2020*, 2020.

Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data Assimilation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2016.

Feng Bao, Zezhong Zhang, and Guannan Zhang. An ensemble score filter for tracking high-dimensional nonlinear dynamical systems. *Computer Methods in Applied Mechanics and Engineering*, 432:117447, 2024a.

Feng Bao, Zezhong Zhang, and Guannan Zhang. A score-based filter for nonlinear data assimilation. *Journal of Computational Physics*, 514:113207, 2024b.

Peter Bauer. What if? numerical weather prediction at the crossroads. *Journal of the European Meteorological Society*, 1:100002, 2024.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619(7970):533–538, 2023.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, 641(8065):1180–1187, 2025.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In *Advances in Neural Information Processing Systems*, 2021.

Dohyun Bu, Chanho Kim, Seokun Choi, and Jong-Seok Lee. Sparsely-supervised data assimilation via physics-informed Schrödinger bridge. *arXiv preprint arXiv:2603.22319*, 2026.

Alberto Carrassi, Marc Bocquet, Laurent Bertino, and Geir Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *Wiley Interdisciplinary Reviews: Climate Change*, 9(5):e535, 2018.

Kun Chen, Lei Bai, Fenghua Ling, Peng Ye, Tao Chen, Jing-Jia Luo, Hao Chen, Yi Xiao, Kang Chen, Tao Han, et al. Towards an end-to-end artificial intelligence driven global weather forecasting system. *arXiv preprint arXiv:2312.12462*, 2023.

Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.

Yongxin Chen, Tryphon T Georgiou, and Michele Pavon. On the relation between optimal transport and schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

PHILIPPE Courtier, J-N Thépaut, and Anthony Hollingsworth. A strategy for operational implementation of 4d-var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519):1367–1387, 1994.

Geir Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, 99(C5): 10143–10162, 1994.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter Dominik Dueben, and Torsten Hoefler. DiffDA: a diffusion model for weather-scale data assimilation. In *Forty-first International Conference on Machine Learning*, 2024.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Lili Lei, Fuzhong Weng, Wansuo Duan, Yaodeng Chen, Lin Zhang, Ruichun Wang, Jun Yang, Xiaohao Qin, Wei Han, Jun Li, et al. Overview and prospect of data assimilation in numerical weather prediction. *Journal of Meteorological Research*, 39(3):559–592, 2025.

Christian Léonard. From the schrödinger problem to the monge–kantorovich problem. *Journal of Functional Analysis*, 262(4):1879–1920, 2012.

Martin Leutbecher and Tim N Palmer. Ensemble forecasting. *Journal of computational physics*, 227 (7):3515–3539, 2008.

Zeyu Li, Hongkun Dou, Shen Fang, Wang Han, Yue Deng, and Lijun Yang. Physics-aligned field reconstruction with diffusion bridge. In *International Conference on Learning Representations*, volume 2025, pp. 58864–58900, 2025.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

AC Lorenc. A global three-dimensional multivariate statistical interpolation scheme. *Monthly Weather Review*, 109(4):701–721, 1981.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2):130 – 141, 1963.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Peter Manshausen, Yair Cohen, Peter Harrington, Jaideep Pathak, Mike Pritchard, Piyush Garg, Morteza Mardani, Karthik Kashinath, Simon Byrne, and Noah Brenowitz. Generative data assimilation of sparse weather station observations at kilometer scales. *Journal of Advances in Modeling Earth Systems*, 17(10):e2024MS004505, 2025.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, et al. Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024.

E. Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. *Annales de l'institut Henri Poincaré*, 2(4):269–310, 1932.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Jing-An Sun, Hang Fan, Junchao Gong, Ben Fei, Kun Chen, Fenghua Ling, Wenlong Zhang, Wanghan Xu, Li Yan, Pierre Gentine, et al. Lo-sda: Latent optimization for score-based atmospheric data assimilation. *arXiv preprint arXiv:2510.22562*, 2025.

Xiaoze Xu, Xiuyu Sun, Wei Han, Xiaohui Zhong, Lei Chen, Zhiqiu Gao, and Hao Li. Fuxi-da: A generalized deep learning data assimilation framework for assimilating satellite observations. *npj Climate and Atmospheric Science*, 8(1):156, 2025.

A APPENDIX

A.1 DATA PROCESSING

We use ERA5 (ECMWF Reanalysis v5) (Hersbach et al., 2020) as the ground-truth atmospheric state in this study. Due to computational constraints, we do not use the original ERA5 spatial resolution of 0.25°. Instead, we adopt the 1.5° resolution version provided by WeatherBench2 (Rasp et al., 2024), corresponding to a global grid of size 120×240. We use four surface variables, namely 2 meter temperature ($t2m$), 10 meter eastward wind speed ($u10$), 10 meter northward wind speed ($v10$), and mean sea-level pressure (msl), together with five atmospheric variables, geopotential (z), eastward wind speed (u), northward wind speed (v), temperature (t), and specific humidity (q), across 13 pressure levels (50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa), resulting in a total of $4 + 5 \times 13 = 69$ variables. The dataset is sampled at a temporal resolution of 6 hours. Data from 1989 to 2017 are used for training, while data from 2018 are reserved for evaluation.

To simulate sparse observations, we randomly sample 1% of the spatial grid points from the ERA5 dataset. Specifically, the atmospheric state at a single time step is represented as a tensor of size $69 \times 120 \times 240$. Among the $120 \times 240 = 28,800$ spatial grid points, 1% (i.e., 288 grid points) are randomly selected, and all 69 variables at each selected location are retained as observations. Consequently, the sparse observations at a single time step are represented by a tensor of size 69×288. To ensure a fair comparison across different methods, the selected observation locations are fixed throughout both the training and inference stages for all models.

For the background field, we use 48-hour forecasts generated by the deep learning weather foundation model Aurora (Bodnar et al., 2025), initialized from ERA5 states. Since the official Aurora release does not provide a model at the 1.5° resolution used in this study, we fine-tune the official 0.25° pretrained model on the 1.5° ERA5 dataset. Specifically, we adopt the short lead-time fine-tuning strategy described in the Aurora paper with a roll-out step of 2, using ERA5 data from 1979 to 1988 for fine-tuning. The resulting model is then used to perform 48-hour autoregressive forecasts for ERA5 data from 1989 to 2018, and these forecasts serve as the background fields for data assimilation. Among them, data from 1989 to 2017 are used for training, while data from 2018 are reserved for evaluation.

Before training, each of the 69 variable-level channels is standardized using the mean and standard deviation computed from the training set of ERA5; the same statistics are used at inference, and predictions are transformed back to physical units for evaluation.

A.2 MODEL TRAINING AND INFERENCE

Algorithm 1 Training of SBDA

Require: ground truth and background pairs $\{(\mathbf{x}^*_{(i)}, \mathbf{x}^b_{(i)})\}_i$, observation matrix $\mathbf{H}$, precomputed Kalman gain $\mathbf{K}$, neural network ϵ_θ, diffusion schedule β_t.

repeat

Sample $\mathbf{x}^*, \mathbf{x}^b \sim \mathcal{U}(\{(\mathbf{x}^*_{(i)}, \mathbf{x}^b_{(i)})\}_i)$

$\mathbf{y} \leftarrow \mathbf{H}\mathbf{x}^*$

$\hat{\mathbf{x}}^b \leftarrow \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^b)$

Sample $t \sim \mathcal{U}([0, 1])$

Sample $\mathbf{x}_t \sim q(\mathbf{x}_t \mid \mathbf{x}^*, \hat{\mathbf{x}}^b)$ according to Equation (4)

$\sigma_t^2 \leftarrow \int_0^t \beta_\tau d\tau$

$\mathcal{L}_\theta \leftarrow \|\epsilon_\theta(\mathbf{x}_t, t) - \frac{\mathbf{x}_t - \mathbf{x}^*}{\sigma_t}\|_2^2$

Take an optimization step on θ to minimize $\mathcal{L}_\theta$

until convergence

The training and inference algorithms of SBDA are presented in Algorithm 1 and Algorithm 2, respectively. The neural network architecture used in our SBDA framework is not restricted to a specific backbone. In this work, we adopt a simplified U-Net equipped with residual blocks and channel-wise linear attention blocks, following (Luo et al., 2023). For the DSB process, we use the

Algorithm 2 Inference of SBDA

Require: background $\mathbf{x}^b$, observations $\mathbf{y}$, observation matrix $\mathbf{H}$, precomputed Kalman gain $\mathbf{K}$, trained neural network ϵ_θ, diffusion schedule β_t
  $\mathbf{x}_N \leftarrow \mathbf{x}^b + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}^\mathbf{b})$
  $\sigma_t^2 \leftarrow \int_0^t \beta_\tau d\tau$
  for $n = N, \ldots, 1$ **do**
    $\mathbf{x}_0^\epsilon \leftarrow \mathbf{x}_n - \sigma_{t_n}\epsilon_\theta(\mathbf{x}_n, t_n)$
    $\hat{\mathbf{x}}_0^\epsilon \leftarrow \mathbf{x}_0^\epsilon + \mathbf{K}(\mathbf{y} - \mathbf{H}\mathbf{x}_0^\epsilon)$
    if deterministic inference **then**
      $\mathbf{x}_{n-1} \leftarrow \mathbb{E}_q[\mathbf{x}_{n-1} \mid \hat{\mathbf{x}}_0^\epsilon, \mathbf{x}_n]$ according to Equation (4)
    else
      Sample $\mathbf{x}_{n-1} \sim q(\mathbf{x}_{n-1} \mid \hat{\mathbf{x}}_0^\epsilon, \mathbf{x}_n)$ according to Equation (4)
    end if
  end for
  return $\mathbf{x}_0$

mirrored β_t schedule proposed in (Liu et al., 2023), with the minimum and maximum values set to 10^{-4} and 0.3, respectively, and discretize the diffusion process into 1000 time steps.

The model is implemented in PyTorch and trained using data parallelism on 8 NVIDIA RTX 4090 GPUs for a total of 20,000 optimization steps. The global batch size is 8, with one sample processed on each GPU, resulting in approximately 4 hours of training. Optimization is performed using the AdamW optimizer (Loshchilov & Hutter, 2019) with a weight decay of 10^{-4}. The learning rate is linearly warmed up from 10^{-6} to 10^{-4} during the first 1,000 steps and then annealed to 10^{-6} using a cosine annealing schedule for the remainder of training. Following common practice, an exponential moving average (EMA) of the model parameters with a decay factor of 0.99 is maintained throughout training and used for inference. For the implementation of optimal interpolation, the Kalman gain $\mathbf{K}$ is first computed according to Equation (10). Assuming the background error covariance and observation error covariance are $\mathbf{B}_{ij} = \sigma_b^2 \exp(-d_{ij}^2/L^2)$, $\mathbf{R} = \sigma_r^2\mathbf{I} = \alpha^2\sigma_b^2\mathbf{I}$, we further define $\mathbf{C}_{ij} = \exp(-d_{ij}^2/L^2)$. Substituting these expressions into the Kalman gain formulation yields the simplified form:

$$\mathbf{K} = \mathbf{C}\mathbf{H}^\top(\mathbf{H}\mathbf{C}\mathbf{H}^\top + \alpha^2\mathbf{I})^{-1},$$

where the factor σ_b^2 cancels out. Consequently, the implementation depends only on two hyperparameters: the correlation length L and the relative observation error coefficient α. In this work, we set $L = 200\text{km}$ and $\alpha = 0.01$. Thus, the prescribed observation standard deviation is 1% of the prescribed background-error scale, corresponding to a variance ratio of 10^{-4}. OI is applied independently to each of the 69 variable-level channels, using the same covariance matrices and Kalman gain.

The objective function in Equation (5) is implemented using the mean squared error (MSE), where the loss is obtained by uniformly averaging the squared errors over all variables and spatial grid points:

$$\mathcal{L}_\theta = \frac{1}{VHW}\sum_{v=1}^{V}\sum_{h=1}^{H}\sum_{w=1}^{W}\left[\epsilon_\theta(\mathbf{x}_t, t)_{v,h,w} - (\frac{\mathbf{x}_t - \mathbf{x}_0}{\sigma_t})_{v,h,w}\right]^2.$$

In all experiments, we set the number of inference time steps to $N = 10$.

A.3 Baseline Setting

Classical Methods. To evaluate SBDA against conventional data assimilation approaches, we adopt 3D-Var (Lorenc, 1981) as the representative traditional baseline. Because the proposed framework assimilates only the background field and observations at a single time step, without utilizing temporal information, we compare it with 3D-Var under the same setting. In our implementation, 3D-Var is realized using the optimal interpolation formulation described above.

Deep learning-based methods. To evaluate SBDA against existing deep learning-based data assimilation approaches, we adopt DiffDA (Huang et al., 2024) as the representative deep learning

baseline. The original DiffDA implementation employs GraphCast (Lam et al., 2023) as the backbone network for its diffusion model. To ensure a fair comparison, we reproduce DiffDA based on its original paper and publicly available source code while replacing the backbone with the same simplified U-Net architecture used in SBDA, thereby eliminating differences arising from network capacity. The only architectural difference is that DiffDA is formulated as a conditional diffusion model and therefore takes the background field as an additional input, resulting in twice as many input channels in the initial convolution layer as SBDA. All other network configurations are identical to those of SBDA, while the diffusion process during both training and inference follows the settings described in the original DiffDA paper.

A.4 Evaluation Metrics

The primary metric used to evaluate both analysis and forecasting performance is the latitude-weighted root mean squared error (LW-RMSE). Let the predicted atmospheric field and the corresponding ground-truth field be denoted by $\hat{\mathbf{x}}, \mathbf{x} \in \mathbb{R}^{H \times W}$, respectively. The LW-RMSE is defined as

$$\text{LW-RMSE}(\hat{\mathbf{x}}, \mathbf{x}) = \sqrt{\sum_{h=1}^{H} \sum_{w=1}^{W} \frac{1}{W} \frac{\cos \phi_h}{\sum_{i=1}^{H} \cos \phi_i} (\hat{x}_{h,w} - x_{h,w})^2},$$

where H and W are the numbers of latitude and longitude grid points, respectively, and ϕ_h is the latitude of the h-th row.

The relative change in LW-RMSE of method A with respect to method B is

$$\Delta_{\text{rel}}(A; B) = \frac{\text{LW-RMSE}_A - \text{LW-RMSE}_B}{\text{LW-RMSE}_B},$$

where negative values indicate error reductions relative to method B.

For each variable, we compute the latitude-weighted continuous ranked probability score (CRPS) at each evaluation time as

$$\text{CRPS} = \sum_{h=1}^{H} \sum_{w=1}^{W} \frac{1}{W} \frac{\cos \phi_h}{\sum_{i=1}^{H} \cos \phi_i} \int_{-\infty}^{\infty} \left[F_{h,w}(\xi) - \mathbf{1}\{x_{h,w} \le \xi\} \right]^2 \, \mathrm{d}\xi,$$

where $F_{h,w}$ is the predictive cumulative distribution function at grid point (h, w), and $\mathbf{1}\{\cdot\}$ is the indicator function. Lower CRPS indicates better performance. For deterministic predictions treated as point-mass distributions, this reduces to the latitude-weighted mean absolute error. Reported CRPS values are averaged over all evaluation times.

The CRPS skill score (CRPSS) measures the relative improvement of the stochastic SBDA ensemble over deterministic SBDA inference:

$$\text{CRPSS} = 1 - \frac{\text{CRPS}_{\text{ens}}}{\text{CRPS}_{\text{det}}},$$

where CRPS_{ens} and CRPS_{det} are the latitude-weighted CRPS defined above for the stochastic ensemble and deterministic prediction, respectively, at the same evaluation time. For each variable, we compute CRPSS at each evaluation time and then average over time. CRPSS is reported as a percentage, with positive values indicating improvement over the deterministic reference.

Following Bi et al. (2023), we estimate the forecast time gain by matching the 48-h LW-RMSE of SBDA to the lead time at which the strongest baseline (DiffDA) attains the same error. Because forecasts are evaluated at 6-h intervals, the matching baseline lead time is estimated by linear interpolation between adjacent points, and the gain is defined as the difference from 48 h.

A.5 Additional Results

Single-step Assimilation across All Variables. Figure 4 shows relative changes in LW-RMSE against the background across all 69 variables. SBDA reduces errors for 68 variables, with a mean relative reduction of 16.6%; $t2m$ is the only variable showing degradation.

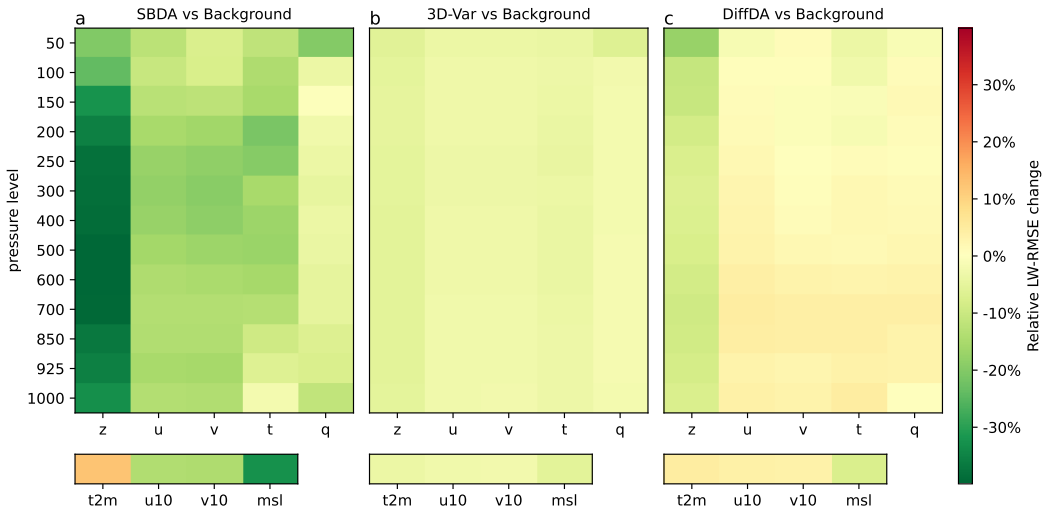


Figure 4: **Relative changes in latitude-weighted RMSE (LW-RMSE). a–c**, SBDA, 3D-Var, and DiffDA analyses relative to the background, respectively. Upper and lower panels show atmospheric and surface variables.

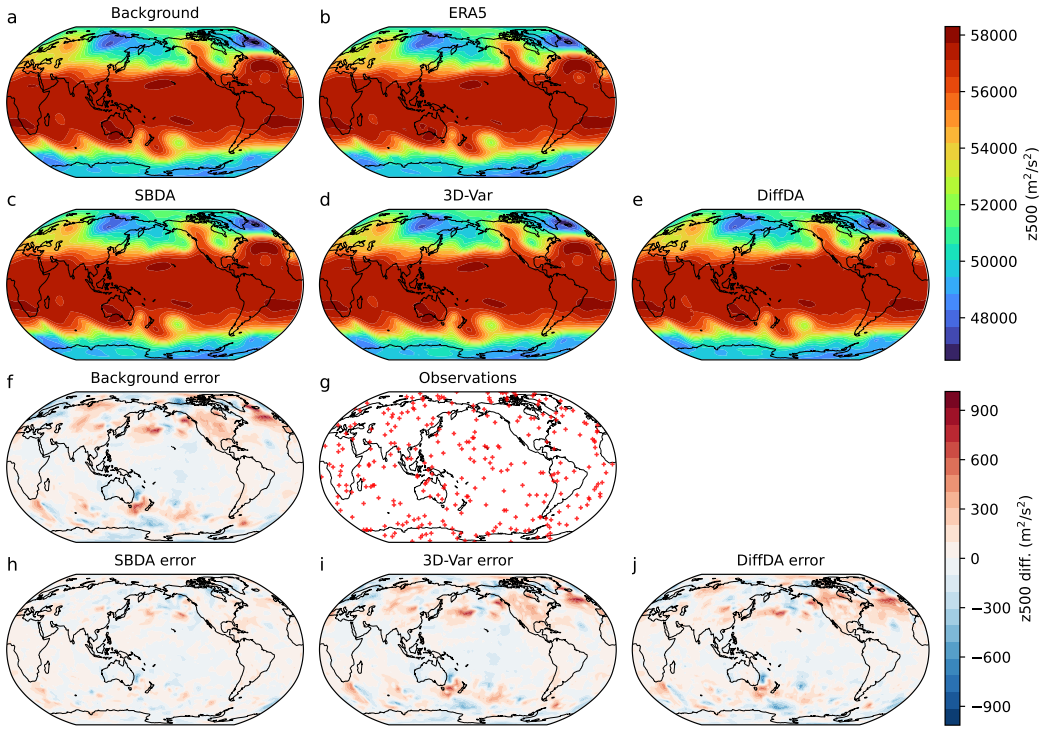


Figure 5: **Geopotential at 500 hPa** ($z500$) **at 12:00 UTC on 15 January 2018.** Background (**a**), ERA5 (**b**), analyses from SBDA (**c**), 3D-Var (**d**), and DiffDA (**e**), background error (**f**), observation locations (**g**, red crosses), and corresponding analysis errors (**h–j**). Errors are computed as ERA5 − Field.

Error Spatial Distributions. Aggregated LW-RMSE does not reveal the spatial structure of assimilation corrections. Figure 5 therefore compares global $z500$ fields at 12:00 UTC on 15 January 2018. The background contains pronounced large-scale errors (Figure 5-f). Although 3D-Var and DiffDA reduce errors near observation locations, their corrections remain localized, including over

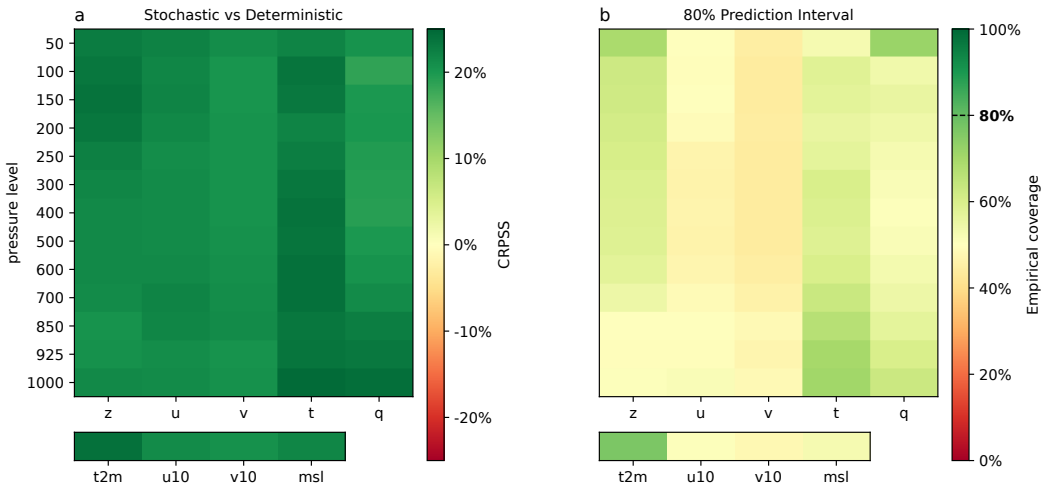


Figure 6: **Probabilistic performance across all 69 variables.** **a**, CRPSS of the 20-member stochastic ensemble relative to deterministic inference. **b**, empirical coverage of the central 80% prediction intervals; the dashed line marks nominal coverage. Upper and lower panels show atmospheric and surface variables.

North America and the Tasman Sea (Figure 5-i,j). By contrast, SBDA produces broader and more spatially coherent error reductions in these regions, including areas far from observations (Figure 5-h). This case therefore suggests that SBDA can use sparse observations to support large-scale error correction beyond their immediate vicinity.

Probabilistic Analysis across All Variables. As shown in Figure 6, the 20-member stochastic ensemble achieves positive CRPSS across all 69 variables, ranging from 18.48% to 24.54% relative to deterministic inference. However, empirical coverage of the central 80% prediction intervals ranges from 43.75% to 76.44%, indicating undercoverage across all variables.

Forecast Skill across All Variables. Figure 10 shows that SBDA achieves the largest forecast LW-RMSE reductions relative to the background across all 69 variables at every evaluated lead time from 6 to 48 h. Relative to forecasts initialized from the background, SBDA achieves mean LW-RMSE reductions of 18.8% at 6 h and 17.6% at 48 h.

Zero-shot Generalization across All Variables. The gains at unseen observation locations persist across all five resampled layouts: SBDA reduces LW-RMSE relative to the background for 67 of 69 variables in every layout (Figure 9). The two exceptions, $t2m$ and $q150$, are consistent across layouts, while the sample standard deviation of SBDA's relative changes remains below 0.85 percentage points for every variable.

Ablation Results across All Variables. Figure 7 shows relative changes in LW-RMSE for the three ablated variants against FULL. All three variants increase errors across all 69 variables, with mean relative increases of 11.1%, 2.8%, and 18.9% for w/o T, w/o I, and w/o TI, respectively.

Diffusion Controls across All Variables. The negligible aggregate change from DiffDA to DiffDA-OI is accompanied by small variable-wise changes, ranging from -1.01% to 0.80% (Figure 8). FULL and w/o I outperform DiffDA-OI on 68 and 67 variables, respectively, supporting the benefit of background-to-analysis transport even without per-step observation refinement.

Data Availability

The ERA5 data used in this study were obtained from WeatherBench2, which provides the Copernicus Climate Change Service (C3S) ERA5 reanalysis for 1959–2023 regridded to a 1.5° equiangular grid at 6-h intervals (gs://weatherbench2/datasets/era5/1959-2023_01_10-6h-240x121_equiangular_with_poles_conservative.zarr). The 48-h Aurora background fields were generated from these data using our fine-tuned 1.5°-resolution Aurora model.

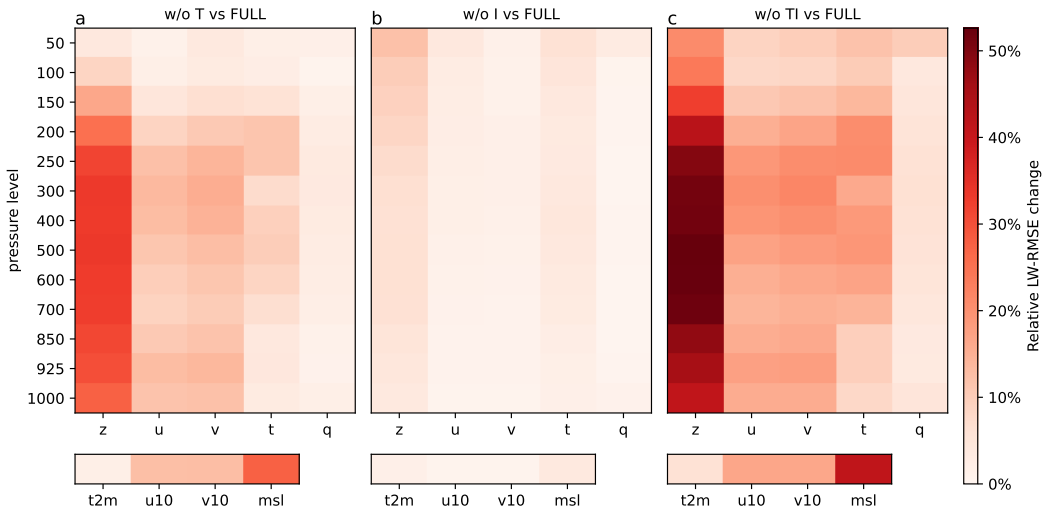


Figure 7: **Relative changes in latitude-weighted RMSE (LW-RMSE) in the ablation study. a–c**, w/o T, w/o I, and w/o TI relative to FULL, respectively. Upper and lower panels show atmospheric and surface variables.

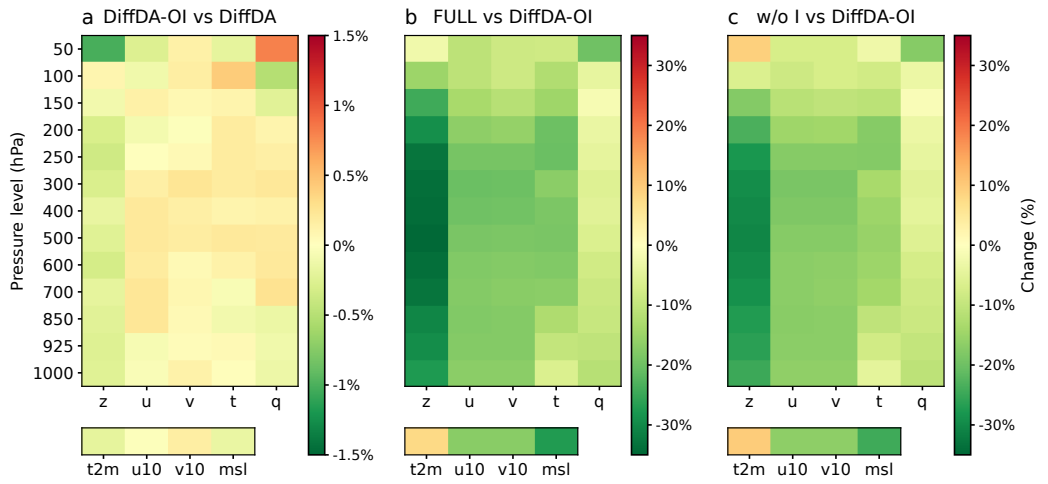


Figure 8: **Diffusion controls across all 69 variables.** Relative LW-RMSE changes for **a**, DiffDA-OI against DiffDA; **b**, FULL against DiffDA-OI; and **c**, w/o I against DiffDA-OI. Negative values indicate lower errors. Panel **a** uses a narrower color scale; panels **b,c** share the same scale. Upper and lower panels show atmospheric and surface variables.

Code Availability

The source code used for model training, evaluation and reproduction of the results reported in this study, including the environment specifications and commands for generating the 48-h Aurora background fields, is available in an anonymized repository hosted by 4open.science (https://anonymous.4open.science/r/SBDA/). The trained SBDA model weights and the fine-tuned 1.5°-resolution Aurora model weights are publicly available on Hugging Face (https://huggingface.co/strayee/SBDA).

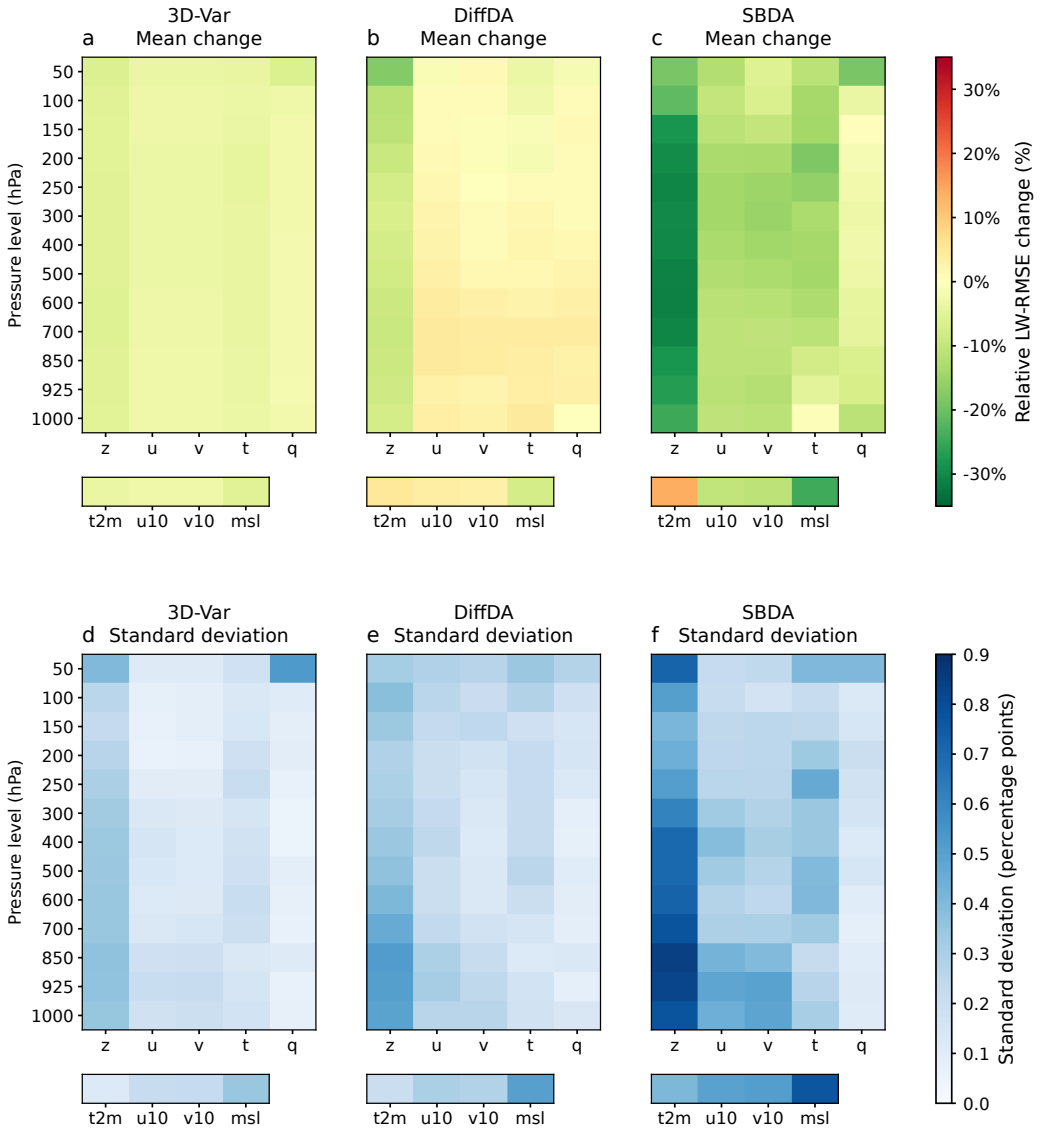


Figure 9: **Zero-shot generalization across all 69 variables. a–c**, mean relative LW-RMSE changes against the background for 3D-Var, DiffDA, and SBDA across five resampled observation layouts. **d–f**, corresponding sample standard deviations, expressed in percentage points. The background reference is identical to that in Table 3. Each panel shows atmospheric variables above and surface variables below. Negative mean changes indicate lower errors.

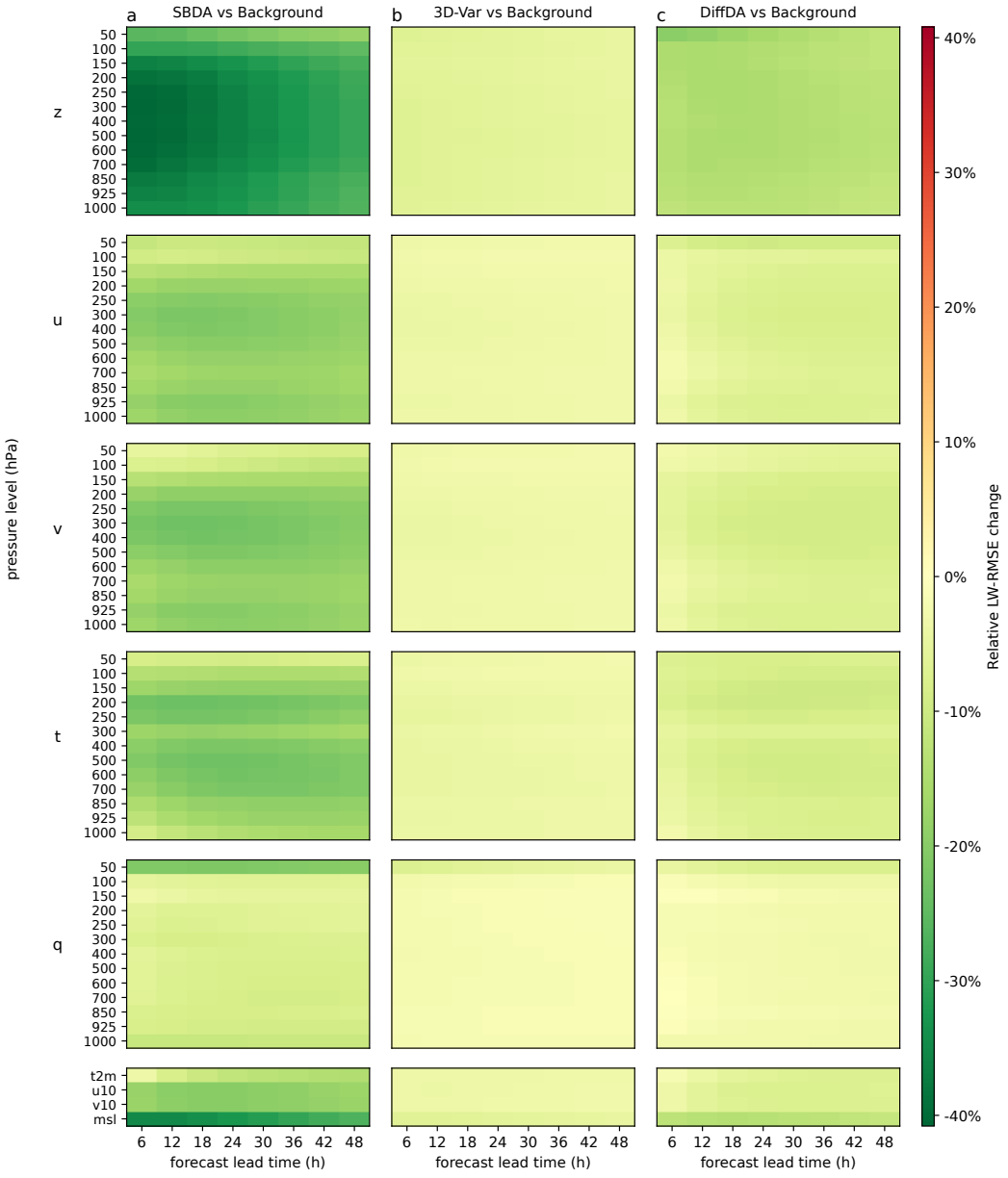


Figure 10: **Relative changes in forecast LW-RMSE across all 69 variables. a–c**, forecasts initialized from SBDA, 3D-Var, and DiffDA analyses relative to forecasts initialized from the background, respectively. Negative values indicate lower errors.