

# Deep Generative Data Assimilation in Multimodal Setting

Yongquan Qu\*  
Columbia University  
yq2340@columbia.edu

Juan Nathaniel\*  
Columbia University  
jn2808@columbia.edu

Shuolin Li  
Columbia University  
s15487@columbia.edu

Pierre Gentine  
Columbia University  
pg2328@columbia.edu

## Abstract

Robust integration of physical knowledge and data is key to improve computational simulations, such as Earth system models. Data assimilation is crucial for achieving this goal because it provides a systematic framework to calibrate model outputs with observations, which can include remote sensing imagery and ground station measurements, with uncertainty quantification. Conventional methods, including Kalman filters and variational approaches, inherently rely on simplifying linear and Gaussian assumptions, and can be computationally expensive. Nevertheless, with the rapid adoption of data-driven methods in many areas of computational sciences, we see the potential of emulating traditional data assimilation with deep learning, especially generative models. In particular, the diffusion-based probabilistic framework has large overlaps with data assimilation principles: both allows for conditional generation of samples with a Bayesian inverse framework. These models have shown remarkable success in text-conditioned image generation or image-controlled video synthesis. Likewise, one can frame data assimilation as observation-conditioned state calibration. In this work, we propose *SLAMS*: Score-based Latent Assimilation in Multimodal Setting. Specifically, we assimilate in-situ weather station data and ex-situ satellite imagery to calibrate the vertical temperature profiles, globally. Through extensive ablation, we demonstrate that *SLAMS* is robust even in low-resolution, noisy, and sparse data settings. To our knowledge, our work is the first to apply deep generative framework for multimodal data assimilation using real-world datasets; an important step for building robust computational simulators, including the next-generation Earth system models. Our code is available at: <https://github.com/yongquan-qu/SLAMS>.

\*Equal contribution; order decided by `np.random.rand(1)`

## 1. Introduction

Data assimilation (DA) is crucial for numerous scientific disciplines that require accurate modeling of chaotic systems. It is particularly important across various areas of the geosciences, including fluid dynamics simulations, oceanography, and atmospheric science [8]. At its core, the state estimation problem of DA seeks to calibrate background trajectory of states,  $x_{1:T}^b = (x_1^b, \dots, x_T^b) \in \mathbb{R}^{T \times N_x}$ , given available observations,  $y$ , which are often noisy, sparse, and/or multimodal. The calibrated state after assimilation, denoted as  $x_{1:T}^a$ , is referred to as the analysis (see Figure 1).

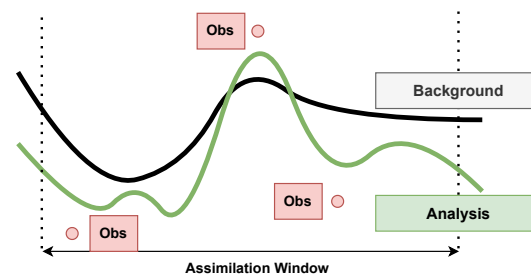


Figure 1. Schematic diagram illustrating the high-level concept of DA. It provides a systematic framework that calibrates background states (model outputs) with multimodal observations (often low-resolution, noisy, and sparse) to produce analysis.

Conventional DA methods often rely on simplifying assumptions (linearity and Gaussianity) or approximations (tangent linear and adjoint) to formulate closed-form expressions for analysis, such as in Kalman filter-based techniques, or to create a manageable optimization target, as in variational approaches. In many applications, however, the system’s underlying dynamics are nonlinear and challenging to encapsulate within a tractable state-space model [44], leading to performance degradation. This necessitates more frequent calibration through assimilation, leading to

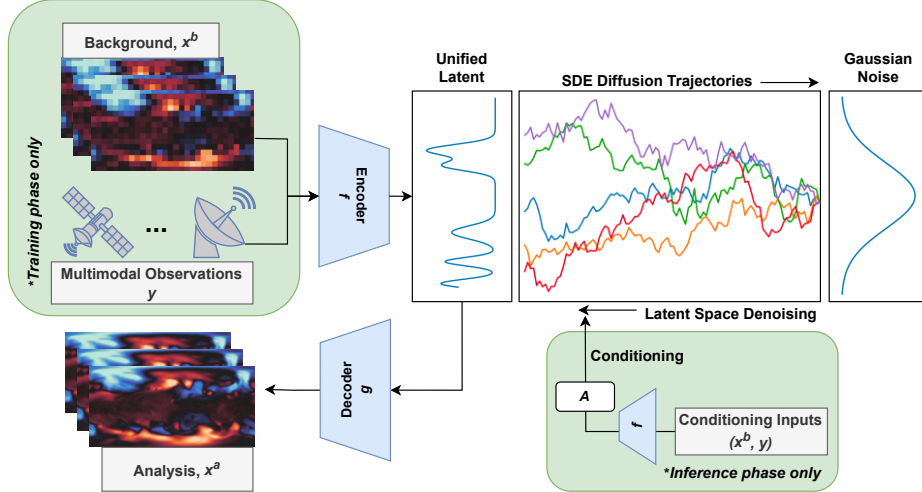


Figure 2. We propose SLAMS - Score-based Latent Assimilation in Multimodal Setting. During training, we fit three set of models, including encoder, decoder, and the score network given multimodal data sources (e.g., background states, in-situ sensors, and ex-situ satellite measurements). During inference, we perform latent space denoising through a reverse SDE process by sampling from prior Gaussian distribution and conditioning on encoded background and observations, synthetically coarsified, noisified, and sparsified by a differentiable measurement function  $\mathcal{A}$  for ablation purposes.

increased computation cost.

With the recent advances of deep learning for Earth system modeling [24, 43], the natural progression is the integration of these data-driven principles into DA for improved state estimation [7, 15], unresolved scale (parameterization) inference [5, 6, 6, 41, 42], and parameter inference [10], either separately or in combination [42]. In particular, the rapid adoption of controllable deep generative modeling in text-conditioned image generation [4, 48] or image-controlled video synthesis [38] could pave the way toward a robust data-driven DA framework. Indeed, the central tenet is similar: we can use observations to condition the generation of analysis. There has been a growing interest to reframe DA through the lens of a deep probabilistic framework [1, 3, 11, 33, 40], particularly with diffusion as a powerful class of generative model capable of fine-grained controlled generation of high-quality samples [19, 47].

However, previous works on data-driven DA used simplified assimilation scenarios: they tend to only assimilate perturbed states as *psuedo-observations* through filtering, noise addition, or sampling [19, 47], ignoring actual observations. Observational datasets in real-world are *proxies* without direct link to the target states. For instance, operational forecasting models may assimilate diverse data sources like weather station outputs, satellite imagery, and LiDAR-derived point clouds [25]. Moreover, these approaches tend to rely on computer vision architectures that favor uniform resolution with limited multi-scale processing capabilities, neglecting rich observations with *heterogeneous* spatiotemporal resolutions [29, 30, 53].

To address these challenges, we introduce SLAMS - Score-based Latent Assimilation in Multimodal Setting (Figure 2). Fundamentally, SLAMS performs its conditional generative process within a unified latent space. This approach allows for the projection of heterogeneous, multimodal datasets into a common  $\mathcal{L}$ -dimensional, latent subspace alongside the target states. As a result, it eliminates the cumbersome, assumption-heavy observation operator,  $\mathcal{H} : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_y}$ , commonly used in traditional DA setup to reconcile heterogeneous spaces [11, 40]. Furthermore, the probabilistic underpinning of SLAMS allows us to generate an ensemble of analysis which naturally facilitates uncertainty quantification. In this study, we adapt score-based latent diffusion model [45, 50] to calibrate real-world weather states using multimodal observation. Specifically, we assimilate both *in-situ* weather station data and *ex-situ* satellite observations to refine vertical temperature profile in the atmosphere as a proof-of-concept. Our ablation demonstrates that SLAMS produces consistent analysis states, even in low-resolution, noisy, and sparse data settings.

## 2. Methodology

Data assimilation at its core attempts to estimate the conditional probability density function of  $p(x_{1:T} | y)$  through a Bayesian inverse formulation as defined in Equation 1,

$$p(x_{1:T} | y) = \frac{p(y | x_{1:T})}{p(y)} p(x_1) \underbrace{\prod_{t=1}^{T-1} p(x_{t+1} | x_t)}_{\text{Markovian background prior}}. \quad (1)$$

We elaborate how the right-hand side of Equation 1 could be reconstructed through the scaffolding of score-based diffusion framework. In particular, we discuss our latent version of the model (SLAMS) that enables multimodal assimilation. Finally, we discuss the datasets, experimental setups, and ablation used throughout this work.

## 2.1. Multimodality in Unified Latent Space

Many multimodal observations,  $Y^{1:K} = (y^1 \in \mathbb{R}^{T \times M_1}, \dots, y^K \in \mathbb{R}^{T \times M_K})$ , such as cloud cover, vegetation, and radiation tend to be indirectly linked as proxies to the target states, such as vertical temperature. For one, capturing these relationships is non-trivial because  $x$  and  $Y$  may have different dimensionality and occupy non-overlapping manifolds which render generative models operating directly in the pixel space less effective. Even if we have access to a mapping function, such as the observation operator that resolves the state-observation heterogeneous spaces,  $\mathcal{H} : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{M_1, \dots, M_K}$ , our imperfect knowledge can exponentiate error especially with high-dimensional  $K$ , and further calibrations to correct biases are required [26].

Fortunately, with deep learning, we are able to circumvent this problem by learning a sufficiently expressive function that is able to project both the states and observations into a unified, reduced-order  $\mathcal{L}$ -dimensional latent space. In this work, we define  $f_\theta : \mathbb{R}^{(N_x + M_1 + \dots + M_K)} \rightarrow \mathbb{R}^{\mathcal{L}}$  as an encoder that takes as input the concatenation of background state and each observation modality along unique channels for a single timestep. The decoder  $g_\phi : \mathbb{R}^{\mathcal{L}} \rightarrow \mathbb{R}^{N_x}$  then maps the latent code back into the pixel-space.

In the following section, we review previous work on score-based diffusion model and its novel application to DA introduced by [47]. In our extension, we define variables  $x$  and  $y$  to represent the latent representations of the unconditional and conditioning inputs, respectively, including *in-situ* and *ex-situ* observations to align with Equation 1.

## 2.2. Score-based Data Assimilation in Latent Space

**Forward diffusion.** In the forward diffusion process, a sample  $x(t) \sim p(x)$  is progressively perturbed through a continuous-time diffusion process expressed as linear stochastic differential equation (SDE) as in Equation 2 [50],

$$dx(t) = \underbrace{f(t)x(t)dt}_{\text{drift term}} + \underbrace{g(t)dw(t)}_{\text{diffusion term}}, \quad (2)$$

where  $f(t), g(t) \in \mathbb{R}$  are the drift and diffusion coefficients,  $w(t) \in \mathbb{R}^{\mathcal{L}}$  the standard Brownian motion (Wiener process), and  $x(t) \in \mathbb{R}^{\mathcal{L}}$  the perturbed sample at time  $t \in [0, 1]$ . Since the source of randomness from the Wiener process is Gaussian, a linear SDE w.r.t.  $x(t)$  ensures that the perturbation kernel is also Gaussian with the form  $p(x(t) | x) = \mathcal{N}(\mu(t)x, \Sigma(t))$  [49]. As a result,  $\mu(t)$  and  $\Sigma(t)$  can be derived analytically from  $f(t), g(t)$  [49].

Furthermore, in a standard diffusion process, we want our sample distribution at  $t = 0$  to be as close to the underlying data distribution  $p(x)$  as possible, such that  $p(x(0)) \approx \mathcal{N}(x, 0)$ . This means that  $\mu(0) = 1$  and  $\Sigma(0) \ll 1$ . On the other hand, we want our sample distribution at  $t = 1$  to be as close to Gaussian noise as possible, such that  $p(x(1)) \approx \mathcal{N}(0, \Sigma(1))$ , implying  $\mu(1) = 0$ . There are multiple SDEs that satisfy these boundary and evolution constraints, including variance-exploding (VE) and variance-preserving (VP) SDEs [50].

**Reverse denoising.** The reverse denoising process is represented by a reverse SDE as defined in Equation 3 [50]. Notice that the drift and diffusion terms are similar to those found in the forward SDE, and that the only difference is the *score*. If we have access to the true score, we can perfectly recover the original data samples from noise as  $t$  from 1 to 0. Intuitively, the score guides the reverse process in the direction where the probability of observing  $x(t)$  is highest.

$$dx(t) = \underbrace{[f(t)x(t) - g(t)^2 \nabla_{x(t)} \log p(x(t))]}_{\text{drift term}} dt + \underbrace{g(t)dw(t)}_{\text{diffusion term}}. \quad (3)$$

**Training score function.** In practice, the score is approximated with a neural network called the score network,  $s_\phi(x(t), t)$ . The objective function would then be to minimize a continuous weighted combination of Fisher divergences between  $s_\phi(x(t), t)$  and  $\nabla_{x(t)} \log p(x(t))$  through score matching [50, 52]. However, the perturbed data distribution  $p(x(t))$  is complex, data-dependent, and hence unscalable. As such, [52] reformulated the objective function by replacing the score with the perturbation kernel where we have access to the analytical form as discussed earlier (Equation 4; expectation is taken over  $p(x), p(t)$ , and  $p(x(t) | x)$ ).

$$\arg \min_{\phi} \mathbb{E} \left[ \Sigma(t) \|s_\phi(x(t), t) - \nabla_{x(t)} \log p(x(t) | x)\|_2^2 \right]. \quad (4)$$

Several studies have noted the instability of this objective function as  $t \rightarrow 0$ , so [54] suggests a reparameterization trick which replaces  $s_\phi(x(t), t) = \epsilon_\phi(x(t), t)/\sigma(t)$ , where  $\Sigma = \sigma^2$  (Equation 5; expectation is taken over  $p(x), p(t), p(\epsilon) \sim \mathcal{N}(0, I)$ ).

$$\arg \min_{\phi} \mathbb{E} \left[ \|\epsilon_\phi(\mu(t)x + \sigma(t)\epsilon, t) - \epsilon\|_2^2 \right]. \quad (5)$$

Following convention used by previous works on score-based diffusion model, we will denote  $\epsilon_\phi(x(t), t)$  with  $s_\phi(x(t), t)$  for cleaner notation.

**Conditioning the generative process.** The case we have discussed so far is the unconditional generative process as we try to sample  $x \sim p(x(0))$  from a prior noise distribution. In order to condition the generative process with observations  $y$ , we seek to sample from  $x \sim p(x(0) | y)$ . This can be done by modifying the score as in Equation 3 with  $\nabla_{x(t)} \log p(x(t) | y)$  and plugging it back to the reverse SDE process.

However, one would need re-training whenever the observation process,  $p(y | x)$ , changes. Nonetheless, several works have attempted to approximate the conditional score with just a single pre-trained score network, eliminating the need for expensive re-training [12, 50].

First, using Bayes rule, we expand the conditional score as in Equation 6,

$$\begin{aligned} \nabla_{x(t)} \log p(x(t) | y) &= \underbrace{\nabla_{x(t)} \log p(x(t))}_{\text{unconditional score}} \\ &+ \underbrace{\nabla_{x(t)} \log p(y | x(t))}_{\text{likelihood score}}. \end{aligned} \quad (6)$$

Since the first term on the right-hand side is already approximated by the score function, the remaining task is to identify the second likelihood score. Assuming differentiable measurement function  $\mathcal{A}$ , and a Gaussian observation process in the latent space,  $p(y | x(t)) = \mathcal{N}(\mathcal{A}(x(t)), \sigma_y^2)$ , the approximation goes as in Equation 7 [12],

$$\begin{aligned} p(y|x(t)) &= \int p(y|x)p(x|x(t))dx \\ &\simeq \mathcal{N}(y | \mathcal{A}(\hat{x}(x(t))), \sigma_y^2), \end{aligned} \quad (7)$$

where  $\hat{x}(x(t))$  can be approximated by the Tweedie’s formula [13] as in Equation 8,

$$\hat{x}(x(t)) \simeq \frac{x(t) + \sigma^2(t)s_\phi(x(t), t)}{x(t)}. \quad (8)$$

The implication is that we can now approximate  $\nabla_{x(t)} \log p(x(t) | y)$  with just a singly trained score network, allowing for a zero-shot assimilation where we do not require constant finetuning when distributions shift.

### 2.3. Scalability and Numerical Stability

We also implemented several performance and stability improvements as proposed in [47]. For brevity, we leave out the theoretical details and refer interested reader to [47].

These improvements include using a subset of embedded trajectories  $\{x_i, x_{i+1}, \dots, x_{i+K-1}\} \in x_{1:T}$  to approximate the entire sequence during training, termed as a Markov blanket (throughout this work, we set  $K = 5$ ). Practically, this strategy helps to reduce the computation cost in cases where the sequence length grows prohibitively

large, as is commonly found in several Earth system processes. Also, we reparameterize Equation 7 to that in Equation 9 to improve the numerical instability of  $\hat{x}(x(t))$  due to the division by  $x(t)$  in Equation 8,

$$p(y|x(t)) \simeq \mathcal{N}(y | \mathcal{A}(\hat{x}(x(t))), \sigma_y^2 + \underbrace{\gamma \frac{\sigma^2(t)}{\mu^2(t)}}_{\text{stability term}} I), \quad (9)$$

where  $\gamma, I$  are scalar constant and identity matrix respectively. This re-parameterization trick reduces numerical instability especially at the beginning of the reverse denoising process where  $\mathbb{E}[x(t)] \rightarrow 0$  as  $t \rightarrow 1$ . Intuitively, the stability term ensures that the likelihood approximation is adjusted according to the noise-signal ratio (i.e.,  $\sigma^2/\mu^2$ ): a noisier state,  $x(t)$ , should result in a more diffused likelihood approximation, and vice versa.

Last, we implement a predictor-corrector procedure to improve the quality of our conditional generative process. The reverse SDE *prediction* process is solved using the exponential integrator (EI) discretization scheme [54] as in Equation 10, and the *correction* uses a few steps of Langevin Monte Carlo (LMC) as in Equation 11,

$$\begin{aligned} x(t - \Delta t) &\leftarrow \frac{\mu(t - \Delta t)}{\mu(t)} x(t) \\ &+ \left( \frac{\mu(t - \Delta t)}{\mu(t)} - \frac{\sigma(t - \Delta t)}{\sigma(t)} \right) \Sigma(t) s_\phi(x(t), t), \end{aligned} \quad (10)$$

$$x(t) \leftarrow x(t) + \tau s_\phi(x(t), t) + \sqrt{2\tau} \epsilon. \quad (11)$$

The correction process is important because it corrects sample generation using information from the gradient of  $\log p(x(t))$  (approximated by  $s_\phi(x(t), t)$ ), which guides the samples towards regions of higher probability, and adds randomness modulated by the Langevin amplitude step  $\tau$ .

## 3. Experimental Setup

### 3.1. Datasets

A standard DA framework generally attempts to calibrate *model states* with available *observations*. The former typically comes from a more homogeneous model output with intrinsic uncertainty, while the latter is more heterogeneous in terms of its modality, resolution, and quality. For this work, we consider two major observation modalities: *in-situ* weather stations and *ex-situ* satellite imagery.

- *Model states.* We use ERA5 reanalysis as our approximation to the true states [18]. ERA5 reanalysis was produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) that recalibrates their historical forecasts or hindcasts with available observations using state-of-the-art DA method. We

select temperature across 10 pressure levels, such that  $t \in \{10, 50, 100, 200, 300, 500, 700, 850, 925, 1000\}$ -hpa, following specifications suggested by [37].

- *In-situ observations.* We use global rain gauges data provided by the Climate Prediction Center (CPC) which has daily resolution and dates back to 1979 [9]. Precipitation is an important phenomenon that captures complex interaction on land, ocean, and atmosphere [31, 32] and which is challenging to model by Earth system models.
- *Ex-situ observations.* We use outgoing longwave radiation (OLR) data collected on-board the NOAA-14, NOAA-16, and NOAA-18 satellites [27]. OLR is an important parameter that is closely related to the planetary energy budget where temperature profile is coupled [31].

We resample all datasets to a 1.5-degree grid at daily resolution. We divide the years 2000-2015 for training, 2016-2021 for validation, and 2022 for testing. We refer to (i) *unimodal* and (ii) *multimodal* as models that incorporate only background state or also take into account *in-situ* and *ex-situ* observations for conditioning. The measurement function  $\mathcal{A}$  is assumed to be differentiable. If this is not the case, a deep-learning emulation of the physical process can be used so that it would be differentiable. Non-latent approaches (i.e., pixel-based) are by construction *unimodal* due to the absence of  $\mathcal{H}$  to reconcile the heterogeneous state-observation spaces discussed earlier.

### 3.2. Details on Autoencoder

We use a convolution-based encoder-decoder structure with 5 layers, ReLU activation [35], kernel size of 5, stride of 1, and a batch normalization [21]. We train over 64 epochs, optimized with ADAMW [28], with a batch size of 64, and follows a linearly decaying scheduler with an initial learning rate of  $2 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-3}$ .

### 3.3. Details on Score Network

We use U-Net [46] with residual blocks [17], SiLU activation [14], and layer normalization [2] for our score network. In particular, the network consists of 3 layers consisting of [3, 3, 3] residual blocks. We specify the hidden channels to be [64, 128, 256] and [32, 64, 128] for either the pixel-based or latent-based score network, with the latter approach enjoying a 8x reduction in model size. Furthermore, we use a circular padding to ensure continuous representation of a spherical globe [22]. We train over 64 epochs, optimized with ADAMW [28], with a batch size of 64, and follows a linearly decaying scheduler with an initial learning rate of  $2 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-3}$ . Finally, we introduce a two-layer linear network to represent the temporal component of the score network with a hidden and embedding size of 256 and 64 respectively.

During inference, we apply 512 denoising steps with 1-step LMC correction,  $\gamma = 10^{-2}$ , and  $\tau = 0.5$ .

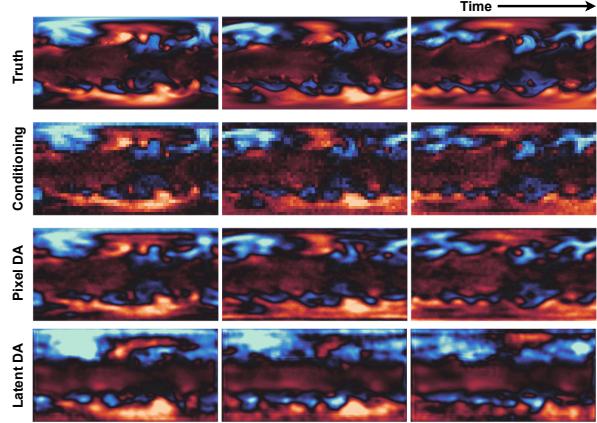


Figure 3. Ideal case where we have **high resolution, low noise, dense inputs** (4x coarsening,  $\sigma^2 = 0.1$ ) to calibrate  $t@200hpa$ . We find that pixel-based DA generates qualitatively better assimilated states.

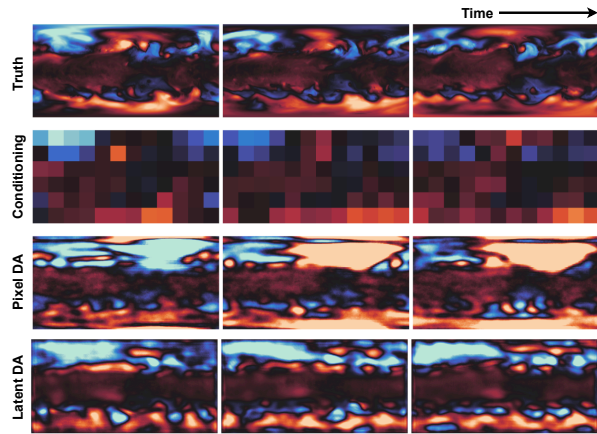


Figure 4. Realistic case where we have **low-resolution inputs** (20x coarsening) to calibrate  $t@200hpa$ . Our latent-based DA approach, SLAMS, is physically more consistent.

## 4. Results and Discussion

In this section, we present an extensive ablation study evaluating the consistency of our proposed latent-based framework, SLAMS, when compared to its pixel-based counterpart. We divide our discussion into the (i) ideal case and (ii) realistic scenario of data assimilation, depending on the quality of conditioning inputs. In addition, we perform multimodal feature importance study to understand the contribution of different observation modality.

### 4.1. Ideal Case: Assimilating High Quality Data

We first begin our discussion in an idealized case where we are given high-resolution, low noise, and dense inputs. We mimic this scenario by applying a coarsening factor of 4, additive Gaussian noise with  $\sigma^2 = 0.1$ , and no masking.

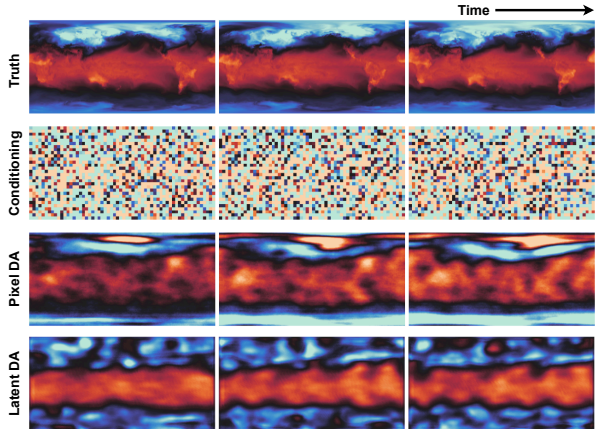


Figure 5. Realistic case where we have **noisy inputs** ( $\sigma^2 = 4$ ) to calibrate  $t@1000hpa$ . Our latent-based DA approach, SLAMS, is physically more consistent.

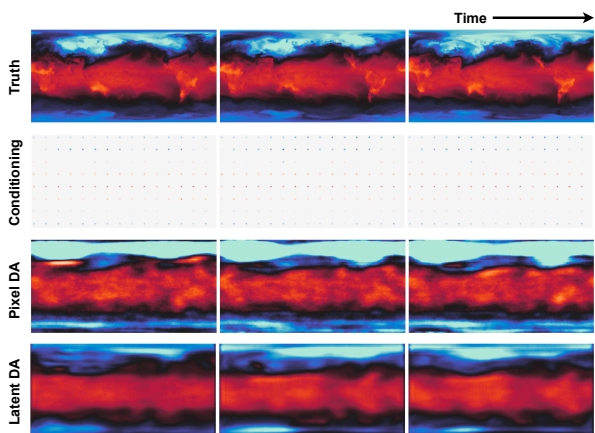


Figure 6. Realistic case where we have **sparse inputs** (16x sampling gap) to calibrate  $t@1000hpa$ . Our latent-based DA approach, SLAMS, is physically more consistent.

As illustrated in Figure 3, we find that pixel-based DA approach generates qualitatively better assimilated states when compared to our latent-based approach. This is akin to introducing additional small noises into the reverse SDE process, and a well-trained score network is robust to denoise such slight perturbation.

## 4.2. Realistic Case: Assimilating Low Quality Data

Unfortunately, many real-world scenarios tend to provide us with (i) low-resolution, (ii) noisy, and (iii) sparse data. We attempt to replicate these scenarios by applying extreme coarsening factor of 20, additive Gaussian noise of  $\sigma^2 = 4.0$ , or sampling gap of 16, respectively.

**Low-resolution data.** As illustrated in Figure 4, we find that our latent-based DA approach tends to be temporally and physically consistent especially in high latitudes where

we do not observe unrealistic abrupt cooling  $\rightarrow$  warming changes, which are otherwise observed in the pixel-based DA approach.

**Noisy data.** As illustrated in Figure 5, we find that our latent-based DA approach produces more realistic continuous pattern in the tropics. Compared to the pixel-based approach, SLAMS also generates samples with more accurate physics, with no unrealistic large-scale near-surface warming in the polar region.

**Sparse data.** As illustrated in Figure 6, we find that our latent-based DA approach does not inaccurately overestimate near-surface temperature in the high-latitude, and generate realistic, smoother calibration in the tropics.

These results suggest the stabilizing capability of our latent-based approach, SLAMS, that can be partly explained by: (i) reduced dimensionality of the latent code that makes it less sensitive to high-dimensional, pixel-level perturbation [55], (ii) stabilizing characteristic of a well-trained decoder exhibiting posterior non-collapse [45], (iii) additional constraints enforced by increasing data modality [16].

## 4.3. Toward Consistent Multimodal Assimilation

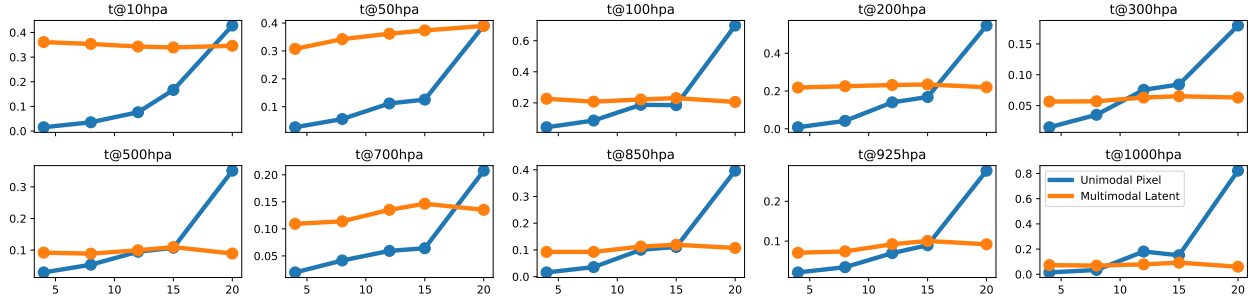
In order to fully showcase the robustness of our latent-based approach, SLAMS, we conducted an in-depth ablation study that varied the coarsening factor, noise level, and sampling gap given (i) *unimodal* (only background), or (ii) *multimodal* (background + observations) conditioning inputs. Specifically, we employed the Wasserstein distance [34] to quantify the distribution divergence between the true and assimilated states. Full results are illustrated in Figure 7.

Overall, our ablation study indicates that SLAMS maintains its robustness (i.e., exhibits lower Wasserstein distances) across the majority of target states, particularly when assimilating inputs that are of low resolution (coarsening factors  $\geq 15$ ), highly noisy ( $\sigma^2 \geq 2$ ), and sparse (sampling gap  $\geq 12$ ).

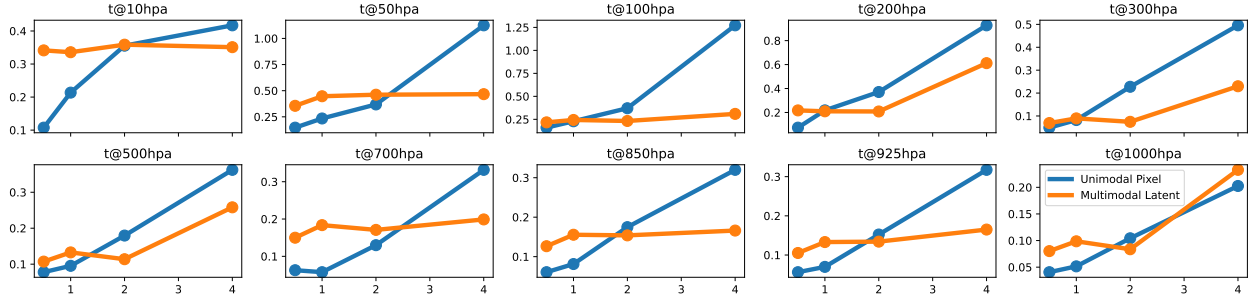
First, an operationally useful grid resolution for Earth system model typically falls below 0.1 degrees [39]. However, many datasets such as those used in our study are generally an order of magnitude coarser, approximately 1.5 to 2.5 degrees. This discrepancy can lead to instability and sensitivity to perturbations in pixel-based DA approaches.

Likewise, the challenge of noisy observations, compounded by model uncertainty, is a widespread issue. Traditional DA systems must account for these through expensive error propagation calculations, which can become intractable in multimodal, high-dimensional scenarios [23]. Thus, a robust DA system, underpinned by a deep probabilistic framework like ours, has to be robust to noise.

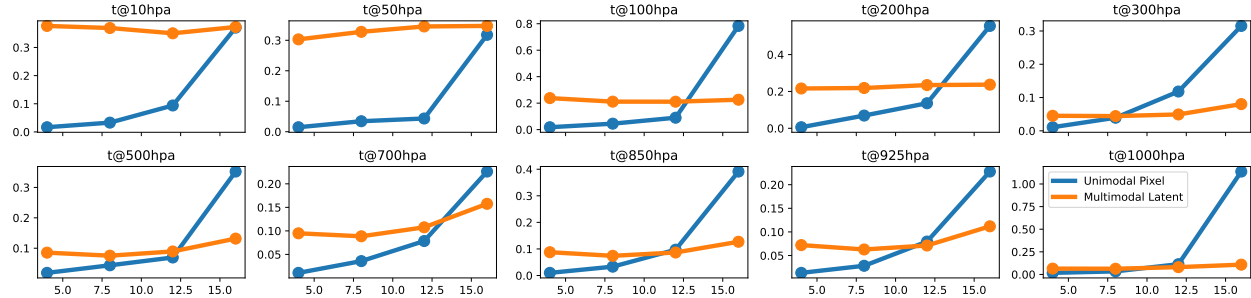
Lastly, data sparsity is a common challenge in DA for Earth system modeling. Sparsity can come in many forms, including spatial (swath width, coverage), temporal (orbit period), and/or representativeness (geostationary orbit)



(a) Wasserstein distance against coarsening factors. Our multimodal latent-based DA approach (SLAMS) is consistent even when coarsening factors  $\geq 15$



(b) Wasserstein distance against noise level. Our multimodal latent-based DA approach (SLAMS) is consistent even when  $\sigma^2 \geq 2$



(c) Wasserstein distance against sampling gap. Our multimodal latent-based DA approach (SLAMS) is consistent even when sampling gap  $\geq 12$

Figure 7. Ablation result evaluating the effects of varying (a) resolution, (b) noise level, and (c) sparsity on Wasserstein distance ( $\downarrow$  is better) of different diffusion-based DA methods. Across most target states, a diffusion-based DA approach operating in latent space (SLAMS) is more stable than that in the pixel space beyond certain threshold.

[36]. Thus, a DA framework that can efficiently extract signals is crucial for robust real-time modeling.

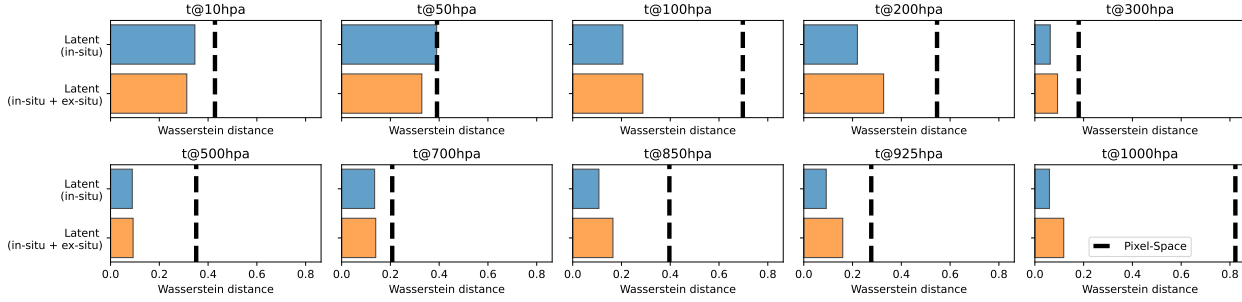
#### 4.4. Multimodal Feature Ablation

So far, we have demonstrated the feasibility of performing DA in a multimodal setting through a unified representational space. The natural next step is to assess the impact of multimodal observations in improving the quality of analysis states. Figure 8 reaffirms our earlier findings that our latent-based DA approach, SLAMS, outperforms pixel-based DA across most target states. Additionally, the inclusion of ex-situ imagery is particularly valuable for constraining top-of-atmosphere (ToA) variables (e.g.,  $t@10hpa$ ,  $t@50hpa$ ). This finding is consistent with the role of OLR as a proxy for the planetary energy budget [20].

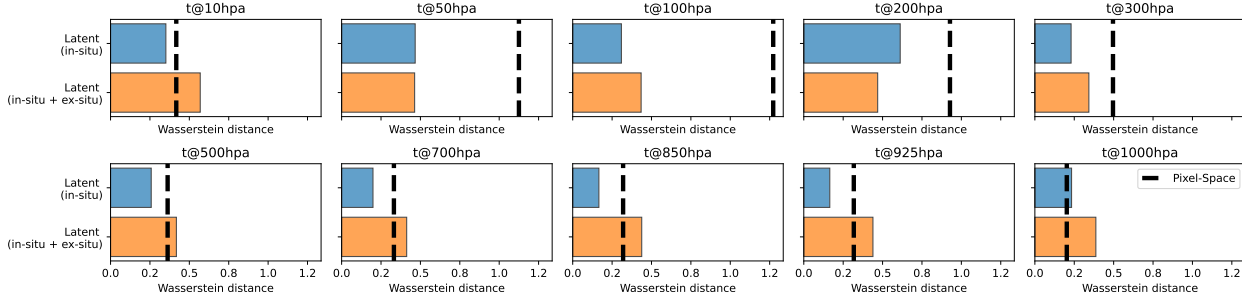
One strategy to maximize the extraction of signals from multimodal observations involves accounting for different latent representations of various data modalities. Future work could look into distinct, more powerful autoencoders (e.g., VQ-VAE [51]) for each modality and introducing an aggregator to boost the expressiveness of each feature [11]. This approach also provides the benefit of making the framework adaptable to different underlying data structures that are not easily gridded as image frames, including LiDAR point clouds, textual, and tabular data.

## 5. Conclusion

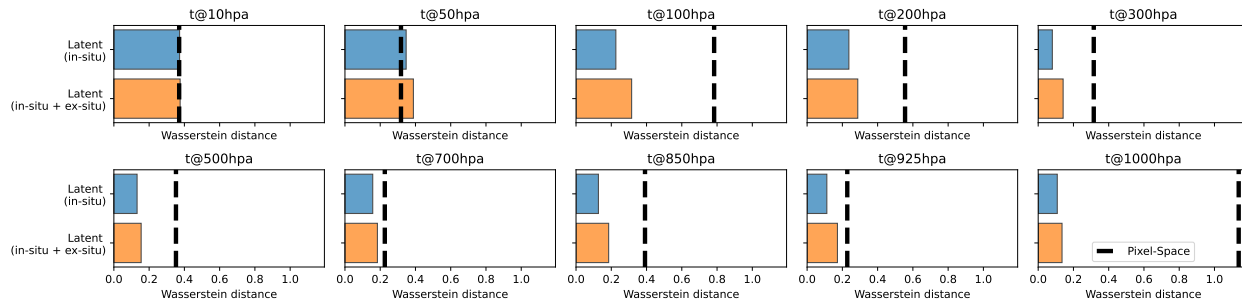
We introduce SLAMS, a score-based latent data assimilation framework that enables multimodality. Our framework re-



(a) Effect of assimilating low-resolution multimodal inputs (20x coarsening factor)



(b) Effect of assimilating noisy multimodal inputs ( $\sigma^2 = 4$ )



(c) Effect of assimilating sparse multimodal inputs (16x sampling gap)

Figure 8. Ablation result on the effect of multimodality in the calibration of target states given (a) low-resolution, (b) noisy, and (c) sparse inputs. The x-axis shows the Wasserstein distance ( $\downarrow$  is better). The results demonstrate that adding ex-situ imagery (orange bar), mostly enhances top-of-atmosphere state calibration (e.g.,  $t@10hpa$ ,  $t@50hpa$ ) compared to only conditioning on in-situ observations (blue bar). Both latent methods (SLAMS) outperform pixel-based approach (vertical dashed line) in most target states.

frames the traditionally compute-intensive DA process using deep probabilistic principles through diffusion as a powerful class of generative model. SLAMS facilitates the assimilation of multimodal data through its latent-centric architecture, capitalizing on recent advancements in dimensionality reduction and automatic feature extraction. The adaptable design of SLAMS also accommodates future integration with data modalities traditionally challenging to represent as image frames, such as point clouds. Moreover, the probabilistic underpinning of our framework allows for ensemble generation of analysis, enabling uncertainty quantification. We validate SLAMS’s efficacy in realistic scenario by assimilating multimodal *in-situ* weather station data and *ex-situ* satellite imagery to calibrate the vertical temperature profile, globally. Extensive ablation con-

firms that SLAMS is robust to low-resolution, noisy, and sparse inputs. To our knowledge, this work is the first to propose a data-driven *probabilistic, multimodal* data assimilation framework for real-world Earth system modeling.

## Acknowledgments

The authors would like to acknowledge funding from the National Science Foundation Science and Technology Center, Learning the Earth with Artificial intelligence and Physics, LEAP (Grant number 2019625), Department of Energy grant.



## References

- [1] Rossella Arcucci, Jiangcheng Zhu, Shuang Hu, and Yi-Ke Guo. Deep data assimilation: integrating deep learning with data assimilation. *Applied Sciences*, 11(3):1114, 2021. 2
- [2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5
- [3] Jichao Bao, Liangping Li, and Arden Davis. Variational autoencoder or generative adversarial networks? a comparison of two deep learning methods for flow and transport data assimilation. *Mathematical Geosciences*, 54(6):1017–1042, 2022. 2
- [4] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 2
- [5] Mohamed Aziz Bhouri and Pierre Gentine. History-based, bayesian, closure for stochastic parameterization: Application to lorenz'96. *arXiv preprint arXiv:2210.14488*, 2022. 2
- [6] Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to infer unresolved scale parametrization. *Philosophical Transactions of the Royal Society A*, 379(2194):20200086, 2021. 2
- [7] Caterina Buizza, César Quilodrán Casas, Philip Nadler, Julian Mack, Stefano Marrone, Zainab Titus, Clémence Le Cornec, Evelyn Heylen, Tolga Dur, Luis Baca Ruiz, et al. Data learning: Integrating data assimilation and machine learning. *Journal of Computational Science*, 58:101525, 2022. 2
- [8] 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. 1
- [9] Mingyue Chen, Wei Shi, Pingping Xie, Viviane BS Silva, Vernon E Kousky, R Wayne Higgins, and John E Janowiak. Assessing objective techniques for gauge-based analyses of global daily precipitation. *Journal of Geophysical Research: Atmospheres*, 113(D4), 2008. 5
- [10] Yuming Chen, Daniel Sanz-Alonso, and Rebecca Willett. Autodifferentiable ensemble kalman filters. *SIAM Journal on Mathematics of Data Science*, 4(2):801–833, 2022. 2
- [11] Sibö Cheng, Jianhua Chen, Charitos Anastasiou, Panagiota Angeli, Omar K Matar, Yi-Ke Guo, Christopher C Pain, and Rossella Arcucci. Generalised latent assimilation in heterogeneous reduced spaces with machine learning surrogate models. *Journal of Scientific Computing*, 94(1):11, 2023. 2, 7
- [12] Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022. 4
- [13] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. 4
- [14] Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. 5
- [15] Alban Farchi, Marcin Chrust, Marc Bocquet, Patrick Laloux, and Massimo Bonavita. Online model error correction with neural networks in the incremental 4d-var framework. *Journal of Advances in Modeling Earth Systems*, 15(9):e2022MS003474, 2023. 2
- [16] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xi-ang, et al. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):3094, 2022. 6
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5
- [18] 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. 4
- [19] Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D Dueben, and Torsten Hoefler. Diffda: a diffusion model for weather-scale data assimilation. *arXiv preprint arXiv:2401.05932*, 2024. 2
- [20] Yi Huang, V Ramaswamy, and Brian Soden. An investigation of the sensitivity of the clear-sky outgoing longwave radiation to atmospheric temperature and water vapor. *Journal of Geophysical Research: Atmospheres*, 112(D5), 2007. 7
- [21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 5
- [22] Md Amirul Islam, Matthew Kowal, Sen Jia, Konstantinos G Derpanis, and Neil DB Bruce. Position, padding and predictions: A deeper look at position information in cnns. *arXiv preprint arXiv:2101.12322*, 2021. 5
- [23] Tijana Janjić, Niels Bormann, Marc Bocquet, JA Carton, Stephen E Cohn, Sarah L Dance, Svetlana N Losa, Nancy K Nichols, Roland Potthast, Joanne A Waller, et al. On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144(713):1257–1278, 2018. 6
- [24] Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Nor-gaard, Jamie Smith, Griffin Mooers, James Lottes, Stephan Rasp, Peter Düben, Milan Klöwer, et al. Neural general circulation models. *arXiv preprint arXiv:2311.07222*, 2023. 2
- [25] William A Lahoz and Philipp Schneider. Data assimilation: making sense of earth observation. *Frontiers in Environmental Science*, 2:16, 2014. 2
- [26] Jianyu Liang, Koji Terasaki, and Takemasa Miyoshi. A machine learning approach to the observation operator for satellite radiance data assimilation. *Journal of the Meteorological Society of Japan. Ser. II*, 101(1):79–95, 2023. 3

- [27] Brant Liebmann and Catherine A Smith. Description of a complete (interpolated) outgoing longwave radiation dataset. *Bulletin of the American Meteorological Society*, 77(6):1275–1277, 1996. [5](#)
- [28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [5](#)
- [29] Gengchen Mai, Chris Cundy, Kristy Choi, Yingjie Hu, Ni Lao, and Stefano Ermon. Towards a foundation model for geospatial artificial intelligence (vision paper). In *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, pages 1–4, 2022. [2](#)
- [30] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, et al. On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*, 2023. [2](#)
- [31] Syukuro Manabe and Robert F Strickler. Thermal equilibrium of the atmosphere with a convective adjustment. *Journal of the Atmospheric Sciences*, 21(4):361–385, 1964. [5](#)
- [32] Syukuro Manabe and Richard T Wetherald. Thermal equilibrium of the atmosphere with a given distribution of relative humidity. 1967. [5](#)
- [33] Syamil Mohd Razak, Atefeh Jahandideh, Ulugbek Djuraev, and Behnam Jafarpour. Deep learning for latent space data assimilation in subsurface flow systems. *SPE Journal*, 27(05):2820–2840, 2022. [2](#)
- [34] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.*, pages 666–704, 1781. [6](#)
- [35] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010. [5](#)
- [36] Juan Nathaniel, Jiangong Liu, and Pierre Gentine. Metaflux: Meta-learning global carbon fluxes from sparse spatiotemporal observations. *Scientific Data*, 10(1):440, 2023. [7](#)
- [37] Juan Nathaniel, Yongquan Qu, Tung Nguyen, Sungduk Yu, Julius Busecke, Aditya Grover, and Pierre Gentine. Chaosbench: A multi-channel, physics-based benchmark for subseasonal-to-seasonal climate prediction. *arXiv preprint arXiv:2402.00712*, 2024. [5](#)
- [38] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023. [2](#)
- [39] Brian C O’Neill, Claudia Tebaldi, Detlef P Van Vuuren, Veronika Eyring, Pierre Friedlingstein, George Hurtt, Reto Knutti, Elmar Kriegler, Jean-Francois Lamarque, Jason Lowe, et al. The scenario model intercomparison project (scenariomip) for cmip6. *Geoscientific Model Development*, 9(9):3461–3482, 2016. [6](#)
- [40] Mathis Peyron, Anthony Fillion, Selime Gürol, Victor Marchais, Serge Gratton, Pierre Boudier, and Gael Goret. Latent space data assimilation by using deep learning. *Quarterly Journal of the Royal Meteorological Society*, 147(740):3759–3777, 2021. [2](#)
- [41] Yongquan Qu and Xiaoming Shi. Can a machine learning-enabled numerical model help extend effective forecast range through consistently trained subgrid-scale models? *Artificial Intelligence for the Earth Systems*, 2(1):e220050, 2023. [2](#)
- [42] Yongquan Qu, Mohamed Aziz Bhouri, and Pierre Gentine. Joint parameter and parameterization inference with uncertainty quantification through differentiable programming. *arXiv preprint arXiv:2403.02215*, 2024. [2](#)
- [43] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and fmm Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. [2](#)
- [44] Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adria Lopez Escoriza, Ruud JG Van Sloun, and Yonina C Eldar. Kalmannet: Neural network aided kalman filtering for partially known dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022. [1](#)
- [45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [2, 6](#)
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. [5](#)
- [47] François Rozet and Gilles Louppe. Score-based data assimilation. *arXiv preprint arXiv:2306.10574*, 2023. [2, 3, 4](#)
- [48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. [2](#)
- [49] Simo Särkkä and Arno Solin. *Applied stochastic differential equations*. Cambridge University Press, 2019. [3](#)
- [50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. [2, 3, 4](#)
- [51] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. [7](#)
- [52] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. [3](#)
- [53] Liang Xue and Dongxiao Zhang. A multimodel data assimilation framework via the ensemble kalman filter. *Water Resources Research*, 50(5):4197–4219, 2014. [2](#)
- [54] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. [3, 4](#)
- [55] Zhendong Zhao, Xiaojun Chen, Yuexin Xuan, Ye Dong, Dakui Wang, and Kaitai Liang. Defeat: Deep hidden feature

backdoor attacks by imperceptible perturbation and latent representation constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15213–15222, 2022. [6](#)