Appa: Bending Weather Dynamics with Latent Diffusion Models for Global Data Assimilation

Gérôme Andry **Sacha Lewin** **François Rozet** **Omer Rochman** **Victor Mangeleer**

Matthias Pirlet **Elise Faulx** **Marilaure Grégoire** **Gilles Louppe**

University of Liège

Abstract

Deep learning has advanced weather forecasting [1–8], but accurate predictions first require identifying the current state of the atmosphere from observational data. In this work, we introduce Appa, a score-based data assimilation model generating global atmospheric trajectories at 0.25° resolution and 1-hour intervals. Powered by a 565M-parameter latent diffusion model trained on ERA5, Appa can be conditioned on arbitrary observations to infer plausible trajectories, without retraining. Our probabilistic framework handles reanalysis, filtering, and forecasting, within a single model, producing physically consistent reconstructions from various inputs. Results establish latent score-based data assimilation as a promising foundation for future global atmospheric modeling systems.

1 Introduction

Data assimilation combines observational data with physical models to estimate atmospheric states. Formally, let $x^{1:L} = (x^1, x^2, \dots, x^L) \in \mathbb{R}^{L \times V \times C}$ denote a trajectory of L atmospheric states, each represented as C physical fields over a mesh of V vertices. Let $p(x^1)$ be the initial state prior and $p(x^{i+1} \mid x^i)$ the transition dynamics. Observations $y \in \mathbb{R}^M$ of the state trajectory $x^{1:L}$ follow an observation process $p(y \mid x^{1:L})$, generally formulated as $y = \mathcal{M}(x^{1:L}) + \eta$, where the measurement function $\mathcal{M} : \mathbb{R}^{L \times V \times C} \mapsto \mathbb{R}^M$ might be non-linear and $\eta \in \mathbb{R}^M$ represents observational error that accounts for instrumental noise and systematic uncertainties. The goal of data assimilation is to infer plausible trajectories $x^{1:L}$ consistent with the observations, that is, to estimate the trajectory posterior

$$p(x^{1:L} \mid y) = \frac{p(y \mid x^{1:L})}{p(y)} \, p(x^1) \prod_{i=1}^{L-1} p(x^{i+1} \mid x^i) \, . \tag{1}$$

While Eq. (1) defines the posterior inference problem generically, assimilation tasks correspond to specific choices of states x^i and observations y to take into account. In this work, we focus on three practical cases:

$$\textbf{Reanalysis:} \quad x^{1:L} \sim p(x^{1:L} \mid y^{1:L}), \tag{2}$$

$$\textbf{Filtering:} \quad x^{L} \sim p(x^{L} \mid y^{1:L}), \tag{3}$$

$$\textbf{Forecasting:} \quad x^{K+1:L} \sim p(x^{K+1:L} \mid y^{1:K}) \;\text{ or }\; p(x^{K+1:L} \mid x^{1:K}). \tag{4}$$

Reanalysis aims to reconstruct full trajectories from partial historical observations of the same time segment. The primary purpose of reanalysis is to create datasets of historical data for the land, atmospheres, and oceans. These datasets enable scientists to better monitor and understand the climate, conduct surveys, and develop new weather models. Filtering, in contrast, only infers the posterior distribution of the current state, obtained as a marginal of the reanalysis posterior. Forecasting, as its

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

name suggests, extends beyond the observed segment, producing posterior distributions over future states. The forecasting problem is often initialized by an external estimation of the current state x^K or decomposed into first estimating this current state from $y^{1:K}$, and then predicting its evolution.

Traditional methods like 4D-Var [9–13] and ensemble Kalman filters [14] are effective but rely on linearizations, require expensive differentiation, and provide point estimates rather than full posterior distributions [15]. Recent data-driven approaches [16–20] integrate deep learning into assimilation or forecast directly from observations, but suffer from limited resolution, lack of uncertainty quantification, and require retraining for new observation configurations.

2 Appa

Appa combines score-based data assimilation [21–23] with latent diffusion models for physics emulation [24], scaled to the global atmospheric system at 0.25° resolution and 1-hour intervals, with 6 surface variables and 5 atmospheric variables across 13 pressure levels.

Architecture Appa consists of a 340M-parameter encoder-decoder pair (E_ψ, D_ψ) that compresses atmospheric states x^i by a factor 530 into a latent representations $z^i \sim \mathcal{N}(z^i \mid E_\psi(x^i), \sigma_z^2 I)$. This reduces the dimensionality from $\mathcal{O}(10^8)$ elements per atmospheric state ($\mathcal{O}(10^{10})$ for 4 days at 1-hour resolution) to $\mathcal{O}(10^5)$ elements per latent state ($\mathcal{O}(10^7)$ for 4 days), enabling efficient generation and inference in the latent space. The encoder-decoder pair is trained with a latitude- and level-weighted mean squared error loss [3, 4].

The autoencoder is paired with a 225M-parameter diffusion transformer (DiT) [25] that operates on windows of $W = 24$ consecutive latent states. Following [22, 26], we train a denoiser to estimate the denoising posterior mean $\mathbb{E}\left[z^{i:i+W} \mid z_t^{i:i+W}\right]$ which, via Tweedie's first-order formula, provides the prior score function $\nabla_{z_t^{i:i+W}} \log p(z_t^{i:i+W})$ needed for the reverse diffusion sampling process. For a variance exploding diffusion process [27], we have

$$\mathbb{E}\left[z^{i:i+W} \mid z_t^{i:i+W}\right] = z_t^{i:i+W} + \sigma_t^2 \nabla_{z_t^{i:i+W}} \log p(z_t^{i:i+W}). \tag{5}$$

As in the SDA [21, 22] framework, the score over trajectories $z_t^{i:i+L}$ is approximated by composing the scores over windows. We generalize SDA's composition algorithm by introducing a time stride $\Delta \geq 1$ between consecutive windows. Using a larger stride reduces the window overlap and, therefore, the number of network evaluations. Detailed algorithms for training and composing local scores can be found in B.2.

Sampling conditionally on weather observations To sample from the posterior $p(z^{1:L} \mid y)$, we replace the prior score in the reverse diffusion sampling process with the posterior score

$$\nabla_{z_t^{1:L}} \log p(z_t^{1:L} \mid y) = \nabla_{z_t^{1:L}} \log p(z_t^{1:L}) + \nabla_{z_t^{1:L}} \log p(y \mid z_t^{1:L})\,. \tag{6}$$

The prior score $\nabla_{z_t^{1:L}} \log p(z_t^{1:L})$ is obtained from the denoiser and combination of scores over windows, while the likelihood score $\nabla_{z_t^{1:L}} \log p(y \mid z_t^{1:L})$ can be approximated without retraining [21, 28–30] under moderate assumptions on the observation process.

The key challenge is that the observation process $p(y \mid x^{1:L})$ is defined for atmospheric states x^i rather than latent states z^i. For observation sequences $y^{1:L}$ of the form $y^i = \mathcal{M}^i(x^i) + \eta^i$, we approximate the mapping from z^i to y^i as the composition of the decoder D_ψ and measurement operator $\mathcal{M}^i$. Formally,

$$p(y^{1:L} \mid z^{1:L}) \approx \mathcal{N}(y^{1:L} \mid \mathcal{A}(z^{1:L}), \Sigma_y)\,, \tag{7}$$

such that $\mathcal{A}(z^{1:L}) = (\mathcal{M}^1(D_\psi(z^1)) \cdots \mathcal{M}^L(D_\psi(z^L)))^\top$ and Σ_y is the covariance of $\eta^{1:L}$. With this formulation, off-the-shelf posterior sampling algorithms [30] can be used for generating atmospheric trajectories conditionally on observational data. In this work, we adapt moment matching posterior sampling (MMPS), originally proposed by Rozet et al. [29] for linear observation operators. In our case, since $\mathcal{A}$ is non-linear, we use its Jacobian A in the estimate of the covariance, yielding the approximation of the perturbed likelihood

$$p(y^{1:L} \mid z_t^{1:L}) \approx \mathcal{N}(y \mid \mathcal{A}(\mathbb{E}[z^{1:L} \mid z_t^{1:L}]), \Sigma_y + A\mathbb{V}[z^{1:L} \mid z_t^{1:L}]A^\top)\,. \tag{8}$$

After inference in the latent space, the generated trajectories $z^{1:L}$ are decoded back to the atmospheric space via the decoder $\hat{x}_i = D_\psi(z_i)$.

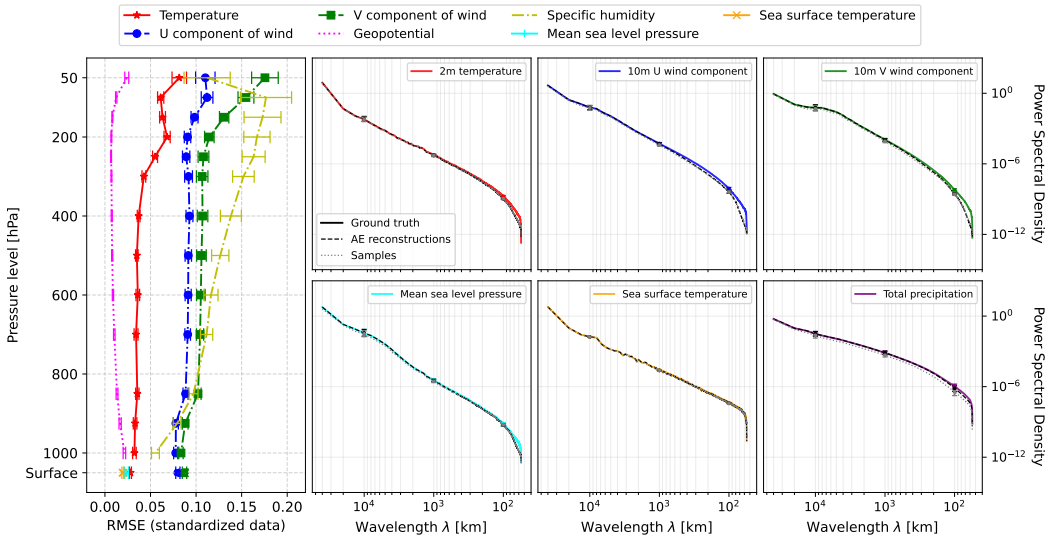

Figure 1. (Left) Standardized autoencoder reconstruction RMSEs. Lower-frequency fields (temperature, geopotential) are reconstructed more accurately than volatile fields (humidity, winds). Near-surface fields benefit from altitude-weighting. (Right) Power spectral density comparison of ground truth, autoencoder reconstructions, and samples generated from Appa's prior. Median and percentile ranges show close alignment across scales, with deviations below 100 km (3 to 4 grid cells).

3 Experiments

We train and evaluate Appa on the ERA5 reanalysis dataset [31] following standard chronological splitting: 1993–2021 for training, 2020 for validation, and 2021 for testing. Below, we evaluate the latent representation quality, assimilation performance across reanalysis, filtering and forecasting tasks, and compare against existing methods.

Latent representation Despite the $530\times$ compression factor, standardized reconstruction RMSEs are mostly below 0.1, with slightly higher values for humidity and winds, and lower ones for surface and low-altitude fields, as shown in Figure 1. Reconstructed power spectra match ground truth closely, except at scales below 100 km where atmospheric energy is minimal. Compared to prior neural compression methods [32], our autoencoder achieves comparable performance.

Assimilation We evaluate Appa across four scenarios: reanalysis, filtering, observational forecasting, and full-state forecasting. For the first three tasks, we assimilate both synthetic ground-station observations of all 6 surface variables and simulated scans of the 5 atmospheric variables across 13 pressure levels. The ground station network consists of 11,000 real-world measurement locations [33] covering roughly 1% of grid points. Ground stations are sparse and globally distributed, while satellite orbital paths provide dense spatial coverage with restricted temporal and spatial reach. Observations are modeled as Gaussian distributions centered on the ERA5 ground truth, with noise levels of 1% for ground stations and 10% for satellite measurements.

Figure 2 summarizes Appa's performance and further qualitative results can be found in D.4. For reanalysis and filtering, conditioning on longer assimilation windows improves both skill and CRPS but gains saturate beyond 24 hours. Forecasting's skill decays gradually with lead time but remains significantly stronger than the persistence baseline. Observational forecasts, conditioned on the last 12 hours of a day-long assimilation, start at skill and CRPS levels comparable to the reanalysis plateau. Full-state forecasts, initialized from two complete states, start lower but eventually converge to similar performance over time. As expected, the initial skill for full-state forecasts is close to autoencoder reconstruction error levels. Overall, these results demonstrate that Appa successfully handles all assimilation and forecasting tasks within a unified probabilistic framework.

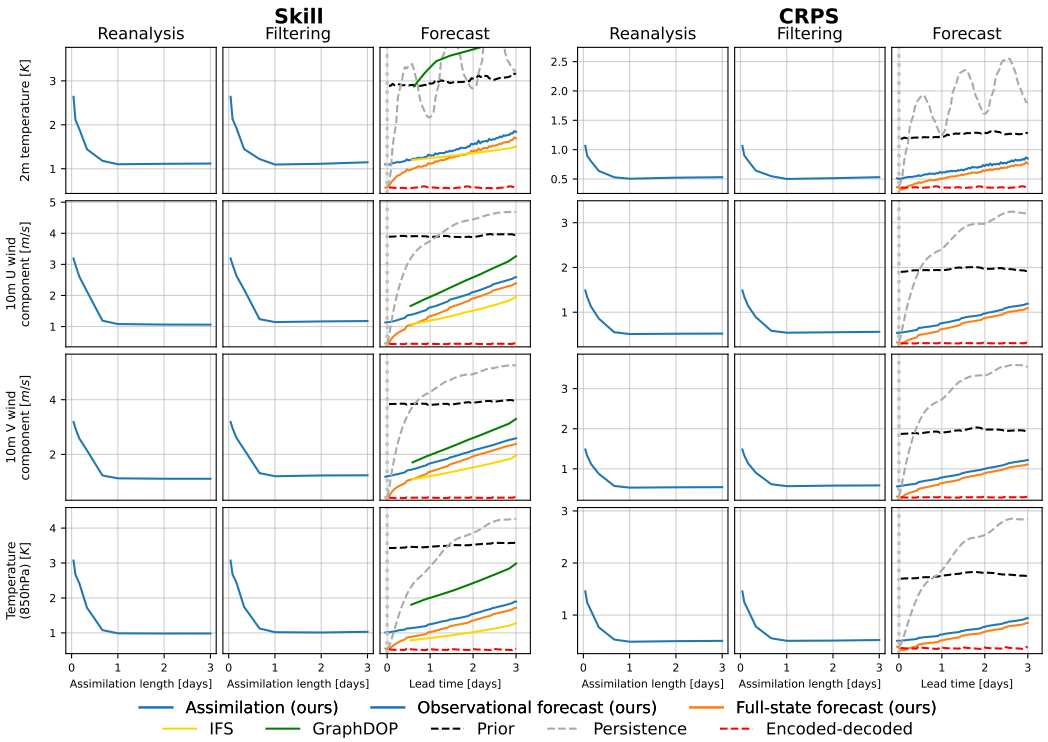

Figure 2. Skill and Continuous Ranked Probability Score (CRPS) for representative variables in January 2023. Detailed experimental setup can be found in B.3. Reanalysis scores are averaged over assimilation windows $x^{1:L}$, while filtering reports the last reanalyzed state. Both improve with longer assimilation windows, but eventually stagnate. Forecasts gradually lose skill over lead time but remain superior to persistence and unconditional baselines. The initial skill level of the full-state forecast matches the compression error level. IFS [34] and GraphDOP [35] are shown for reference.

For short lead times, our forecasts reach skill levels comparable to IFS [34] while performing better than GraphDOP [35]. Notably, the forecast error growth lies between the two baselines: slightly steeper than IFS but below GraphDOP and closely tracking its slope. This suggests that compression does not introduce strong dynamical artifacts that impede the learning of the dynamics, even under different temporal setups (1-hour for Appa vs. 12-hour resolution for IFS) which corroborates with findings of Rozet et al. [24]. While these results remain preliminary, they provide promising evidence that Appa can capture atmospheric dynamics at a level in between purely physical and purely data-driven approaches.

4 Discussion

Summary We introduce Appa, a latent score-based data assimilation framework that produces global atmospheric trajectories by operating in a compressed latent space. Appa can be conditioned on various types of observations without retraining, providing access to the full posterior distribution of consistent trajectories. Our results show that Appa flexibly handles reanalysis, filtering, and forecasting within a single framework, producing competitive performance across scenarios without task-specific training or architectural modifications.

Limitations and future work While Appa demonstrates strong assimilation and forecasting capabilities, it remains a proof of concept and further improvements are needed to make it operational. First, we should consider moving from simplified synthetic observations to realistic measurement, such as satellite radiances. Improving physical consistency is also critical, as compression inevitably degrades fine-scale information. Strategies such as localized assimilation or refined conditioning mechanisms may help. In terms of statistical assessment, the calibration of posterior distributions

deserves further validation. Indeed, the approximations present in our method, notably while estimating the prior and likelihood scores, introduce errors which are complex to quantify. The computational efficiency of observational tasks remains a challenge as well, since conditioning requires repeated decoding steps, projecting observations directly into latent space could mitigate this bottleneck. Finally, our comparison to other models is still preliminary. Baselines for assimilation remain scarce, and fair evaluation against IFS [34], GraphDOP [35] and other models [3, 4, 36] would help position Appa within the spectrum of global atmospheric models.

Acknowledgments and Disclosure of Funding

Gérôme Andry, François Rozet, Sacha Lewin, and Elise Faulx are research fellows of the *National Fund for Scientific Research* (F.R.S.-FNRS) and acknowledge its financial support. Omer Rochman gratefully acknowledges the financial support of the *Walloon Region* under Grant No. 2010235 (ARIAC by Digital Wallonia 4.AI). Victor Mangeleer is a research fellow part of the *Multiple Threats on Ocean Health* (MITHO) project and gratefully acknowledges funding from the *European Space Agency* (ESA).

The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

References

[1] Jaideep Pathak et al. "FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators". 2022.

[2] Lei Chen et al. "FuXi: a cascade machine learning forecasting system for 15-day global weather forecast". In *npj Climate and Atmospheric Science* 6.1 (2023).

[3] Remi Lam et al. "Learning skillful medium-range global weather forecasting". In *Science* 382.6677 (2023).

[4] Ilan Price et al. "Probabilistic weather forecasting with machine learning". In *Nature* 637.8044 (2025).

[5] Simon Lang et al. "AIFS – ECMWF's data-driven forecasting system". 2024.

[6] Cristian Bodnar et al. "A Foundation Model for the Earth System". 2024.

[7] Chong Nan et al. "LangYa: Revolutionizing Cross-Spatiotemporal Ocean Forecasting". 2025.

[8] Surya et al. "Samudra: An AI Global Ocean Emulator for Climate". In arXiv:2412.03795 (2024).

[9] A. C. Lorenc. "Analysis methods for numerical weather prediction". In *Quarterly Journal of the Royal Meteorological Society* 112.474 (1986).

[10] François-Xavier Le Dimet and Olivier Talagrand. "Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects". In *Tellus A* 38A.2 (1986).

[11] Yannick Tr'emolet. "Accounting for an imperfect model in 4D-Var". In *Quarterly Journal of the Royal Meteorological Society* 132.621 (2006).

[12] Yannick Trémolet. "Model-error estimation in 4D-Var". In *Quarterly Journal of the Royal Meteorological Society* 133.626 (2007).

[13] Mike Fisher et al. "Weak-constraint and long-window 4D-Var". In *ECMWF Technical Memoranda* 655 (2011).

[14] Brian R Hunt et al. "Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter". In *Physica D: Nonlinear Phenomena* (2007).

[15] Alberto Carrassi et al. "Data assimilation in the geosciences: An overview of methods, issues, and perspectives". In *WIREs Climate Change* 9.5 (2018).

[16] Sibo Cheng et al. "Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review". In *IEEE/CAA Journal of Automatica Sinica* (2023).

[17] Langwen Huang et al. "DiffDA: a diffusion model for weather-scale data assimilation". In *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. ICML'24. Vienna, Austria: JMLR.org, 2024.

[18] Xiaoze Xu et al. "Fuxi-DA: A Generalized Deep Learning Data Assimilation Framework for Assimilating Satellite Observations". 2024.

[19] Ronan Fablet et al. "Joint interpolation and representation learning for irregularly sampled satellite-derived geophysical fields". In *Frontiers in Applied Mathematics and Statistics* (2021).

[20] Marcin Andrychowicz et al. "Deep Learning for Day Forecasts from Sparse Observations". 2023.

[21] François Rozet and Gilles Louppe. "Score-based data assimilation". In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[22] François Rozet and Gilles Louppe. "Score-based data assimilation for a two-layer quasi-geostrophic model". In *Machine learning and the physical sciences workshop (NeurIPS)*. 2023.

[23] Luca Jonathan et al. "A Generative Framework for Probabilistic, Spatiotemporally Coherent Downscaling of Climate Simulation". In *Journal of Advances in Modeling Earth Systems* arXiv:2412.15361 (2024).

[24] François Rozet et al. "Lost in Latent Space: An Empirical Study of Latent Diffusion Models for Physics Emulation". 2025.

[25] William Peebles and Saining Xie. "Scalable Diffusion Models with Transformers". In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023.

[26] Tero Karras et al. “Elucidating the Design Space of Diffusion-Based Generative Models”. In *Advances in Neural Information Processing Systems* 35 (2022).

[27] Yang Song et al. “Score-Based Generative Modeling through Stochastic Differential Equations”. In *International Conference on Learning Representations*. 2020.

[28] Hyungjin Chung et al. “Diffusion Posterior Sampling for General Noisy Inverse Problems”. In *The Eleventh International Conference on Learning Representations*. 2022.

[29] François Rozet et al. “Learning Diffusion Priors from Observations by Expectation Maximization”. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 2024.

[30] Giannis Daras et al. “A Survey on Diffusion Models for Inverse Problems”. 2024.

[31] H. Hersbach et al. “ERA5 hourly data on single levels from 1940 to present”. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). DOI: 10.24381/cds.adbb2d47, accessed 10-July-2025. 2018.

[32] Piotr Mirowski et al. “Neural Compression of Atmospheric States”. 2024.

[33] NOAA National Centers for Environmental Information. “Global surface summary of the day - GSOD”. Version 1.0. 1999.

[34] ECMWF. “IFS Documentation CY48R1 - Part II: Data Assimilation”. In 2. ECMWF, June 2023.

[35] Mihai Alexe et al. “GraphDOP: Towards skilful data-driven medium-range weather forecasts learnt and initialised directly from observations”. 2024.

[36] Ron Raphaeli et al. “SILO: Solving Inverse Problems with Latent Operators”. arXiv:2501.11746 [cs]. Jan. 2025.

[37] Stephan Rasp et al. “WeatherBench 2: A benchmark for the next generation of data-driven global weather models”. In *Journal of Advances in Modeling Earth Systems* 16.6 (2024), e2023MS004019.

[38] Vincent Fortin et al. “Why should ensemble spread match the RMSE of the ensemble mean?” In *Journal of Hydrometeorology* 15.4 (2014), pp. 1708–1713.

[39] Tilmann Gneiting and Adrian E Raftery. “Strictly proper scoring rules, prediction, and estimation”. In *Journal of the American statistical Association* 102.477 (2007), pp. 359–378.

A Data

A.1 ERA5

ERA5 is a global deterministic reanalysis dataset from ECMWF that provides high-resolution (0.25°) hourly estimates of atmospheric, land, and oceanic variables from 1959 onward [31]. It assimilates observations into a numerical weather prediction model using 4D-Var data assimilation.

For this work, we use a subset of ERA5 data, defined on a 0.25° equiangular grid with 13 pressure levels: 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa. Due to storage limitations, we restrict the temporal coverage of the dataset to the 1993–2021 period, with data split into training (1993–2019), validation (2020), and testing (2021).

Table 1 lists the selected variables. Some serve as both input and predicted features, while others provide contextual information (only input). Context variables are not predicted but help define the temporal and dynamic conditions under which predictions are made, improving model performance.

Table 1. Input variables.

Type	Variable Name	Role
Atmospheric	Temperature	Input/Predicted
Atmospheric	U-Wind Component	Input/Predicted
Atmospheric	V-Wind Component	Input/Predicted
Atmospheric	Geopotential	Input/Predicted
Atmospheric	Specific Humidity	Input/Predicted
Single	2m Temperature	Input/Predicted
Single	10m U-Wind Component	Input/Predicted
Single	10m V-Wind Component	Input/Predicted
Single	Mean Sea Level Pressure	Input/Predicted
Single	Sea Surface Temperature	Input/Predicted
Single	Total Precipitation	Input/Predicted
Clock	Local time of day	Input (Diffusion)
Clock	Elapsed year progress	Input (Diffusion)

A.2 Data pre-processing

Standardization Although the dynamics across the atmospheric column are correlated, each pressure level exhibits distinct statistical behavior. Thus, we computed the mean and standard deviation separately for each variable and at each pressure level, on the whole training dataset. We used these statistics to standardize our entire dataset and to rescale the output of Appa.

Handling missing values Since Sea Surface Temperature is undefined over land (NaN values), we replace these with zeros as a neutral placeholder after standardization.

Data availability ERA5 data was downloaded from the WeatherBench2 platform, where Google has made them publicly available via Google Cloud Storage.

B Technical details

This section provides further technical details for training and inference. Our code will be made available with full reproducibility steps for both training and evaluation.

B.1 Architectures

We adapted architectures from Rozet et al. [24] for both autoencoder and latent diffusion model. The encoder and decoder are fully convolutional neural networks and the diffusion model is adapted from a diffusion transformer (DiT) [25].

Autoencoder The autoencoder compresses atmospheric states from the high-dimensional N320 grid (721×1440 pixels with 71 channels) into a compact latent space (23×47 pixels with 128 channels) via progressive downsampling and channel expansion in a fully convolutional architecture. To accommodate any spatial compression factor, the input is padded to the nearest compatible grid size. We apply periodic padding along longitude to respect global wrap-around, and constant zero padding along latitude to handle polar boundaries. The encoder-decoder pair is trained with a latitude-level-weighted mean squared error loss, following Lam et al. [3].

Latent denoiser The denoiser is a DiT that operates on 24 consecutive latent states. We first patch the latent sequence by a factor of 2 along the temporal axis, then flatten the spatial dimensions, yielding $23 \times 47 \times {}^{24}/_{2} = 12{,}972$ tokens, each with 256 channels, which are passed to the DiT.

Table 2. Autoencoder training configuration

Parameter	**Value**
Loss function	Latitude- and level-weighted mean squared error
Latent noise	$\sigma = 0.01$ for regularization
Optimizer	SOAP with initial learning rate 3×10^{-5} and linear decay
Batch size	64 samples per step
Training duration	95000 update steps (approximately 2 days)
Hardware	64× NVIDIA A100 40GB GPUs

Table 3. Denoiser training configuration

Parameter	**Value**
Loss	Denoising score matching with rectified noise schedule
Noise range	$\sigma_{\text{min}} = 0.001$, $\sigma_{\text{max}} = 1000$
Optimizer	Adam with initial learning rate 1×10^{-4}
Batch size	256 samples per step
Training duration	125000 update steps (approximately 5 days)
Hardware	64× NVIDIA A100 40GB GPUs

B.2 Generating trajectories

Algorithm 1 Training $d_\phi(z_t^{i:i+W})$

for $n = 1$ to N **do**
  $x^{1:L} \sim p(x^{1:L})$
  $i \sim \mathcal{U}(\{1, \dots, L - W\})$
  $t \sim \mathcal{U}(0, 1), \varepsilon \sim \mathcal{N}(0, I)$
  for $j = i$ to $i + W$ **do**
    $z_j \leftarrow E_\psi(x_j)$
  $z_t^{i:i+W} \leftarrow z^{i:i+W} + \sigma_t \, \varepsilon$
  $\ell \leftarrow \left\| d_\phi(z_t^{i:i+W}) - z^{i:i+W} \right\|_2^2$
  $\phi \leftarrow \mathrm{SGD}(\phi, \nabla_\phi \, \ell)$

Algorithm 2 Composing $d_\phi(z_t^{i:i+W})$

function $d_\phi(z_t^{1:L})$
  $a \leftarrow {}^{(W-\Delta)}/_2$
  $b \leftarrow a + \Delta$
  $E_{1:a} \leftarrow d_\phi(z_t^{1:1+W})[:a]$
  for $n = 0$ to ${}^{(L-W)}/_\Delta + 1$ **do**
    $i \leftarrow 1 + n\Delta$
    $E_{i+a:i+b} \leftarrow d_\phi(z_t^{i:i+W})[a:b]$
  $E_{L-W+b:L} \leftarrow d_\phi(z_t^{L-W:L})[b:]$
  return $E_{1:L}$

B.3 Assimilation tasks

Forecasting Appa is trained to generate state windows of a given size, we use 24 hours. We first split the total window in two, the first part being the condition, and the second the part to be generated. Then, we use either the last states of a fully assimilated window (observational forecasting) or full encoded latent states (full-state forecasting). At each autoregressive steps, we use a sliding window to move n steps forward after these steps were predicted. This mechanism offers a balance between conditioning window size and generation speed, as the former can be extended to provide more context but more limited speed (more generations required) or less context but faster total generation.

Evaluation setup On Figure 2, we report IFS [34] and GraphDOP [35] skills using data from Alexe et al. [35] as baselines. However a direct and fair comparison is difficult due to lack of experimental setup details. Alexe et al. [35] mention that skills are computed over January 2023 forecasts for 6 different variables. To match the time period, we report the performance of our method over a 10-member ensemble of 3-day forecasts. We selected the first 8 days of January 2023 at midnight as starting timestamps. This makes a total of 8 dates with 10 members for reported metrics.

C Evaluation metrics

We follow conventional metrics computation for assimilation and forecasting performance. For a fair comparison with the literature, evaluation is performed using WeatherBench2 [37]. For assimilation, we average performance over the time steps.

C.1 Skill

Skill is computed as the root mean square error of the posterior mean of an ensemble compared to the ground-truth trajectory. For K ensembles each consisting of M predicted states $\hat{x}$ of resolution $H \times W$, ground truth x, the skill of a single time step is computed as

$$\text{Skill} = \sqrt{\frac{1}{KHW} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(x_{i,j}^{k} - \frac{1}{M} \sum_{m=1}^{M} \hat{x}_{i,j}^{k,m} \right)^2}.$$

C.2 Spread

Ensemble spread is computed as the square root of the ensemble variance [38]:

$$\text{Spread} = \sqrt{\frac{1}{KHW} \sum_{k=1}^{K} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{x}_{i,j}^{k,m} - \frac{1}{M} \sum_{n=1}^{M} \hat{x}_{i,j}^{k,n} \right)^2}.$$

C.3 Spread-skill ratio

A well-calibrated forecast should have a (corrected for ensemble size) spread-skill ratio of 1, which is a necessary but not sufficient condition. Ratios below one indicate overconfident estimations. The correct ratio is defined as

$$\text{Ratio} = \sqrt{\frac{M+1}{M}} \frac{\text{Spread}}{\text{Skill}}.$$

C.4 Continuous ranked probability score (CRPS)

The CRPS [39] is defined as

$$\text{CRPS} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{1}{M} \sum_{m=1}^{M} ||\hat{x}^{k,m} - x^{k}||_{L_1} - \frac{1}{2M(M-1)} \sum_{m=1}^{M} \sum_{n=1}^{M} ||\hat{x}^{k,m} - \hat{x}^{k,n}||_{L_1} \right).$$

The first term penalizes the average divergence from the ground truth while the second term encourages spread. Therefore, the CRPS is lowest when the distribution of the ensemble matches the ground-truth distribution. Note the L_1 norm used, which means that in the deterministic case, this reduces to the MAE.

D Additional results

D.1 Power spectral density

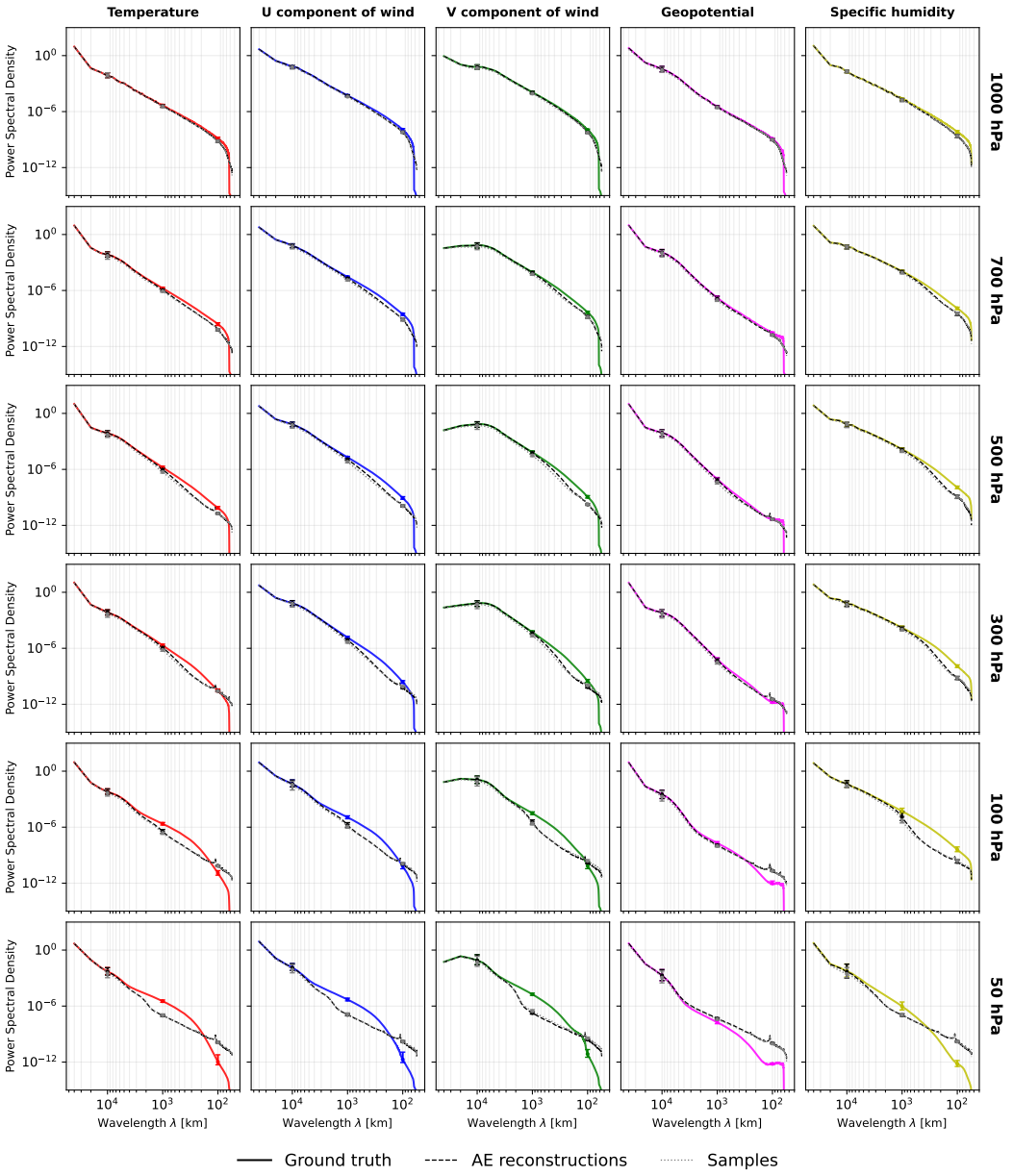

Figure 3. Power spectral density across wavelengths for atmospheric variables at selected pressure levels. Lines show median values and error bars indicate the 5th to 95th percentiles. The close alignment between the curves demonstrates that both the autoencoder and the diffusion model preserve the energy distribution across most spatial scales. Deviations begin to appear at wavelengths around 1000km, which corresponds to roughly 40 grid cells at our 0.25-degree resolution at the equator. These differences become more pronounced at smaller scales, suggesting that while large-scale atmospheric patterns are well-preserved, features spanning fewer than 40 grid cells show some energy loss in the compression and generation processes. Deviations become more pronounced at lower pressure levels, as the model prioritizes surface and low-altitude variables.

D.2 Physical consistency

To further evaluate physical consistency, we examine whether our model preserves important physical relationships between variables. First, we analyze the consistency between two different estimators of altitude at given pressure levels. Using the geopotential Φ, altitude can be derived as

$$H = \frac{\Phi R_e}{g_0 R_e - \Phi} \tag{9}$$

where R_e is Earth's radius and g_0 is the Earth gravitational acceleration at the surface. Alternatively, the equation below (which relies on the ideal gas law and hydrostatic equation) relates altitude to pressure and temperature as

$$\log \frac{p_0}{p_H} = \frac{M g_0}{R} \int_0^H \frac{1}{T_h} \partial h, \tag{10}$$

where R is the universal gas constant, M is an approximation of the atmosphere's molar mass, p_h, T_h are pressure and temperature at height h, and p_0, T_0 are the theoretical pressure and temperature at sea level. This integral can be approximated to extract H using several assumptions about the temperature profile. When comparing these two estimators, Figure 4 shows that our generated samples maintain the same systematic differences ΔH as seen in ground-truth data. This remarkable consistency indicates that our model successfully preserves this physical relationships between temperature, pressure, and geopotential, allowing altitude to be estimated through two independent methods with nearly identical accuracy to the original ERA5 data.

Second, we examine the geostrophic balance

$$\frac{\partial \Phi}{\partial x} = \frac{4\pi \Omega R_e}{N_x} \sin \phi \cos \phi u_g$$
$$\frac{\partial \Phi}{\partial y} = -\frac{2\pi \Omega R_e}{N_y} \sin \phi v_g$$

which is the theoretical equilibrium between pressure gradient forces and Coriolis forces that governs large-scale atmospheric motion. In the above system, ϕ is the latitude, N_x and N_y are the number of pixels along longitude and latitude, Ω is the magnitude of the Earth's angular velocity, and u_g and v_g denote the eostrophic components of the wind. In this balance, in the absence of vertical motion, friction, and isobaric curvature, wind direction should be perpendicular to geopotential gradients, with wind speed proportional to gradient magnitude. This relationship can be expressed by comparing two quantities: (1) the angle θ between wind and geopotential gradients, which should approach 90° in geostrophic conditions, and (2) the correlation between wind speed magnitude and geopotential gradient magnitude, which should approach 1 in perfect geostrophic balance. Figure 4 shows that our generated samples accurately reproduce both aspects of this relationship. At 500 hPa, the approximate level of non-divergence with minimal surface friction effects, both ERA5 data and our generated samples show angles concentrated around 90°. Near the surface at 1000 hPa, where additional forces become significant, both datasets show a systematic deviation in angle. Similarly, the correlation between wind speed and geopotential gradient magnitudes in our samples closely matches the patterns observed in ERA5 data, exhibiting imperfect correlation only at lower pressure levels (explained by the level-weighted training) and following the same decreasing trend as pressure increases toward the surface, where ageostrophic components become more prominent.

These results demonstrate that our latent diffusion model not only preserves the statistical properties of atmospheric fields but also maintains important physical relationships between variables producing trajectories that are physically consistent and realistic. While these analyses confirm strong spatial consistency and physical fidelity, future work should extend our evaluation to more thoroughly quantify the temporal consistency of generated trajectories.

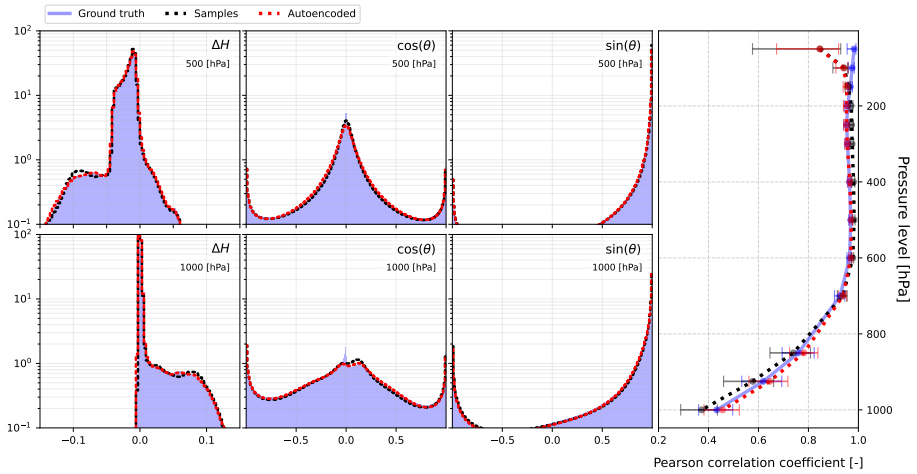

Figure 4. Physical consistency analysis of generated atmospheric states. (Top row) Analysis of altitude consistency at 500 hPa showing the difference ΔH between two independent altitude estimators, and geostrophic balance assessment through the cosine and sine of the angle θ between wind direction and geopotential gradients, demonstrating angles concentrated around 90°. (Bottom row) Same metrics at 1000 hPa demonstrating the presence of a significant ageostrophic component near the surface. (Right) Correlation coefficient between wind magnitude and geopotential gradient magnitude across pressure levels, showing strong correlation at upper levels (near 1) with a consistent decrease toward the surface in both ERA5 data (blue) and generated samples (black dots), confirming Appa's ability to capture complex physical relationships.

D.3 Quantitative evaluation

D.3.1 Compression error

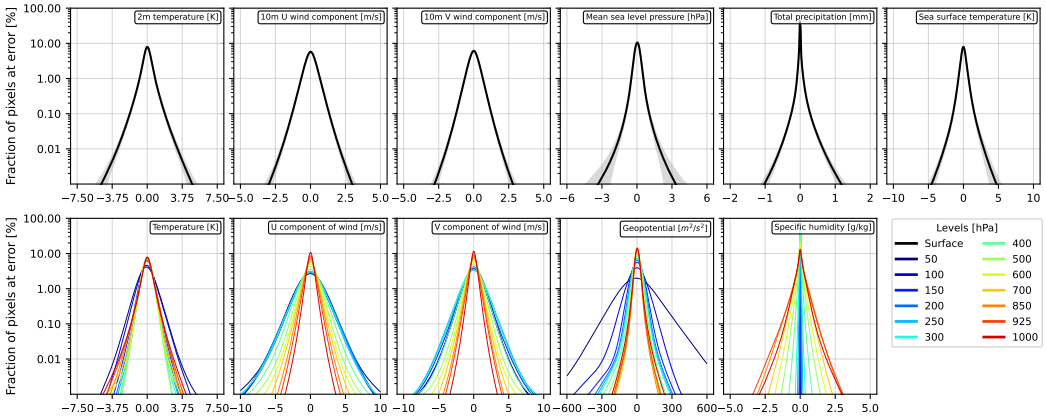

Figure 5. Signed reconstruction errors for surface variables (top) across all grid points and atmospheric variables (bottom) across all pressure levels. Shaded gray area for surface variables corresponds to error spread. In both cases, the concentrated distributions centered around zero demonstrate unbiased and precise predictions. Given $721 \times 1440 = 1,038,240$ grid points, a 0.01% fraction on the y-axis corresponds to approximately 100 grid points, indicating that large errors are rare.

D.3.2 Additional variables

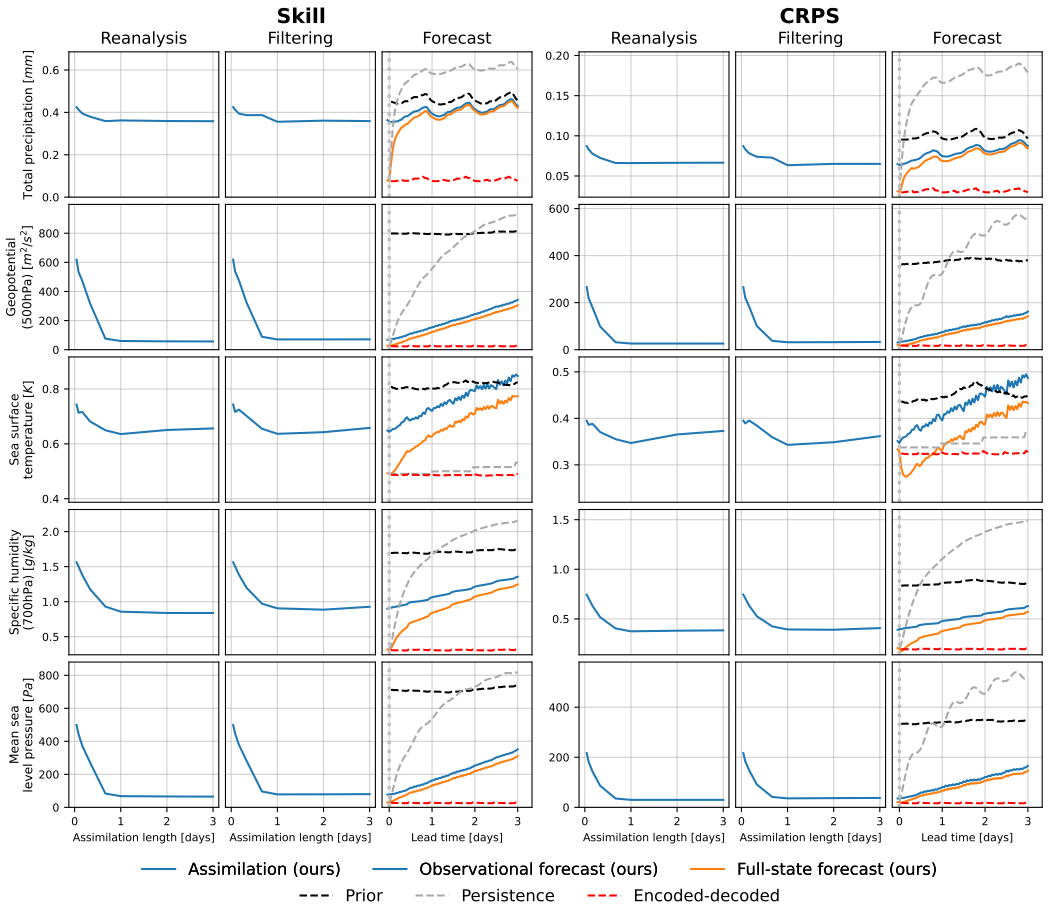

Figure 6. Quantitative evaluation of reanalysis, filtering, and forecasting for additional representative variables. (Left) Skill score and (Right) Continuous Ranked Probability Score (CRPS). Reanalysis scores are averaged over assimilation windows, while filtering reports the last reanalyzed state. Both improve with longer windows. Forecasts gradually lose skill over lead time but remain above persistence and unconditional baselines.

D.3.3 Spread and spread-skill ratios

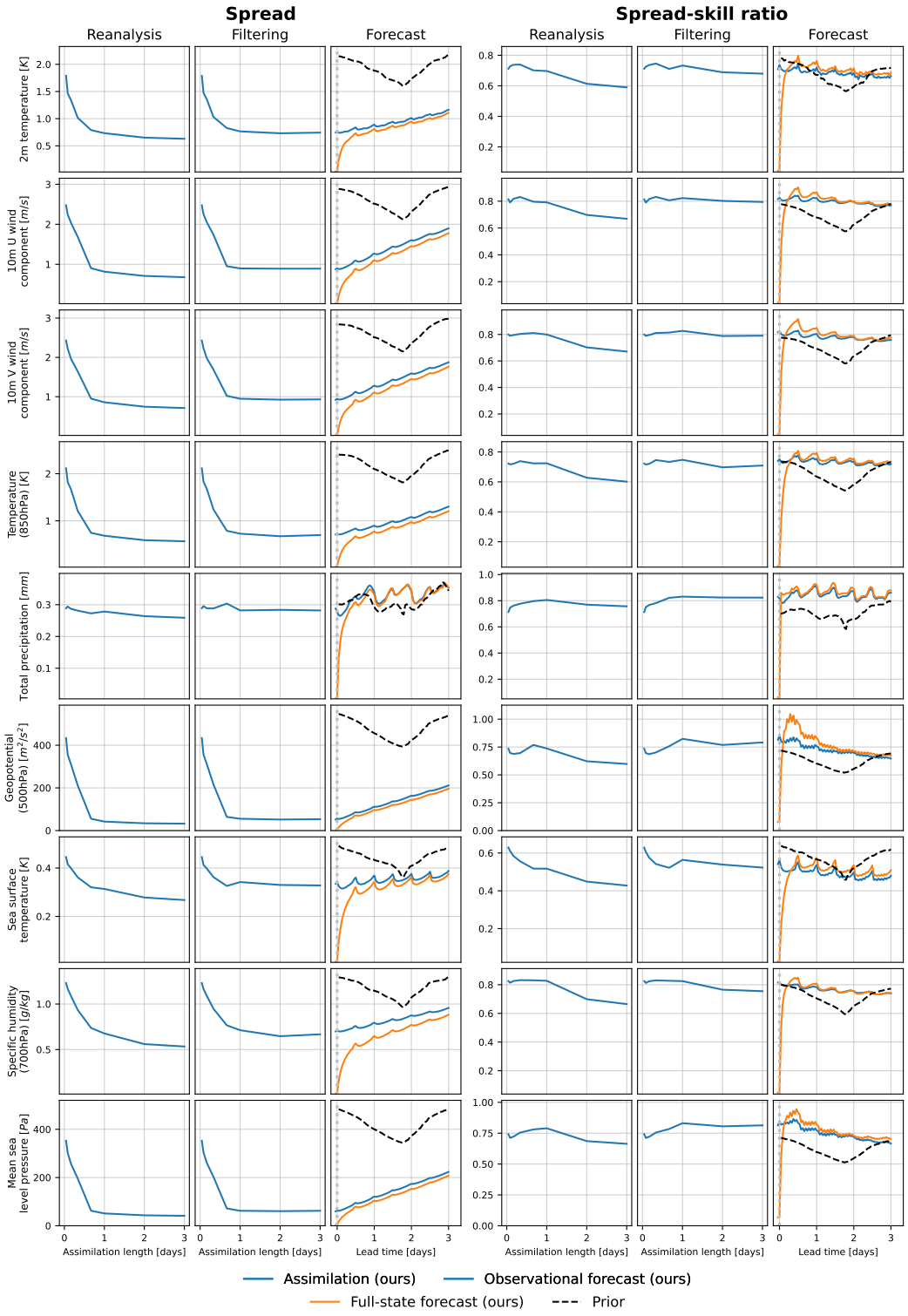

Figure 7. Quantitative evaluation of reanalysis, filtering, and forecasting. (Left) Spread and (Right) Spread-skill ratio for representative variables. Reanalysis scores are averaged over assimilation windows, while filtering reports the last reanalyzed state. Ensemble spread decreases with longer windows, while ratios remain fairly unchanged. A ratio below one indicates overconfidence.

D.4 Qualitative snapshots

In Figures 8 to 13, we display decoded sampled trajectories generated through reanalysis over a window of three days as well as through observational forecasting initialized from reanalyzed states, for six representative variables. The second row of each gallery shows the observed pixels, if the state was observed. Each mask is displayed as the corresponding variable was observed during inference, i.e., as ground stations for surface variables and satellite scans for atmospheric variables.

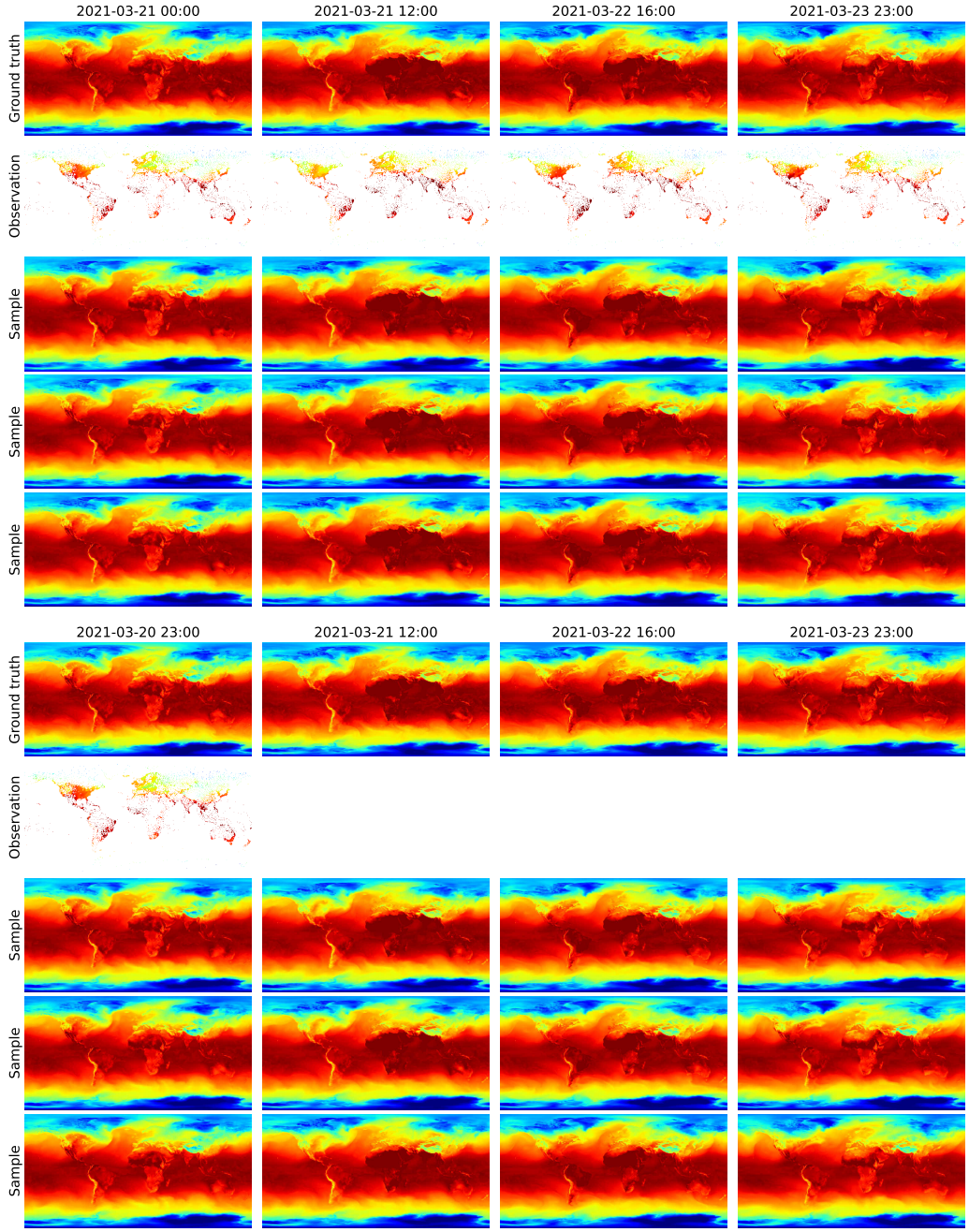

Figure 8. Reconstructed sampled trajectories for surface temperature assimilation. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.

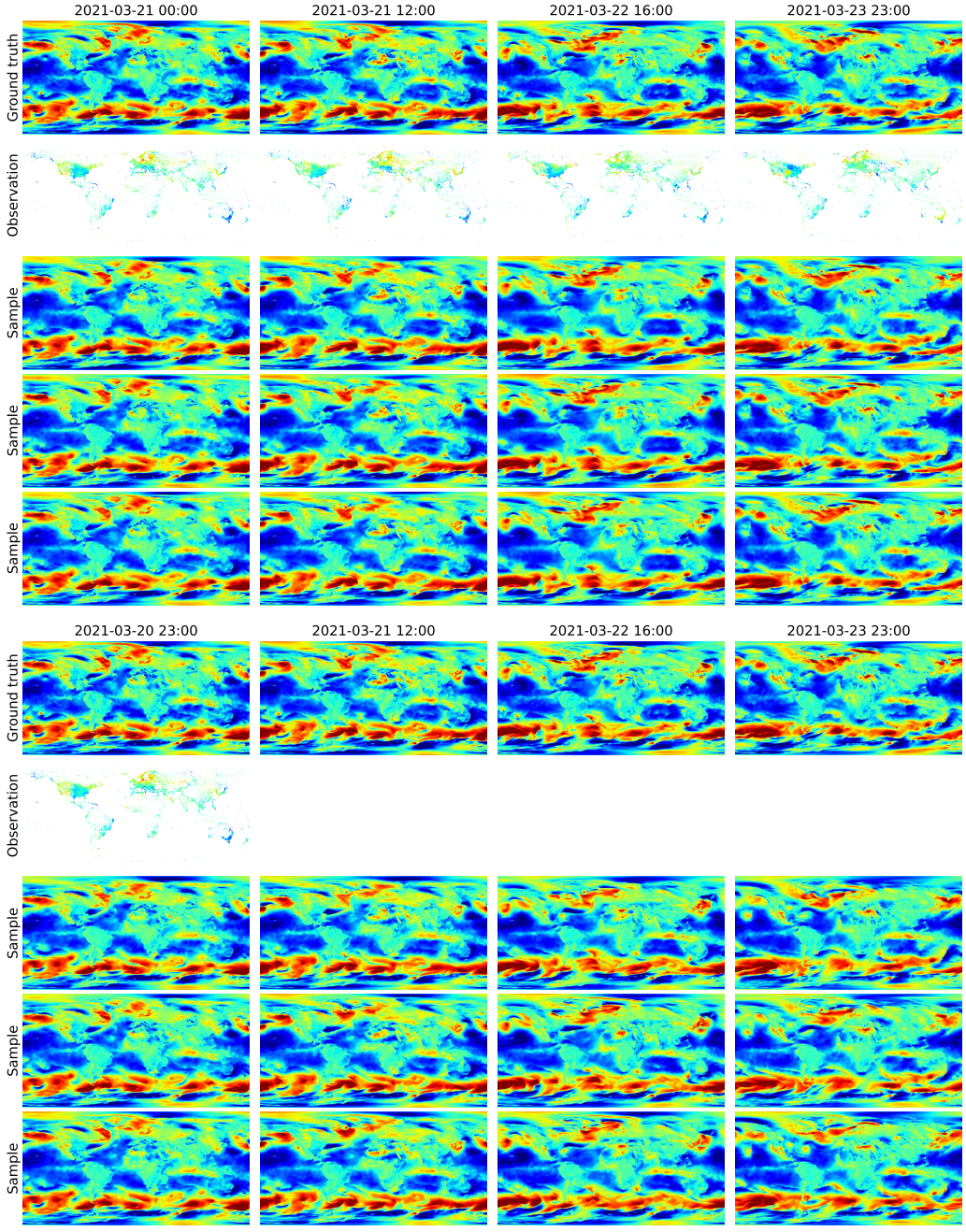

Figure 9. Reconstructed sampled trajectories for surface eastward wind speed assimilation. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.

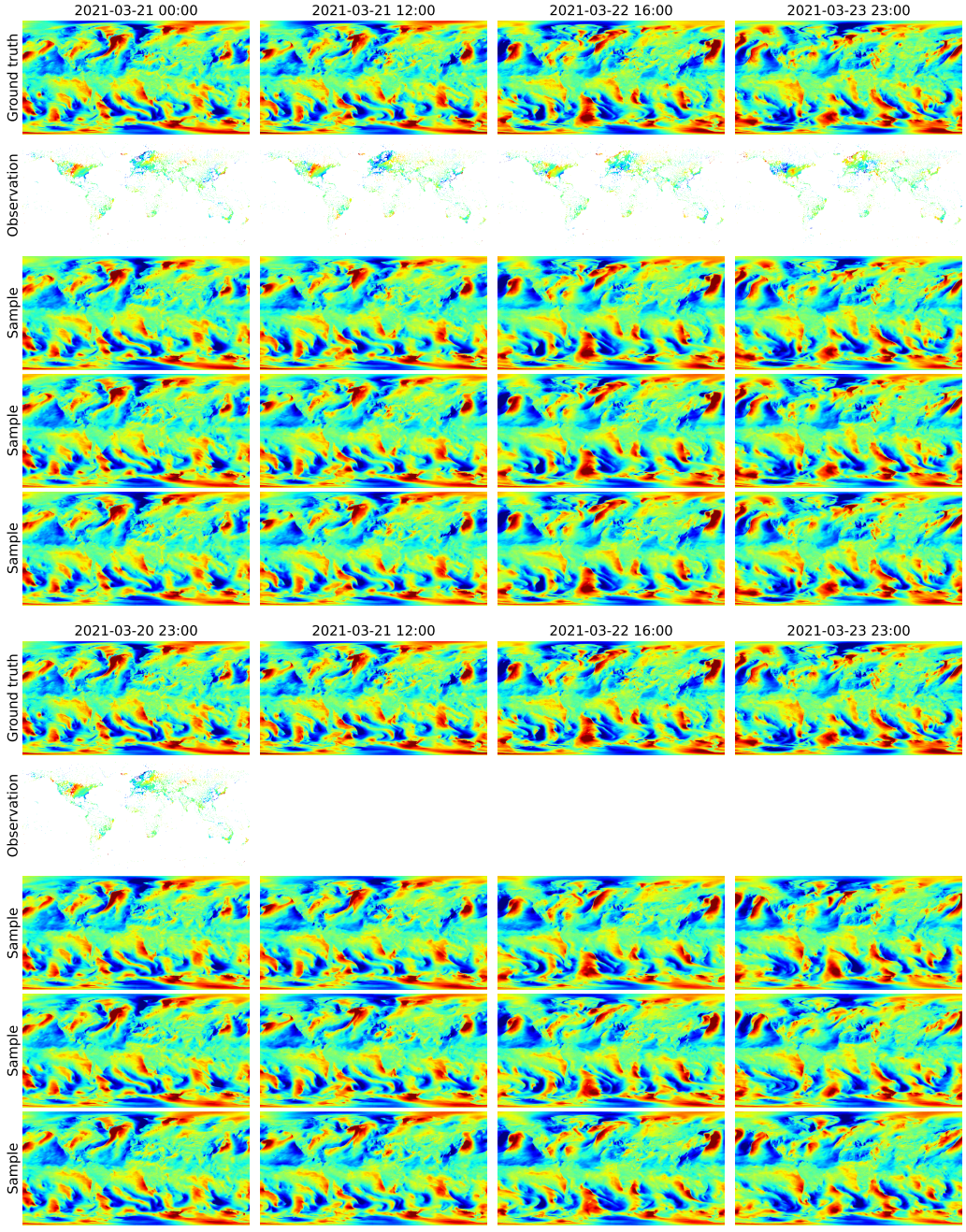

Figure 10. Reconstructed sampled trajectories for surface northward wind speed assimilation. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.

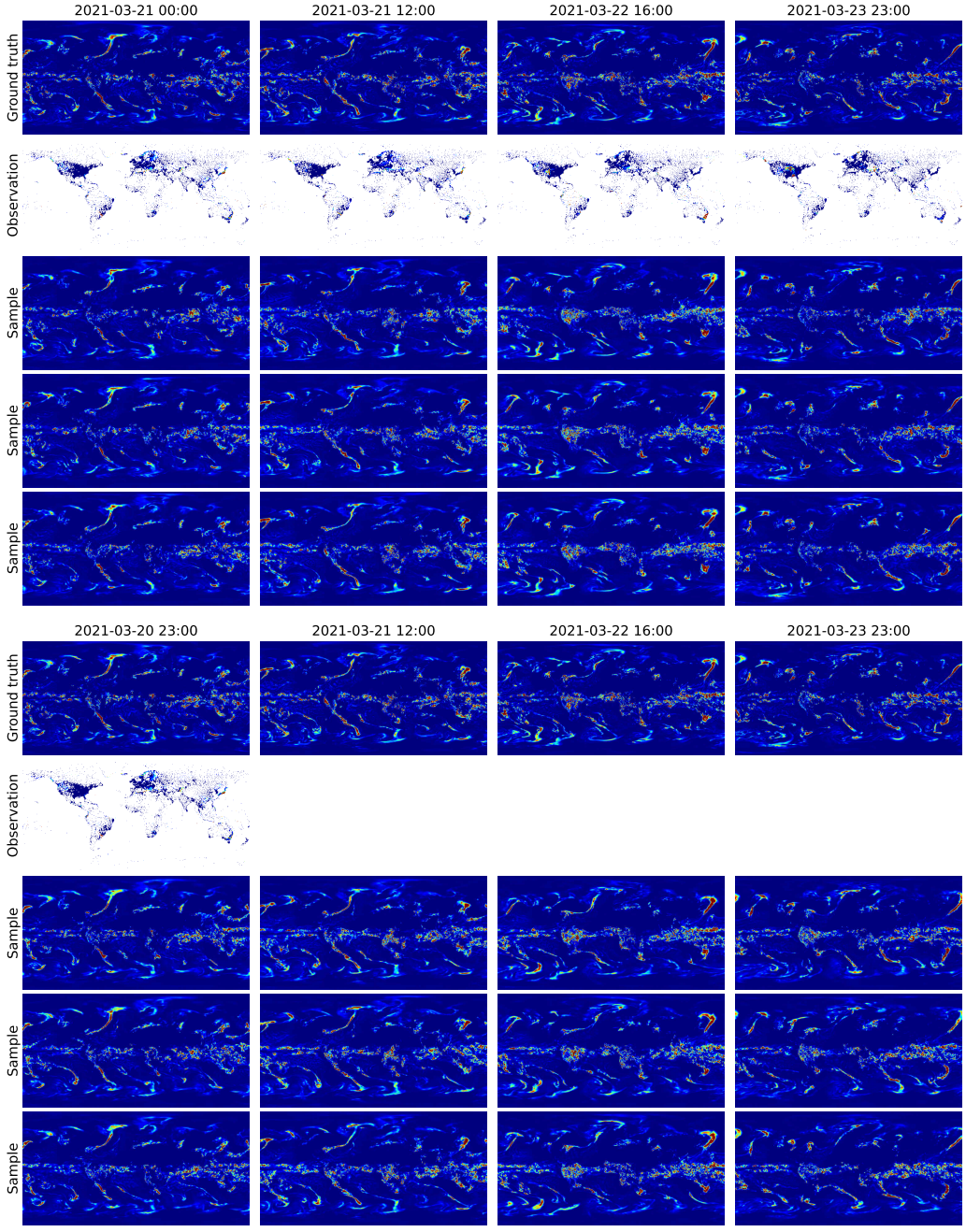

Figure 11. Reconstructed sampled trajectories for total precipitation assimilation. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.

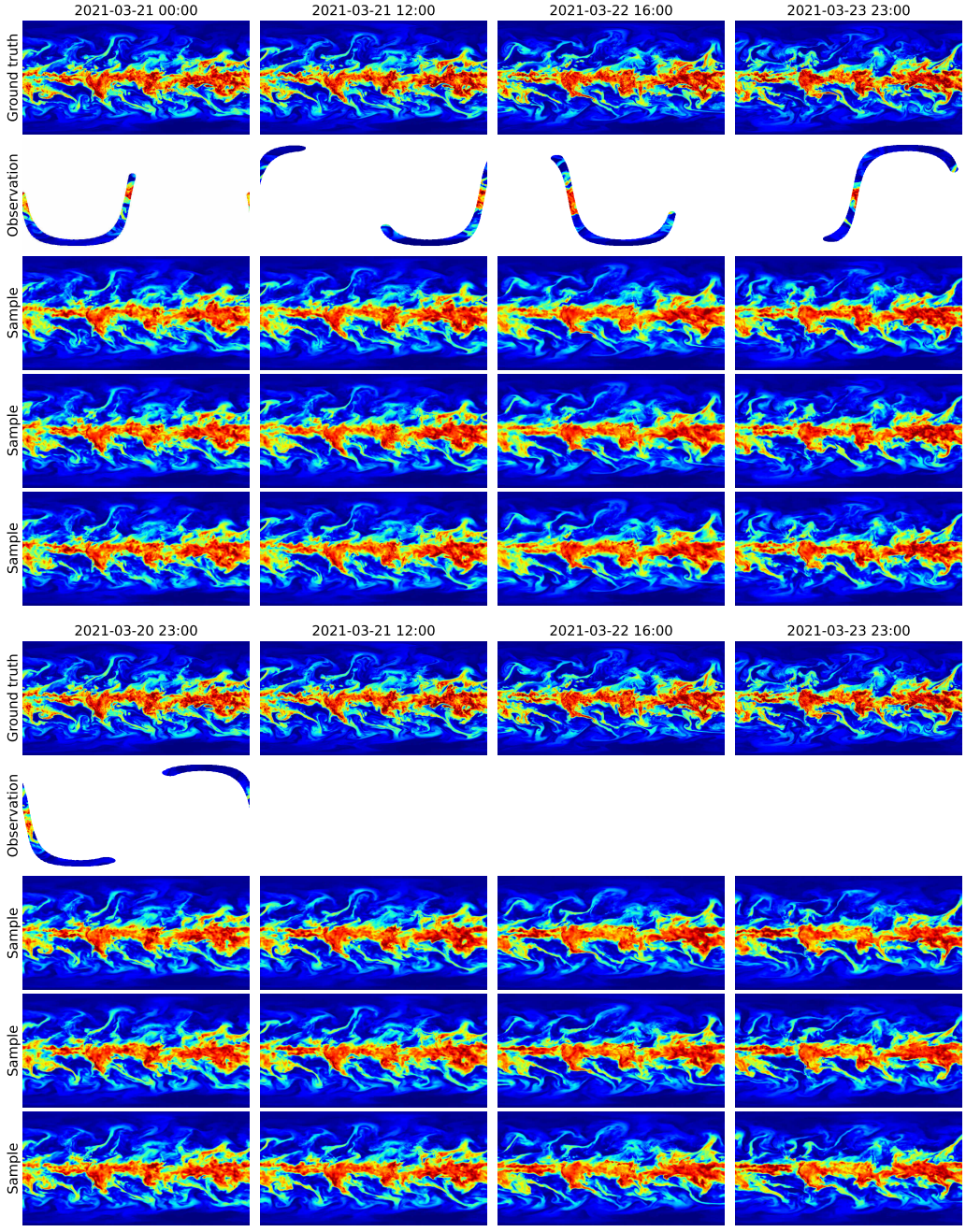

Figure 12. Reconstructed sampled trajectories for specific humidity assimilation at 700hPa. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.

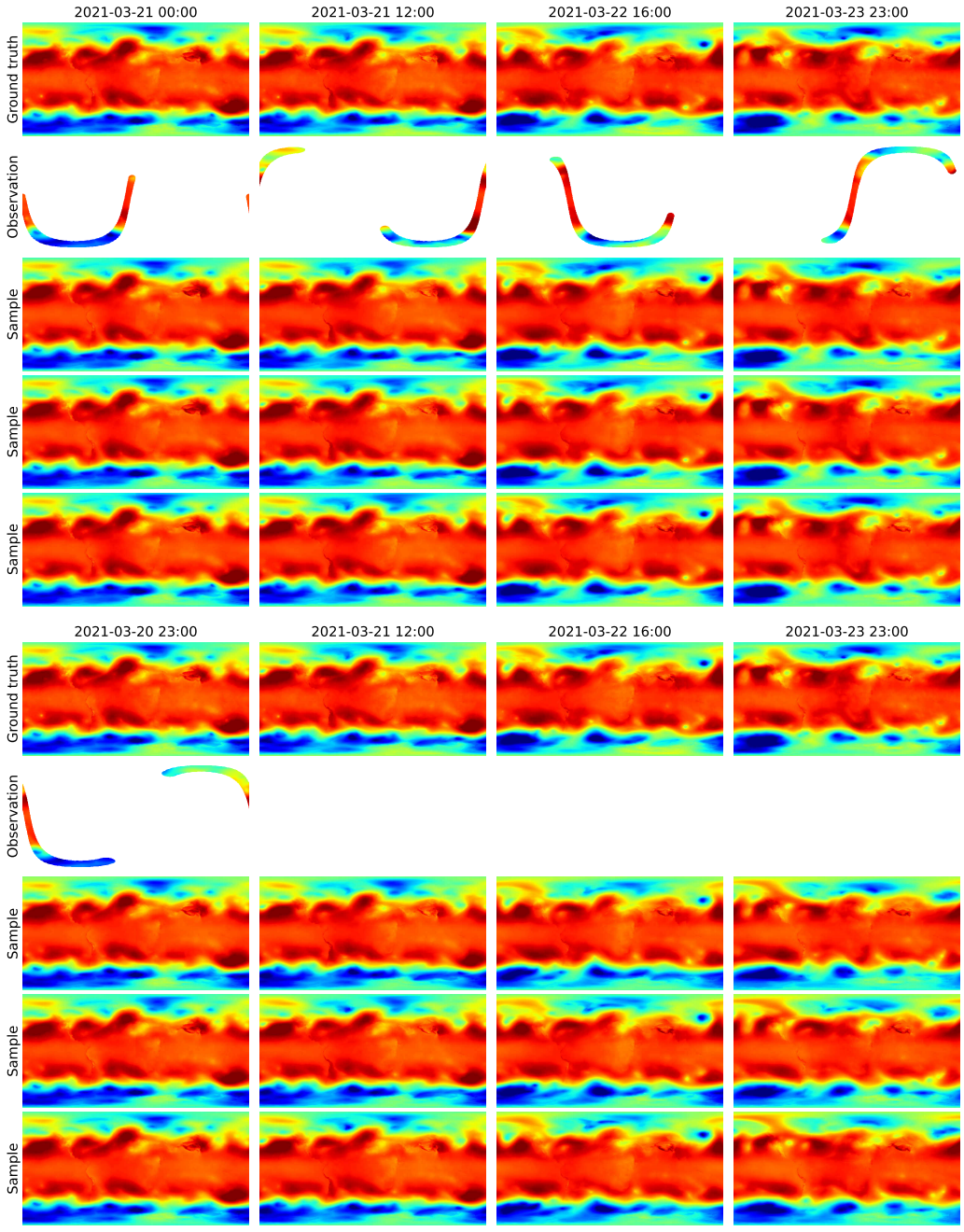

Figure 13. Reconstructed sampled trajectories for geopotential assimilation at 850hPa. (Top) Reanalysis over a window of 72 hours. (Bottom) Observational forecasting over 3 days initialized with the last 12 states of an assimilation over 24 hours.