RISDA: RELIABILITY-INFORMED SCORE-BASED DATA ASSIMILATION OF MULTI-SOURCE ATMOSPHERIC OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Score-based atmospheric data assimilation separates a learned atmospheric prior from observation guidance, making it naturally flexible to changing observations. However, this flexibility breaks down for real-world heterogeneous multi-source observations, where the main bottleneck shifts to the guidance itself: different observations cannot always be directly matched to atmospheric states, and the estimated guidance can vary greatly in reliability across sources, regions, and latent channels. Therefore, we introduce **RISDA** (**R**eliability-**I**nformed **S**core-based **D**ata **A**ssimilation), a framework designed for the assimilation of real-world multi-source observations. First, an atmospheric encoder and source-specific observation encoders map atmospheric states and heterogeneous observations into a shared latent space, bridging variable mismatches and enabling direct guidance for posterior sampling. Second, empirical diagnostics of latent errors motivate Variance-Calibrated Latent Guidance (VCLG), which constructs more accurate observation guidance by accounting for channel-specific uncertainties within the shared latent space. Third, to reasonably combine guidance from multiple sources during diffusion, we introduce Priority-Preserving Gradient Fusion (PPGF), which merges source-specific gradients according to reliability-informed priorities, ensuring that guidance from more reliable sources is preserved during sampling. Assimilation experiments incorporating five types of satellite observations with the surface observation demonstrate that RISDA achieves superior analysis accuracy and higher inference efficiency compared to baseline methods.

1 INTRODUCTION

Data assimilation (DA) combines prior knowledge with observations to estimate the initial conditions for numerical weather prediction (Bauer et al., 2015; Carrassi et al., 2018). A central goal is to improve atmospheric state estimation by integrating complementary information from multiple observation sources, such as satellite and surface observations. Score-based DA offers a promising approach by learning a generative prior over atmospheric states and incorporating observations through likelihood guidance during posterior sampling (Song et al., 2021; Rozet & Louppe, 2023; Andry et al., 2025). This pattern allows the same prior to accommodate different observation settings without retraining. Recent studies have shown the potential of this approach for simulated global observations, sparse weather-station measurements, and satellite-derived products (Huang et al., 2024; Si & Chen, 2025; Xiao et al., 2026; Martin et al., 2025; Manshausen et al., 2025). However, most existing score-based DA methods still focus on single-source or simplified observation settings (Si & Chen, 2025; Andry et al., 2025), and reliable fusion of real-world heterogeneous observations from multiple sources remains insufficiently addressed.

Reliably integrating real multi-source observations into score-based data assimilation is challenging because observations must first be turned into useful guidance and then combined reasonably. This involves two main challenges. First, observations from different instruments are not always expressed in the same physical variables as the atmospheric state. For example, microwave satellites observe the atmosphere through brightness temperatures, while atmospheric states are described by variables such as temperature, humidity, and wind. A common solution is to use a forward observation operator to map the atmospheric state into the observation space (Chung et al., 2023; Rozet et al., 2024).

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.

However, the guidance then depends on gradients through this operator, making the update sensitive to errors in the mapping and its derivatives (Li et al., 2024). Second, even when guidance from each source is available, these likelihoods and gradients are empirically estimated in diffusion, and their reliability varies across instruments, variables, and regions. Combining these approximate guidance signals in practice is nontrivial. Simply combining source guidance can allow less reliable gradients to weaken more reliable ones.

These challenges suggest that reliable multi-source assimilation requires more than simply adding observations from different sources: their guidance should first be made comparable, and their differences in reliability should then be reflected in how that guidance is combined. First, heterogeneous observations and atmospheric states can be represented in a shared latent space, providing a common basis for constructing observation guidance without relying on a separate observation-space mapping for each source (Si & Chen, 2025; Qu et al., 2024). Second, once different sources are represented in the same space, their errors relative to a common atmospheric reference provide a natural way to compare their reliability. These reliability estimates can then guide multi-source fusion, so that more reliable guidance is preserved when different sources disagree during sampling.

We instantiate these two insights in **RISDA** (**R**eliability-**I**nformed **S**core-based **D**ata **A**ssimilation) through three concrete components. First, we design source-specific observation-to-latent (O2L) encoders that map observations into a shared atmospheric latent space defined by an autoencoder pretrained on atmospheric states. This bridges the variable-space mismatch and enables efficient unified gradient guidance through latent matching. Second, we analyze the errors between encoded observations and encoded atmospheric states, finding approximately Gaussian marginal distributions and weak cross-channel correlations. These empirical diagnostics of latent errors motivate Variance-Calibrated Latent Guidance (VCLG), which constructs more accurate observation guidance by accounting for channel-specific uncertainties within the shared latent space. Finally, to further reliably combine guidance from multiple sources during diffusion, we use residual mean square errors estimated separately at each latent location and channel as an empirical proxy for source reliability, and introduce Priority-Preserving Gradient Fusion (PPGF), which merges source-specific gradients according to reliability-informed priorities, ensuring that guidance from more reliable sources is preserved during sampling. Together, these components enable multi-source observations to jointly guide atmospheric posterior sampling while accounting for differences in source reliability. Experimental results demonstrate that our method improves atmospheric analysis accuracy compared to baseline methods and achieves higher inference efficiency. Our main contributions are threefold:

- We develop source-specific observation-to-latent interfaces that map heterogeneous observations into a shared atmospheric latent space, enabling efficient unified observation guidance without relying on source-specific forward operators during sampling.
- Based on the approximately Gaussian marginal error distributions and weak cross-channel correlations in the latent space, we propose Variance-Calibrated Latent Guidance (VCLG), which accounts for channel-specific uncertainties within the shared latent space for more accurate observation guidance.
- The latent residuals quantify source reliability in the shared space. Based on the source reliability, we propose Priority-Preserving Gradient Fusion, which uses the estimated reliability to order and fuse source-specific guidance, preserving more reliable gradients when different sources disagree.

2 Related Work

2.1 Traditional and ML-based Atmospheric Data Assimilation

Traditional DA estimates the state of a dynamical system by combining a numerical forecast with observations within a Bayesian framework. For high-dimensional atmospheric systems, four-dimensional variational (4D-Var) methods and ensemble Kalman filters (EnKF) are widely used in operational numerical weather prediction (Bonavita et al., 2012; Berre et al., 2015; Houtekamer & Zhang, 2016). These methods rely on approximations to background and observation errors and can require substantial computational resources. Recent end-to-end data assimilation approaches substantially accelerate this process by directly learning the mapping from observations and background states to atmospheric analysis fields (Allen et al., 2025; Xu et al., 2025; Sun et al., 2025c; Gupta

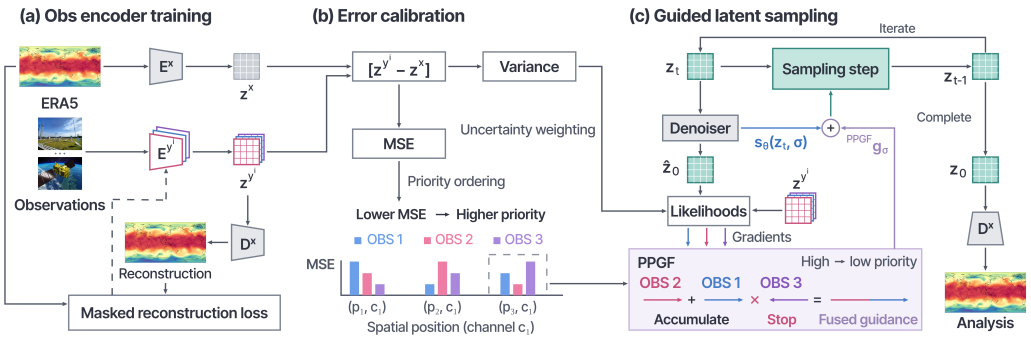


Figure 1: Overview of RISDA. (a) Source-specific observation encoders are trained through the frozen atmospheric decoder using a masked reconstruction loss. (b) Observation–ERA5 latent errors provide variances for likelihood calibration and MSEs for source-priority ordering. (c) PPGF fuses source-specific likelihood gradients in priority order. The fused guidance is combined with the prior score during latent sampling, and the final latent is decoded into an atmospheric analysis.

et al., 2026; Xu et al., 2026). However, they suffer from high training overheads and lack operational flexibility, as incorporating new observation types usually necessitates complete model retraining.

2.2 Generative score-based data assimilation

Score-based data assimilation (Rozet & Louppe, 2023) formulates data assimilation as observation-guided sampling of state trajectories based on diffusion models (Sun et al., 2025a; Song et al., 2021). By decoupling prior learning from the observation model, it can accommodate changing observation configurations without retraining the prior. Hybrid diffusion further combines training-time initial-state conditioning with inference-time guidance for joint forecasting and assimilation (Shysheya et al., 2024). Applications include sparse weather-station assimilation (Manshausen et al., 2025) and surface-ocean reconstruction (Martin et al., 2025). Latent methods extend this framework to high-dimensional systems: Appa (Andry et al., 2025) learns priors over global atmospheric trajectories, LO-SDA (Sun et al., 2025b) combines background-conditioned priors with latent optimization, and Rozet et al. (2025) study compressed physics emulation with observation-guided sampling. Building on these, RISDA addresses the assimilation of real-world heterogeneous multi-source observations. Through source-specific encoders, it unifies multi-source observations into a shared latent space and ensures reasonable multi-source observational guidance via reliability-informed ranking.

3 Method

In this section, we present **RISDA**, a framework for assimilating real multi-source observations through latent observation alignment and reliability-informed gradient fusion, as shown in Figure 1. We first introduce the score-based DA formulation and the shared atmospheric latent space. We then use latent error statistics to construct Variance-Calibrated Latent Guidance (VCLG), before introducing Priority- Preserving Gradient Fusion (PPGF) to combine the source gradients according to their estimated reliability.

3.1 Preliminaries

3.1.1 Score-based data assimilation with latent diffusion

Let $\boldsymbol{x}$ denote the atmospheric state and $\boldsymbol{y}$ the observations. Data assimilation combines a prior $p(\boldsymbol{x})$ with an observation likelihood $p(\boldsymbol{y} \mid \boldsymbol{x})$ to infer the posterior $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{x})p(\boldsymbol{y} \mid \boldsymbol{x})$. Following Appa (Andry et al., 2025), we adopt latent diffusion models for score-based data assimilation. First, an atmospheric autoencoder ($E^{\boldsymbol{x}}$, $D^{\boldsymbol{x}}$) trained on ERA5 maps atmospheric states to latent representations. For atmospheric states $\boldsymbol{X} = \{\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^L\}$ at L sequential timestamps, their corresponding latent representations are denoted as $\boldsymbol{Z} = \{\boldsymbol{z}^1, \boldsymbol{z}^2, \ldots, \boldsymbol{z}^L\} = \{E^{\boldsymbol{x}}(\boldsymbol{x}^1), E^{\boldsymbol{x}}(\boldsymbol{x}^2), \ldots, E^{\boldsymbol{x}}(\boldsymbol{x}^L)\}$. A latent

score-based diffusion model then learns a prior $p(\boldsymbol{Z})$ over the latents at noise level σ. At diffusion time t, Bayes' rule gives the posterior-score decomposition

$$\nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Z}_t \mid \boldsymbol{y}) = \nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{Z}_t) + \nabla_{\boldsymbol{Z}_t} \log p(\boldsymbol{y} \mid \boldsymbol{Z}_t). \tag{1}$$

The score model approximates the first term. The second term provides the observation-likelihood score. It is usually assumed that observations $\boldsymbol{y}$ of the state trajectory $\boldsymbol{x}^{1:L}$ follow a Gaussian observation process $p(\boldsymbol{y}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{y}|H(\boldsymbol{X}), \boldsymbol{R})$ (Rozet & Louppe, 2023; Andry et al., 2025), where H denotes the observation operator and $\boldsymbol{R}$ is the observation-error covariance. Extending this formulation to the latent space for observation sequences $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^L\}$ of the form $\boldsymbol{y}^i = H^i(D^{\boldsymbol{x}}(\boldsymbol{z}^i)) + \epsilon^i$ following Appa (Andry et al., 2025), the observation likelihood conditioned on the noisy latent trajectory is approximated as

$$p(\boldsymbol{Y} \mid \boldsymbol{Z}_t) \approx \mathcal{N}\big(\boldsymbol{Y} \;\big|\; A(\mathbb{E}[\boldsymbol{Z} \mid \boldsymbol{Z}_t]), \boldsymbol{R} + \boldsymbol{A}\, \mathbb{V}[\boldsymbol{Z} \mid \boldsymbol{Z}_t]\, \boldsymbol{A}^\top\big)\,. \tag{2}$$

Here, A denotes the composition of the decoder D^x and the observation operator H, and $\boldsymbol{A}$ is the Jacobian of A. $\mathbb{E}[\boldsymbol{Z} \mid \boldsymbol{Z}_t]$ can be provided by the denoising model (Chung et al., 2023). $\mathbb{V}[\boldsymbol{Z} \mid \boldsymbol{Z}_t]$ denotes the uncertainty in estimating the clean state $\boldsymbol{Z}$ from the noisy state $\boldsymbol{Z}_t$. Further details are provided in Appendix B.

3.1.2 Multi-source score-based data assimilation

Consider multi-source observation sequences $\boldsymbol{Y}^{1:N} = \{\boldsymbol{Y}^1, \boldsymbol{Y}^2, \ldots, \boldsymbol{Y}^N\}$ gathered from N distinct sources. Assuming conditional independence across sources given the atmospheric latent $\boldsymbol{Z}$, for tractable guidance, we use a composite-likelihood approximation formed from separately approximated source likelihoods $\widetilde{p}_{\text{comp}}(\boldsymbol{Y}^{1:N} \mid \boldsymbol{Z}_t) := \prod_{i=1}^{N} p^i(\boldsymbol{Y}^i \mid \boldsymbol{Z}_t)$. The resulting composite log-likelihood gradient decomposes as

$$\nabla_{\boldsymbol{Z}_t} \log \widetilde{p}_{\text{comp}}(\boldsymbol{Y}^{1:N} \mid \boldsymbol{Z}_t) = \sum_{i=1}^{N} \nabla_{\boldsymbol{Z}_t} \log p^i(\boldsymbol{Y}^i \mid \boldsymbol{Z}_t). \tag{3}$$

During diffusion sampling, we apply Equation 2 separately to each source to obtain $\boldsymbol{g}_t^i = \nabla_{\boldsymbol{Z}_t} \log \widetilde{p}^i(\boldsymbol{Y}^i \mid \boldsymbol{Z}_t)$, and then approximate the joint likelihood guidance by summing the source-wise gradients. Therefore, we have the joint score under multi-source guidance:

$$\boldsymbol{s}(\boldsymbol{Z}_t, t) = \boldsymbol{s}_\theta(\boldsymbol{Z}_t, t) + \sum_{i=1}^{N} \boldsymbol{g}_t^i. \tag{4}$$

3.2 Score-Based Data Assimilation in Unified Latent Space

3.2.1 Observation guidance in a shared latent space

The raw satellite observations are brightness temperatures (TBs), which do not directly correspond to the variables in ERA5. We investigate two approaches to address this challenge. One is utilizing source-specific observation encoders $E^{\boldsymbol{y}}$ to map them into a **unified latent space** aligned with the atmosphere, as shown in Figure 1, enabling direct guidance for posterior sampling in the latent space. The other is using an observation operator P to map the atmospheric latent to the satellite observation space, and the guidance can be obtained by backpropagation through the operator.

Observation-to-Latent (O2L) operator We describe the route for a given observation source $\boldsymbol{y}^i$. Let $\boldsymbol{y}^i_{grid} \in \mathbb{R}^{H \times W \times C}$ denote the gridded observations. We can obtain its latent representation through a trained observation encoder $\boldsymbol{z}^{\boldsymbol{y}^i} = E^{\boldsymbol{y}^i}(\boldsymbol{y}^i_{grid}) \in \mathbb{R}^{\hat{H} \times \hat{W} \times \hat{C}}$, where $\boldsymbol{z}^{\boldsymbol{y}^i}$ has the same spatial and channel dimensions as the ERA5 latent. Further training details are provided in Appendix C. During posterior guided sampling, the likelihood guidance is computed using solely the subset $\boldsymbol{z}^{\boldsymbol{y}^i}_{disc} \subset \boldsymbol{z}^{\boldsymbol{y}^i}$ that corresponds to the discrete observation locations. Specifically, for each observation source i, we extract the observed subset $\boldsymbol{z}^{\boldsymbol{y}^i}_{disc} = \big[\boldsymbol{z}^{\boldsymbol{y}^i}\big]_{(h,w)\in\Omega} \in \mathbb{R}^{K \times \hat{C}}$. Here, $\Omega = \Big\{(h, w) \mid \mathbf{m}^i_{h,w} = 1,\ 0 \le h < \hat{H},\ 0 \le w < \hat{W}\Big\}$, $K = |\Omega|$, and $\boldsymbol{m}^i \in \mathbb{R}^{\hat{H} \times \hat{W}}$ is the binary latent observation mask. Following Gaussian observation process (Rozet & Louppe, 2023) and

MMPS (Rozet et al., 2024; Andry et al., 2025) approximation as in Equation 2, we obtain the likelihood in the latent space at diffusion time t as

$$p\Big(\boldsymbol{Z}_{disc}^{\boldsymbol{Y}^i}|\boldsymbol{Z}_t\Big) \approx \mathcal{N}\Big(\boldsymbol{Z}_{disc}^{\boldsymbol{Y}^i} \;\Big|\; \mathbb{E}[\boldsymbol{Z} \mid \boldsymbol{Z}_t], \boldsymbol{R}_{O2L}^i + \mathbb{V}[\boldsymbol{Z} \mid \boldsymbol{Z}_t]\Big)\,. \tag{5}$$

Here, $\boldsymbol{R}_{O2L}^i$ is the observation-error covariance of observations in latent space. In this likelihood, both the conditional mean and covariance are restricted to the same selected latent elements. Similar to Equation 3, the joint log-likelihood gradient for multi-source observations in latent space can be expressed as:

$$\nabla_{\boldsymbol{Z}_t} \log \widetilde{p}_{\text{comp}}(\boldsymbol{Z}_{disc}^{\boldsymbol{Y}^{1:N}} \mid \boldsymbol{Z}_t) = \sum_{i=1}^{N} \nabla_{\boldsymbol{Z}_t} \log p^i(\boldsymbol{Z}_{disc}^{\boldsymbol{Y}^i} \mid \boldsymbol{Z}_t) = \sum_{i=1}^{N} \boldsymbol{g}_{O2L,t}^i. \tag{6}$$

Latent-to-Observation (L2O) operator We also give the other scheme based on an observation operator P, which maps atmospheric latents to satellite observation space. We adapt Appa to use this learned L2O operator as an observation-space guidance baseline. More details are provided in Appendix C.

3.2.2 Latent error analysis

To assess the Gaussian error approximation used by O2L guidance, we examine residuals between encoded observations and the corresponding ERA5 latents. For a given instrument, the difference between the observation latent and corresponding ERA5 latent is:

$$\boldsymbol{e}_{latent}^i = \big[\boldsymbol{z}^{\boldsymbol{y}^i} - \boldsymbol{z}\big]_{(h,w)\in\Omega} \in \mathbb{R}^{K\times\hat{C}}. \tag{7}$$

where $\Omega = \Big\{(h,w) \mid \mathbf{m}_{h,w}^i = 1,\ 0 \le h < \hat{H},\ 0 \le w < \hat{W}\Big\}$ denotes the set of valid positions. Here, $K = |\Omega|$ and $\hat{C} = 128$ denotes the number of latent channels. To visualize the distribution of $\boldsymbol{e}_{\text{latent}}^i$, we select 1,000 discrete samples from different time steps and valid positions, and compute pairwise Pearson correlations across different channels.

Figure 2 illustrates that the error correlations among MHS latent channels are predominantly concentrated along the diagonal elements, exhibiting weak cross-channel correlations. This characteristic justifies a simplified diagonal approximation for the observation error covariance matrix R_i^{O2L} within the O2L route (Fan et al., 2026). For the displayed MHS channel, the fitted Gaussian density and Q-Q plot suggest approximately Gaussian marginal errors, supporting the Gaussian approximation in Equation 5. Appendix D provides diagnostics for other observation sources.

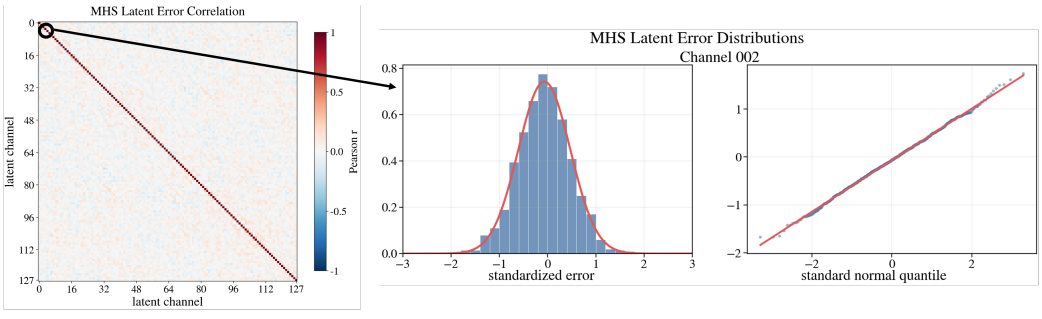


Figure 2: MHS observation–ERA5 latent errors. Left: Pearson correlations across all 128 latent error channels. Right: error histogram with a fitted Gaussian density and Q–Q plot for latent channel 002.

3.3 Variance-Calibrated Latent Guidance

Motivated by the error analysis in Section 3.2.2, we introduce **Variance-Calibrated Latent Guidance (VCLG)** to construct observation guidance with channel-specific uncertainty in the shared latent space. The Gaussian marginal errors in the latent space support the Gaussian observation process

and likelihood in Equation 5. Observing weak cross-channel correlation in the latent space, we can furthermore approximate $\boldsymbol{R}^i_{O2L}$ in the likelihood as a diagonal matrix.

Specifically, for each source i and latent channel c, we collect valid residuals from Equation 7 on a calibration set and estimate their centered variance v^i_c. The channel-wise variance estimates are computed offline across temporal and spatial positions. More details are provided in Appendix E.

Rather than constructing a full matrix, we directly parameterize the diagonal observation error covariance $\mathbf{R}^i_{O2L}$ using these calibrated channel-wise variances, i.e., $\mathbf{R}^i_{O2L} = \mathrm{diag}(v^i_1, v^i_2, \dots)$. This explicitly incorporates channel-specific uncertainties into Equation 5, and this diagonal parameterization provides a tractable approximation to the latent observation-error covariance. At diffusion time t, the calibrated approximate log-likelihood provides source-wise guidance $\boldsymbol{g}^i_{O2L,t}$ through differentiation with respect to $\boldsymbol{Z}_t$. Adding the guidance from each source following Equation 6, we obtain additive gradient fusion (AGF) and the sampling score

$$^{\mathrm{AGF}}\boldsymbol{s}(\boldsymbol{Z}_t, t) = \boldsymbol{s}_\theta(\boldsymbol{Z}_t, t) + \sum_{i=1}^{N} \boldsymbol{g}^i_{O2L,t}. \tag{8}$$

3.4 Priority-Preserving Gradient Fusion

AGF+VCLG combines source-wise approximate likelihood gradients under a composite-likelihood formulation. In practical diffusion sampling, however, the source-wise likelihood approximations also depend on learned observation encoders and empirically calibrated error statistics. The resulting guidance gradients are therefore affected by encoder approximation errors and calibration errors, whose impact can vary across sources, spatial locations, and latent channels. Consequently, directly adding these gradients at each diffusion step may allow less reliable contributions to impair useful guidance from more reliable sources when they disagree. Inspired by gradient-conflict handling in multi-task learning (Yu et al., 2020), we introduce **Priority-Preserving Gradient Fusion (PPGF)** as a reliability-informed guidance strategy for diffusion sampling. At each sampling step, PPGF ranks source-wise gradient components by estimated reliability at each latent location and channel, then accumulates them from high to low priority, stopping at the first opposing contribution. The fused guidance is combined with the prior score to update the noisy latent state, protecting higher-priority contributions from cancellation by lower-priority sources.

Reliability-based source ordering We similarly assemble the valid residuals in Equation 7, and estimate the mean squared error (MSE) as a proxy of the source reliability. MSE measures the source-wise latent difference from the reference. Lower MSE indicates closer agreement with the target atmospheric latent. Moreover, source reliability varies across spatial positions. Specifically, for each source i, position p, and latent channel, we compute the residual MSE q^i_p, omitting the channel index for clarity. In PPGF, we also estimate spatial variance $v^i_{p,c}$ separately at each position p. We further visualize the variation of source-reliability across different positions in Figure 10 and Figure 11. Details of the estimation procedure are provided in Appendix E. PPGF orders sources deterministically by increasing MSE, independently at each latent position and channel:

$$\boldsymbol{\pi}_p = \mathrm{argsort}^{\uparrow}_i \, q^i_p. \tag{9}$$

Sequential gradient fusion At a certain diffusion step, trajectory frame, latent position p, and channel, let $g^1_p, \dots, g^N_p$ denote the scalar components of the VCLG reordered by $\boldsymbol{\pi}_p$, from highest to lowest priority. Starting from $G^0_p = 0$, we sequentially accumulate them for $i \in \{1, \dots, N\}$ as

$$G^i_p = \begin{cases} G^{i-1}_p + g^i_p, & G^{i-1}_p g^i_p \geq 0, \\ G^{i-1}_p \quad \text{(stop)}, & G^{i-1}_p g^i_p < 0. \end{cases} \tag{10}$$

Let $G^\star_p$ denote the accumulated value when fusion terminates. If no contribution opposes the accumulated guidance, PPGF reduces to AGF. Otherwise, it discards the first opposing contribution and all remaining lower-priority contributions. Replacing the additive term in Equation 8 gives

$$^{\mathrm{PPGF}}\boldsymbol{s}(\boldsymbol{Z}_t, t) = \boldsymbol{s}_\theta(\boldsymbol{Z}_t, t) + {}^{\mathrm{PPGF}}\boldsymbol{g}_t, \tag{11}$$

where $^{\mathrm{PPGF}}\boldsymbol{g}_t$ is formed by assembling $G^\star_p$ over all trajectory frames, latent positions, and channels.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Atmospheric states and observations. We consider global atmospheric data assimilation on a 720×1440 grid at 0.25° resolution, using ERA5 (Hersbach et al., 2020) as the gridded reference following previous DA studies (Xu et al., 2025; 2026; Huang et al., 2024; Sun et al., 2025a). The atmospheric state comprises 69 channels: geopotential (Z), relative humidity (R), temperature (T), zonal wind (U), and meridional wind (V) at 13 pressure levels, together with 2-m temperature (`t2m`), 10-m zonal and meridional winds (`u10`, `v10`), and mean sea-level pressure (`msl`). The pressure levels are 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, and 1000 hPa. We assimilate brightness-temperature observations from five types of microwave instruments (ATMS, AMSU-A, MHS, MWTS, and MWHS), together with conventional surface observations (CONVI).

Implementation details. We follow Appa (Andry et al., 2025) for the atmospheric autoencoder, diffusion model, and their training and sampling settings, except that we set the number of latent channels to 128. Each observation encoder uses the same backbone as the atmospheric encoder, with an input layer adapted to its observation source. We train each observation encoder using the masked atmospheric reconstruction loss in Equation 14, while keeping the atmospheric decoder frozen. Satellite observation inputs include observation masks and viewing information, and missing observations are excluded from the guidance. We train the atmospheric autoencoder and latent diffusion model on ERA5 data from 2017–2024. The observation encoders are trained on data from 2016–2024 for ATMS, AMSU-A, MHS, and CONVI, and from 2022–2024 for MWTS and MWHS. We calibrate observation-error statistics and determine source priorities using data from January–March 2025. Further implementation details are provided in Appendix C.

Evaluation protocol. We evaluate all methods over April–June 2025 at six-hourly synoptic times (00, 06, 12, and 18 UTC) and report latitude-weighted RMSE against ERA5. Appendix C provides details on observation processing, trajectory setup, the sampler, and the evaluation metric.

4.2 COMPARISON WITH BASELINES

Appa (Andry et al., 2025) uses the learned L2O route; all other settings use O2L. *AGF+VCLG* directly sums the source gradients constructed by VCLG. For comparison, *AGF*^{m} replaces each source's channel-wise variances with their mean across all latent channels and retains additive fusion. *PCGrad+VCLG* applies PCGrad (Yu et al., 2020) to the gradients obtained by VCLG during posterior sampling. *PPGF* fuses the calibrated gradients using priorities that vary with latent location and channel, whereas *PPGF*^{SD} uses a channel-wise priority order shared across locations. All settings use the same six observation sources. In Table 1, all VCLG settings except PPGF and PPGF^{SD} share the same channel-wise variances v_c^i across spatial positions.

Table 1: RMSE (↓) comparison among methods using all six observation sources.

Method	Pressure level (300/500/850 hPa)					Surface			
	Z ($\mathrm{m^2\,s^{-2}}$)	R (%)	T (K)	U ($\mathrm{m\,s^{-1}}$)	V ($\mathrm{m\,s^{-1}}$)	T2M (K)	U10M ($\mathrm{m\,s^{-1}}$)	V10M ($\mathrm{m\,s^{-1}}$)	MSL (Pa)
Appa	290.68/213.45/162.22	21.14/18.66/18.49	1.59/1.50/1.77	6.46/4.63/3.27	6.01/4.29/3.02	1.25	2.07	2.02	212.64
AGF^m	488.50/329.31/205.98	39.91/46.94/23.39	2.68/1.89/2.18	7.11/4.87/3.16	6.57/4.79/2.90	1.72	1.72	1.73	261.42
AGF+VCLG	194.27/163.40/140.73	20.61/20.44/14.37	1.27/0.99/1.29	3.91/3.05/2.31	3.79/2.91/2.20	1.26	1.36	1.38	184.90
PCGrad+VCLG	210.59/174.79/144.17	21.19/21.29/14.65	1.34/1.03/1.31	3.96/3.08/2.35	3.82/2.96/2.23	1.29	1.37	1.39	187.52
PPGF^{SD}+VCLG	178.04/154.25/138.06	19.15/18.01/13.82	**1.14**/0.94/1.25	3.77/2.97/2.27	3.66/2.83/2.16	1.24	**1.34**	**1.36**	182.64
PPGF+VCLG	**170.41/145.85/133.77**	**17.67/15.53/13.49**	1.15/**0.93/1.24**	**3.75/2.94/2.25**	**3.63/2.81/2.15**	**1.20**	1.35	1.37	**178.48**

Table 1 shows that AGF+VCLG achieves lower RMSE than Appa in 17 of the 19 reported variables, with exceptions at R500 and T2M, and it also reduces RMSE across all 19 variables relative to AGF^m, supporting the effectiveness of channel-wise variances to calibrate latent guidance. Compared with AGF+VCLG, PCGrad+VCLG increases RMSE by 0.72%–8.40% across all 19 reported variables, suggesting that resolving gradient conflicts without explicitly accounting for source reliability during fusion fails to improve analysis accuracy. In contrast, PPGF^{SD}/PPGF+VCLG show improvement by 1.22%–11.89% and 0.72%–24.02%, respectively, relative to AGF+VCLG. These results

Table 2: Mean per-sample inference time under the same condition.

Metric	Appa (L2O)	AGF+VCLG (O2L)	PPGF+VCLG (O2L)
Mean time (min/sample) ↓	30.19	3.51	3.51

support accounting for source reliability during gradient fusion. With spatially varying priorities, PPGF+VCLG achieves the lowest RMSE in 16 of the 19 variables, and performs consistently better than AGF+VCLG. Appendix A.1 provides the complementary comparison across all 69 atmospheric variables. We also provide the spatial RMSE maps in Appendix A.2 showing lower errors over broad regions for PPGF+VCLG than for Appa. All these improvements support the effectiveness of our proposed methods.

Inference efficiency To compare inference efficiency, we measure mean per-sample inference time over 80 assimilation samples on a single NVIDIA H800 GPU for Appa, AGF+VCLG, and PPGF+VCLG, with the same observation inputs for guidance and under the same hardware and software environment. Table 2 shows that both O2L settings require 3.51 minutes per sample, compared with 30.19 minutes for Appa, yielding an $8.60\times$ speedup. O2L constructs likelihood guidance directly in latent space, avoiding repeated evaluation of the learned L2O operators during sampling. PPGF+VCLG matches AGF+VCLG in mean runtime at the reported precision, while improving analysis accuracy (Table 1).

4.3 Ablation study

We isolate the observation guidance route and gradient fusion design through controlled ablations. Within each ablation, all settings other than the component under comparison are held fixed. For ablations using the same observation interface, calibrated variance and reliability statistics are reused, except for explicitly defined changes.

Table 3: RMSE (↓) comparison between two routes using ATMS as the only observation source.

Route	Pressure level (300/500/850 hPa)					Surface			
	Z ($\mathrm{m^2\,s^{-2}}$)	R (%)	T (K)	U ($\mathrm{m\,s^{-1}}$)	V ($\mathrm{m\,s^{-1}}$)	T2M (K)	U10M ($\mathrm{m\,s^{-1}}$)	V10M ($\mathrm{m\,s^{-1}}$)	MSL (Pa)
L2O	973.51/750.80/575.42	31.67/28.36/24.72	2.63/2.91/3.33	12.60/8.89/6.41	12.69/8.71/6.13	2.34	4.38	4.42	738.98
O2L	**509.94/369.44/267.67**	**25.56/23.06/18.49**	**1.75/1.81/2.12**	**7.62/5.35/3.82**	**7.65/5.24/3.74**	**1.68**	**2.53**	**2.62**	**350.27**

O2L versus L2O. We compare the two guidance routes using ATMS as the only observation source. We separately use the calibrated variance in the corresponding space to obtain the gradients: the unified latent space for O2L and the brightness-temperature observation space for L2O. Table 3 shows that O2L achieves lower RMSE for all 19 reported variable–level combinations. This comparison supports the effectiveness of our O2L guidance route for the satellite observations.

Gradient fusion strategies. We compare PPGF+VCLG with three controlled fusion variants using spatially resolved VCLG. For this ablation, we use the spatial variances $v^i_{p,c}$ that vary across latent positions and channels. All variants and the AGF+VCLG baseline obtain the same guidance from each source but have different fusion strategies. *Random* ordering replaces reliability-based priorities with random source orders while retaining the same stopping rule. *Top-1 only* uses only the highest-priority source guidance at each latent position and channel and discards all remaining source guidance. *Skip-and-continue* skips opposing contributions and continues adding subsequent compatible ones, rather than stopping at the first conflict. We report relative RMSE improvements across all atmospheric channels against this spatially calibrated AGF+VCLG baseline, which directly sums the source gradients.

Figure 3 shows that PPGF+VCLG achieves the largest RMSE reductions relative to the spatially calibrated AGF+VCLG baseline across all displayed channels. Random ordering effectively discards a subset of source gradients without using reliability information. Although it degrades some channels, it improves others, suggesting that some source contributions may be unhelpful when

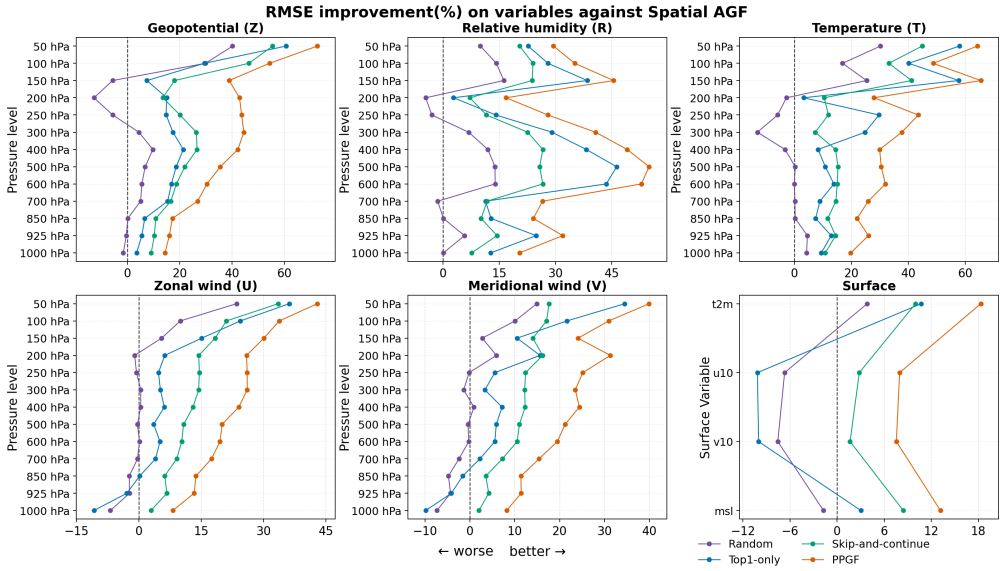


Figure 3: Relative RMSE reductions for PPGF+VCLG and three controlled fusion variants against spatially calibrated AGF+VCLG across all 69 atmospheric channels.

directly summed. Top-1 only achieves substantial gains over random ordering for most channels, supporting the value of reliability-based source selection. However, discarding all remaining sources may also remove useful compatible contributions and leave the guidance insufficient. This limitation is particularly evident in some lower-level and surface wind variables, where Top-1 only underperforms AGF+VCLG. Skip-and-continue improves over AGF+VCLG across all displayed channels but remains below PPGF+VCLG, and its ranking relative to Top-1 only varies across variables. Retaining more compatible gradients therefore does not consistently improve assimilation. A possible explanation is that lower-priority contributions amplify inaccurate guidance even when their signs agree with the accumulated direction. The RMSE time series in Appendix A.3 (Figure 8) complement these results: PPGF+VCLG generally achieves lower RMSE than the three variants over the evaluation period, indicating that the advantage on these variables is not confined to isolated evaluation times. Together, these comparisons suggest that PPGF benefits from retaining useful compatible contributions while limiting the accumulation of less reliable guidance.

5 Conclusion and Discussion

We introduced RISDA, a framework for assimilating real-world heterogeneous multi-source atmospheric observations through calibrated latent guidance and reliability-informed gradient fusion. Source-specific observation encoders map heterogeneous observations into a shared atmospheric latent space. Empirical residual diagnostics motivate a diagonal Gaussian approximation, enabling VCLG to calibrate likelihood guidance using latent error variances. PPGF then uses residual-based reliability estimates to preserve higher-priority guidance when source gradients conflict. Assimilation experiments incorporating multiple satellite instruments and conventional surface observations demonstrate that RISDA achieves a lower RMSE against ERA5 and shorter inference times compared to Appa. These results confirm that providing guidance within a unified latent space, combined with VCLG and PPGF, substantially improves atmospheric analysis accuracy. However, offline calibration may require updating when observation coverage or error statistics change, and both calibration and evaluation depend on ERA5. Future work could explore adaptive calibration under changing observation conditions and the use of complementary independent observational references. Despite these limitations, RISDA helps extend SDA beyond simplified observation settings to real global multi-source assimilation, providing a foundation for further development.

AI USE STATEMENT

We used generative AI tools to support literature search and summarization and to assist with drafting and editing portions of this manuscript. We have reviewed all AI-assisted work: the AI-assisted edits were checked line by line against our original drafts by the authors, and every claim, number, figure, and artifact in this paper was produced and verified by the authors. The authors take responsibility for verifying all AI-assisted content and for the final methods, experimental results, and conclusions.

ETHICS STATEMENT

This work uses meteorological observations and reanalysis data for atmospheric data assimilation. It does not involve human subjects or personal data.

REPRODUCIBILITY STATEMENT

The paper and appendix provide the method formulation, model configurations, calibration procedures, training hyperparameters, baseline settings, and evaluation protocols. We also report the training, calibration, and evaluation periods, sampling settings, and RMSE aggregation procedure to support reproduction of the experimental results.

REFERENCES

Anna Allen, Stratis Markou, Will Tebbutt, James Requeima, Wessel P. Bruinsma, Tom R. Andersson, Michael Herzog, Nicholas D. Lane, Matthew Chantry, J. Scott Hosking, and Richard E. Turner. End-to-end data-driven weather prediction. *Nature*, 641:1172–1179, 2025. doi: 10.1038/s41586-025-08897-0.

Gérôme Andry, Sacha Lewin, François Rozet, Omer Rochman, Victor Mangeleer, Matthias Pirlet, Elise Faulx, Marilaure Grégoire, and Gilles Louppe. Appa: Bending weather dynamics with latent diffusion models for global data assimilation. *arXiv preprint arXiv:2504.18720*, 2025.

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

Loïk Berre, Hubert Varella, and Gérald Desroziers. Modelling of flow-dependent ensemble-based background-error correlations using a wavelet formulation in 4d-var at météo-france. *Quarterly Journal of the Royal Meteorological Society*, 141(692):2803–2812, 2015.

Massimo Bonavita, Lars Isaksen, and Elías Hólm. On the use of eda background error variances in the ecmwf 4d-var. *Quarterly journal of the royal meteorological society*, 138(667):1540–1559, 2012.

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.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Hang Fan, Lei Bai, Ben Fei, Yi Xiao, Kun Chen, Yubao Liu, Yongquan Qu, Fenghua Ling, and Pierre Gentine. Physically consistent global atmospheric data assimilation with machine learning in latent space. *Science Advances*, 12(1):eaea4248, 2026.

Aayush Gupta, Akshay Subramaniam, Michael S Pritchard, Karthik Kashinath, Sergey Frolov, Kelsey Lieberman, Christopher Miller, Nicholas Silverman, and Noah D Brenowitz. Healda: Highlighting the importance of initial errors in end-to-end ai weather forecasts. *arXiv preprint arXiv:2601.17636*, 2026.

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.

P. L. Houtekamer and Fuqing Zhang. Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144(12):4489–4532, 2016. doi: 10.1175/MWR-D-15-0440.1.

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter D Dueben, and Torsten Hoefler. Diffda: a diffusion model for weather-scale data assimilation. In *Proceedings of International Conference on Machine Learning*, pp. 19798–19815, 2024.

Zeting Li, Wei Han, Xiaoze Xu, Xiuyu Sun, and Hao Li. All-sky microwave radiance observation operator based on deep learning with physical constraints. *Journal of Geophysical Research: Atmospheres*, 129(23):e2024JD042436, 2024.

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.

Scott A Martin, Georgy E Manucharyan, and Patrice Klein. Generative data assimilation for surface ocean state estimation from multi-modal satellite observations. *Journal of Advances in Modeling Earth Systems*, 17(8):e2025MS005063, 2025.

Yongquan Qu, Juan Nathaniel, Shuolin Li, and Pierre Gentine. Deep generative data assimilation in multimodal setting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 449–459. IEEE, 2024.

François Rozet and Gilles Louppe. Score-based data assimilation. *Advances in Neural Information Processing Systems*, 36:40521–40541, 2023.

François Rozet, Gérôme Andry, François Lanusse, and Gilles Louppe. Learning diffusion priors from observations by expectation maximization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2783.

François Rozet, Ruben Ohana, Michael McCabe, Gilles Louppe, François Lanusse, and Shirley Ho. Lost in latent space: An empirical study of latent diffusion models for physics emulation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4492.

Aliaksandra Shysheya, Cristiana Diaconu, Federico Bergamin, Paris Perdikaris, José Miguel Hernández-Lobato, Richard E. Turner, and Emile Mathieu. On conditional diffusion models for PDE simulations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0732.

Phillip Si and Peng Chen. Latent-ensf: A latent ensemble score filter for high-dimensional data assimilation with sparse observation data. In *International Conference on Learning Representations*, volume 2025, pp. 94879–94895, 2025.

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, and LEI BAI. Align-DA: Align score-based atmospheric data assimilation with multiple preferences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

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*, 2025b.

Xiuyu Sun, Xiaohui Zhong, Xiaoze Xu, Yuanqing Huang, Hao Li, J. David Neelin, Deliang Chen, Jie Feng, Wei Han, Libo Wu, and Yuan Qi. A data-to-forecast machine learning system for global weather. *Nature Communications*, 16:6658, 2025c. doi: 10.1038/s41467-025-62024-1.

Pengpeng Xiao, Phillip Si, and Peng Chen. LD-EnSF: Synergizing latent dynamics with ensemble score filters for fast data assimilation with sparse observations. In *International Conference on Learning Representations*, 2026.

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, 2025. doi: 10.1038/s41612-025-01039-3.

Xiaoze Xu, Xiuyu Sun, Songling Zhu, Xiaohui Zhong, Yuanqing Huang, Zijian Zhu, Jun Liu, and Hao Li. Fuxiweather2: Learning accurate atmospheric state estimation for operational global weather forecasting. *arXiv preprint arXiv:2603.15358*, 2026.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

A SUPPLEMENTARY EXPERIMENTAL RESULTS

A.1 COMPARISON WITH APPA

Across all 69 channels in Figure 4, PPGF+VCLG reduces RMSE relative to Appa, whereas the gains from PPGF^{SD}+VCLG vary with variable and pressure level.

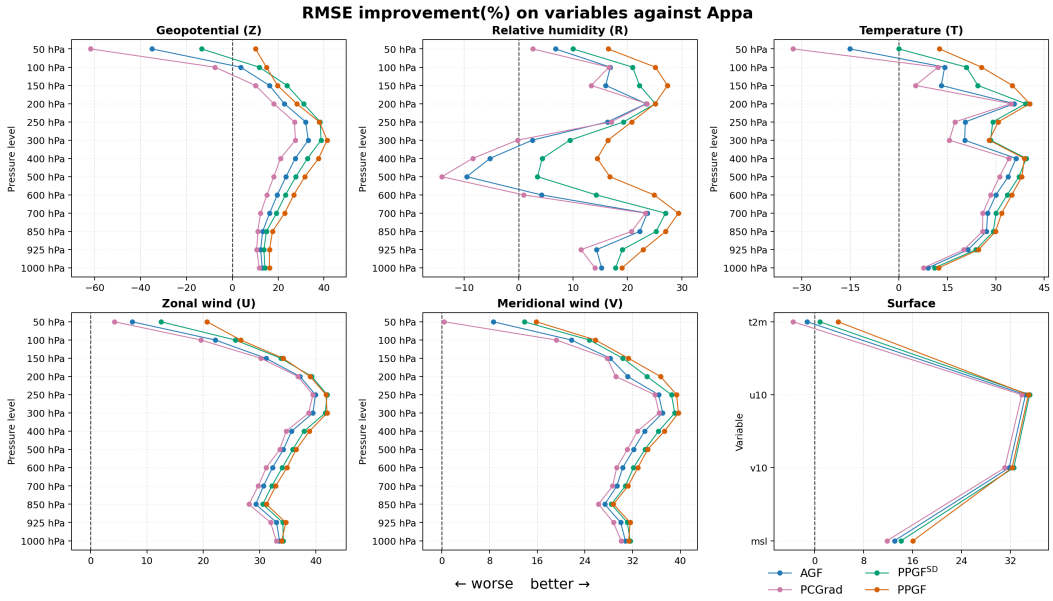


Figure 4: Relative RMSE reductions for AGF+VCLG, PCGrad+VCLG, PPGF^{SD}+VCLG, and PPGF+VCLG against Appa across all 69 atmospheric channels over April–June 2025. The plotted labels omit the shared VCLG component. The horizontal axis shows $(\mathrm{RMSE}_{\mathrm{Appa}} - \mathrm{RMSE}_{\mathrm{method}})/\mathrm{RMSE}_{\mathrm{Appa}} \times 100$, using latitude-weighted RMSE against ERA5. Negative values indicate higher RMSE than Appa.

A.2 Spatial RMSE comparison with Appa

Figures 5–7 support the spatial comparison in Section 4.2. Each map reports RMSE at each grid point over the April–June 2025 evaluation times. Within each variable, Appa and PPGF+VCLG share a color scale; the plotted PPGF label omits the VCLG component. Darker colors indicate lower absolute RMSE.

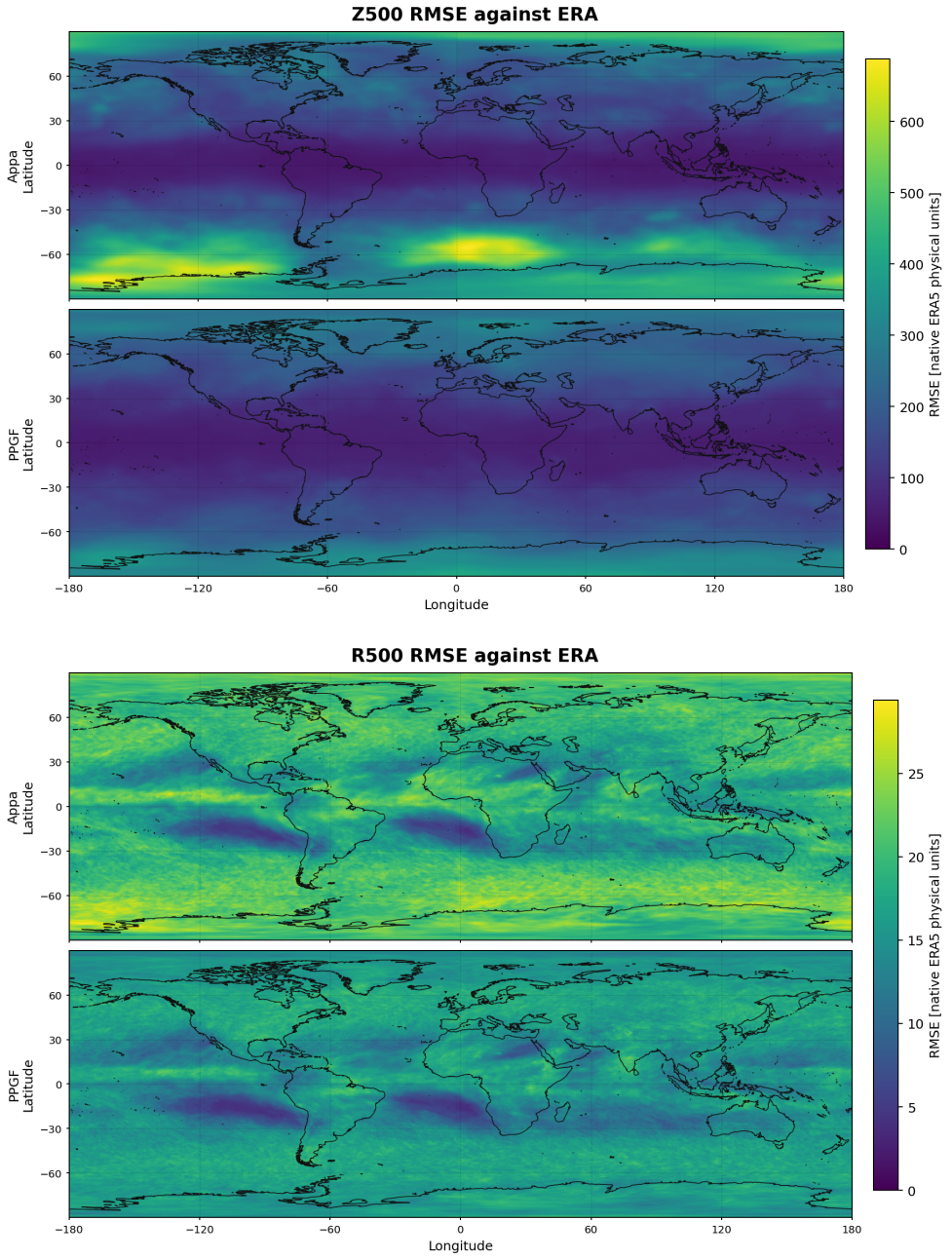


Figure 5: Spatial RMSE against ERA5 over April–June 2025 for 500-hPa geopotential (Z500; $\mathrm{m}^2\,\mathrm{s}^{-2}$) and relative humidity (R500; %). Within each variable, Appa is shown above and PPGF+VCLG below, with a shared color scale.

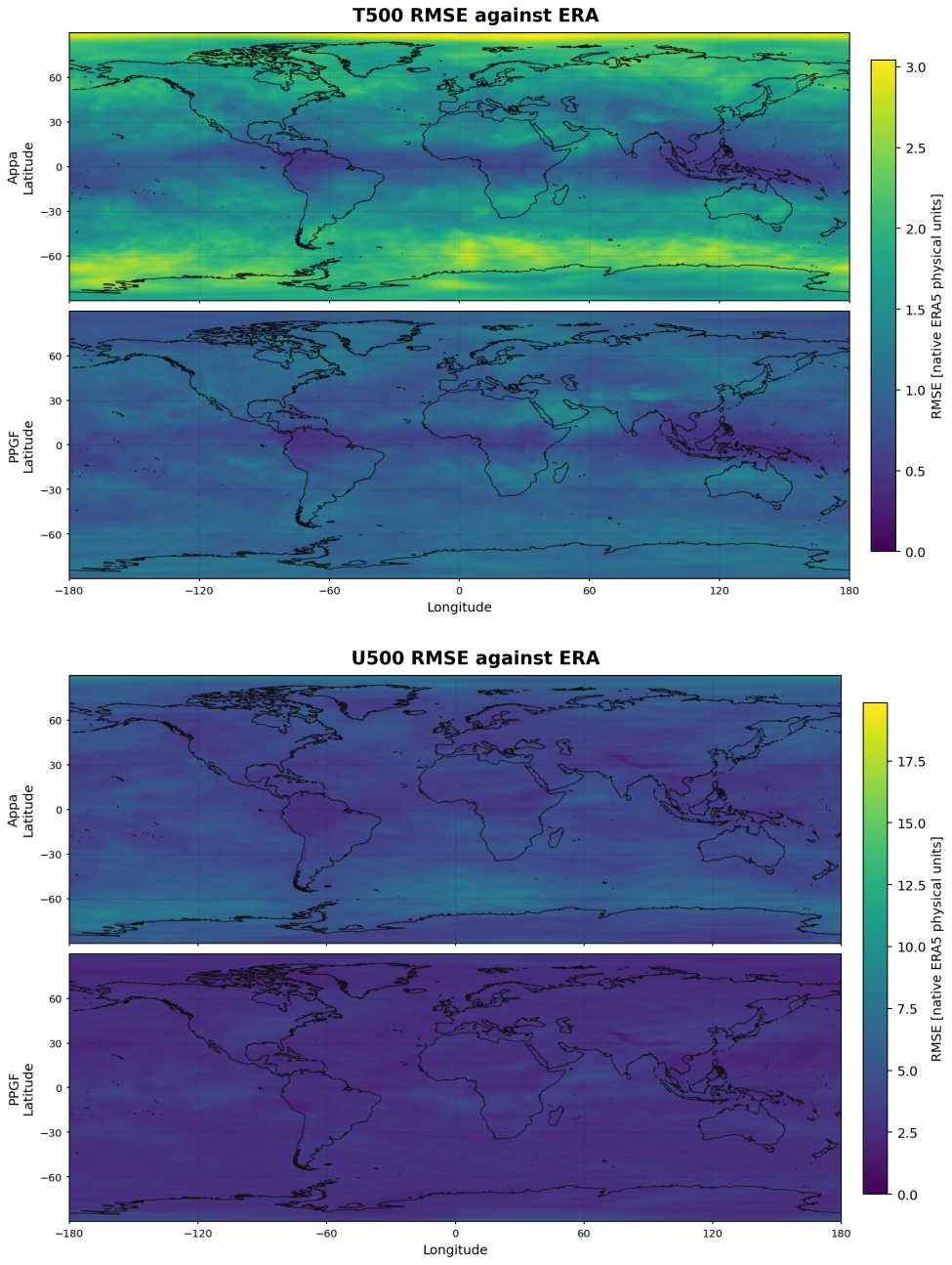


Figure 6: Spatial RMSE against ERA5 over April–June 2025 for 500-hPa temperature (T500; K) and zonal wind (U500; $\mathrm{m\,s^{-1}}$). Within each variable, Appa is shown above and PPGF+VCLG below, with a shared color scale.

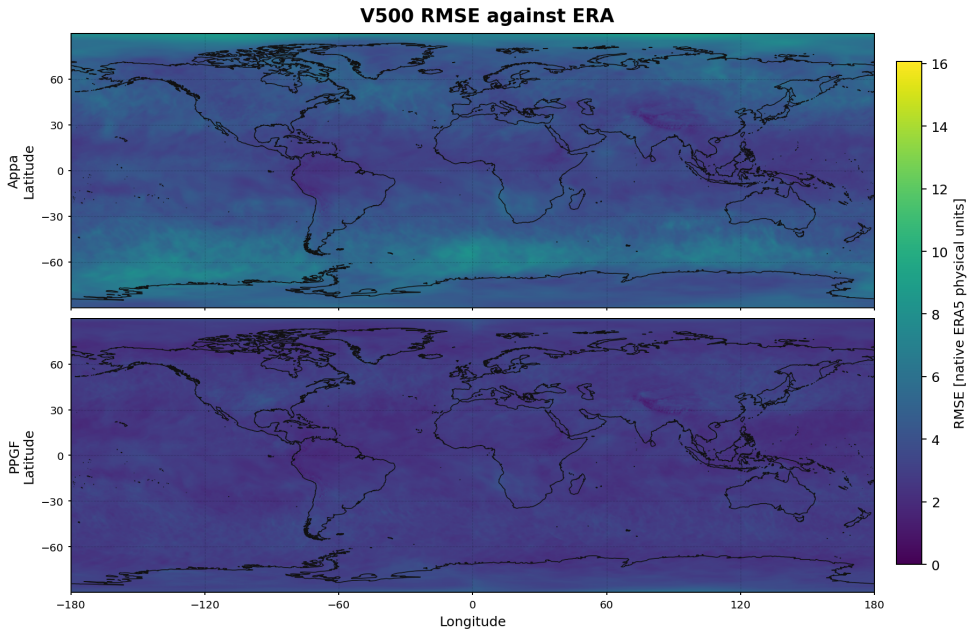
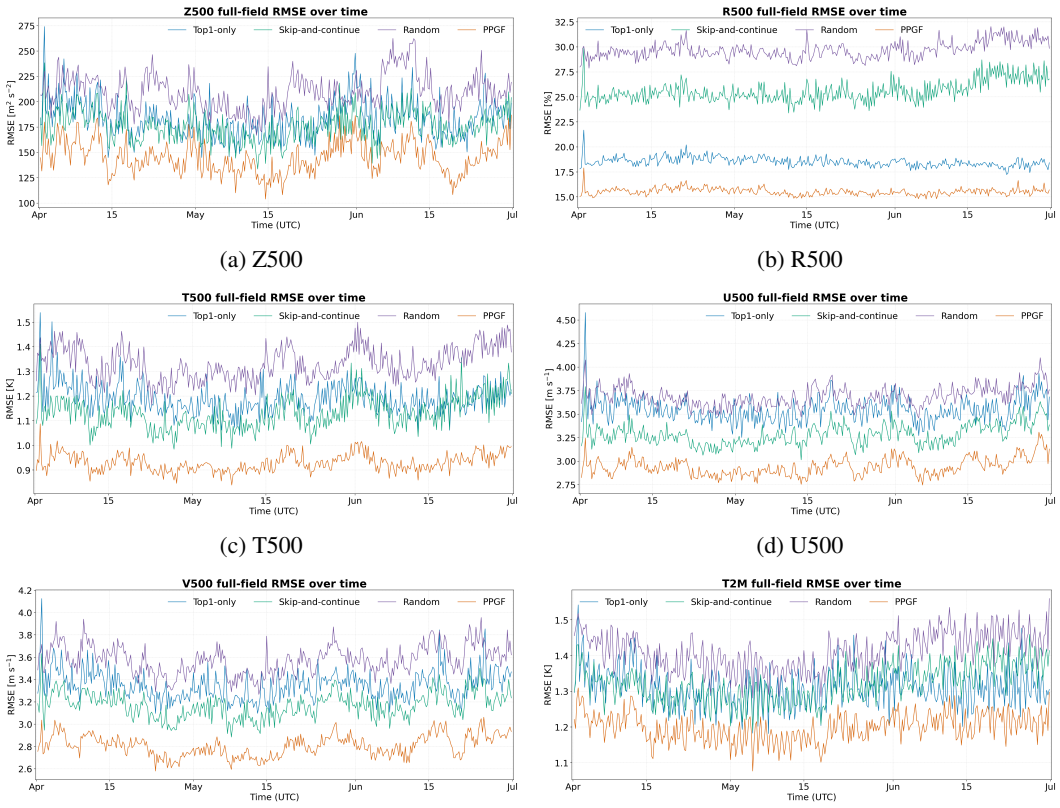


Figure 7: Spatial RMSE against ERA5 over April–June 2025 for 500-hPa meridional wind (V500; $\mathrm{m\,s^{-1}}$). Within each variable, Appa is shown above and PPGF+VCLG below, with a shared color scale.

A.3 GRADIENT FUSION STRATEGIES

The time series in Figure 8 complement the gradient fusion ablation in Section 4.3. They compare PPGF+VCLG with Random ordering, Top-1 only, and Skip-and-continue, all using the same spatially resolved VCLG variances $v^i_{p,c}$ as in Figure 3, over April–June 2025 for Z500, R500, T500, U500, V500, and T2M. Each curve reports latitude-weighted spatial RMSE against ERA5 at each evaluation time (UTC).



(a) Z500 (b) R500

(c) T500 (d) U500

(e) V500 (f) T2M

Figure 8: Temporal RMSE comparison over April–June 2025 for (a) 500-hPa geopotential (Z500; $\mathrm{m}^2\,\mathrm{s}^{-2}$), (b) 500-hPa relative humidity (R500; %), (c) 500-hPa temperature (T500; K), (d) 500-hPa zonal wind (U500; $\mathrm{m\,s}^{-1}$), (e) 500-hPa meridional wind (V500; $\mathrm{m\,s}^{-1}$), and (f) 2-m temperature (T2M; K). Curves compare Random ordering, Top-1 only, Skip-and-continue, and PPGF+VCLG against ERA5. All four settings share spatially resolved VCLG variances, as in Figure 3; VCLG is omitted from the plotted labels. The horizontal axis gives evaluation times (UTC); lower RMSE indicates better analysis accuracy.

B BACKGROUND ON SCORE-BASED DATA ASSIMILATION

Let $\boldsymbol{x}$ denote the atmospheric state and $\boldsymbol{y}$ the observations. Data assimilation combines a prior $p(\boldsymbol{x})$ with an observation likelihood $p(\boldsymbol{y} \mid \boldsymbol{x})$ to infer the posterior $p(\boldsymbol{x} \mid \boldsymbol{y}) \propto p(\boldsymbol{x})p(\boldsymbol{y} \mid \boldsymbol{x})$.

Score-based data assimilation introduces observations during sampling while keeping the learned atmospheric prior fixed (Rozet & Louppe, 2023; Andry et al., 2025; Manshausen et al., 2025). At diffusion noise level σ, let $\boldsymbol{x}_\sigma = \mu_\sigma \boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and μ_σ is the signal coefficient. Bayes' rule gives the posterior-score decomposition

$$\nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma \mid \boldsymbol{y}) = \nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma) + \nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{y} \mid \boldsymbol{x}_\sigma). \tag{12}$$

A pretrained score-based diffusion model approximates the first term, while the second provides observation-likelihood guidance.

The observation likelihood is commonly modeled as Gaussian (Andry et al., 2025; Sun et al., 2025b), $p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid H(\boldsymbol{x}), \boldsymbol{R})$, where H is a differentiable observation operator and $\boldsymbol{R}$ is the observation-error covariance. During diffusion sampling, however, the conditional variance of $\boldsymbol{x}_0$ given $\boldsymbol{x}_\sigma$ generally cannot be neglected at nonzero noise levels (Rozet & Louppe, 2023). Following SDA, we therefore account for the uncertainty in estimating the clean state $\boldsymbol{x}_0$ from the noisy state $\boldsymbol{x}_\sigma$ when approximating the observation likelihood. Specifically, we adopt the moment matching posterior sampling (MMPS) approximation (Rozet et al., 2024) and linearize H as in Appa (Andry et al., 2025), obtaining

$$p(\boldsymbol{y} \mid \boldsymbol{x}_\sigma) \approx \widetilde{p}(\boldsymbol{y} \mid \boldsymbol{x}_\sigma) := \mathcal{N}\big(\boldsymbol{y} \;\big|\; H(\widehat{\boldsymbol{x}}_0), \boldsymbol{R} + \boldsymbol{A}\, \mathbb{V}[\boldsymbol{x}_0 \mid \boldsymbol{x}_\sigma]\, \boldsymbol{A}^\top\big)\,. \tag{13}$$

Here, $\widehat{\boldsymbol{x}}_0 = \mathbb{E}[\boldsymbol{x}_0 \mid \boldsymbol{x}_\sigma]$ is the denoised estimate provided by the diffusion model, $\mathbb{V}[\boldsymbol{x}_0 \mid \boldsymbol{x}_\sigma]$ denotes the conditional covariance, and $\boldsymbol{A}$ is the Jacobian of H evaluated at $\widehat{\boldsymbol{x}}_0$.

C Training and implementation details

Observation-to-latent (O2L) encoder. Each observation encoder uses the atmospheric encoder backbone with an input layer adapted to its source. Omitting the source index for clarity, we train $E^{\boldsymbol{y}}$ through the frozen atmospheric decoder $D^{\boldsymbol{x}}$. For satellite observations, the reconstruction $\widehat{\boldsymbol{x}} = D^{\boldsymbol{x}}(E^{\boldsymbol{y}}(\boldsymbol{y}))$ is supervised by the time-aligned ERA5 atmospheric state. For CONVI, the input consists of four surface observation variables, and supervision is restricted to the corresponding ERA5 surface variables: 2-m temperature, 10-m zonal and meridional winds, and mean sea-level pressure. We use the masked mean squared error

$$\mathcal{L} = \frac{\sum_j M_j (\widehat{x}_j - x_j)^2}{\sum_j M_j}, \tag{14}$$

where j indexes the supervised grid-point and variable entries, and $M_j \in \{0, 1\}$ indicates valid observation coverage for that entry. The coverage mask is broadcast to the supervised atmospheric channels, and the denominator counts valid supervised entries. With the atmospheric decoder fixed, this reconstruction supervision promotes implicit alignment with the atmospheric latent space.

Latent-to-observation (L2O) operator. For satellite observations, we train a source-specific operator P that maps atmospheric latents $\boldsymbol{z} = E^{\boldsymbol{x}}(\boldsymbol{x})$ and viewing information to brightness temperatures. It uses the atmospheric decoder backbone with an output head adapted to the instrument's observation channels:

$$\widehat{\boldsymbol{y}} = P(\boldsymbol{z}, \boldsymbol{a}), \tag{15}$$

where $\boldsymbol{a}$ denotes the satellite zenith angle. For CONVI, the operator predicts the four corresponding surface observation variables without satellite viewing information. We train each operator using

$$\mathcal{L}_{\text{op}} = \frac{\sum_j M_j (\widehat{y}_j - y_j)^2}{\sum_j M_j}, \tag{16}$$

where j indexes observation-grid and channel entries, and the binary mask M_j selects valid observations. During sampling, the L2O baseline evaluates the observation-space likelihood using Equation 2 and backpropagates through P and the denoiser to guide the sampled latent.

Training hyperparameters All six observation encoders are optimized for 100,000 updates using AdamW with an initial learning rate of 3×10^{-5}, a batch size of 8, and linear learning-rate decay. We use $(\beta_1, \beta_2) = (0.9, 0.99)$ and no weight decay. The learned L2O operators use the same optimizer, learning rate, decay schedule, batch size, and number of updates. They are trained with Equation 16, keeping the atmospheric encoder frozen.

Observation processing. Raw satellite brightness-temperature observations are gridded onto the atmospheric analysis grid. Their valid-coverage masks are pooled to the latent resolution to construct the mask used by the latent observation likelihood. Missing observations are excluded from calibration and likelihood evaluation.

Trajectories and sampling. All evaluated settings, including the adapted Appa baseline, use a six-hour assimilation window with one-hour temporal resolution. We evaluate the fourth frame of each trajectory at the six-hourly synoptic times specified in Section 4.1. Sampling uses `PCSample` with 64 predictor–corrector steps. Each step consists of one predictor update and two corrector updates. All remaining sampling settings follow Appa (Andry et al., 2025).

RMSE aggregation. For an atmospheric channel c, let $\widehat{x}_{n,s,c}$ and $x^{\mathrm{ERA5}}_{n,s,c}$ denote the evaluated analysis and reference at grid location s for sample n, and let w_s denote the latitude weight. The reported RMSE is

$$\mathrm{RMSE}_c = \left[\frac{\sum_{n=1}^{N} \sum_s w_s (\widehat{x}_{n,s,c} - x^{\mathrm{ERA5}}_{n,s,c})^2}{N \sum_s w_s} \right]^{1/2}. \tag{17}$$

All errors are evaluated in physical units.

D Additional diagnostics of observation errors

We extend the MHS analysis in Section 3.2.2 to ATMS, AMSU-A, MWTS, MWHS, and conventional surface observations (CONVI). Following Equation 7, the errors compare observation and ERA5 latents after applying the same ERA5-based per-channel standardization. Only latent positions with valid observation coverage are included. Figure 9 combines error histograms and Gaussian Q–Q plots for two channels per source with Pearson correlation matrices across all 128 latent error channels.

The displayed channels have approximately Gaussian error distributions. CONVI exhibits a similar pattern to the satellite sources, with generally weak off-diagonal latent error correlations. These diagnostics motivate the diagonal approximation used for latent variance calibration.

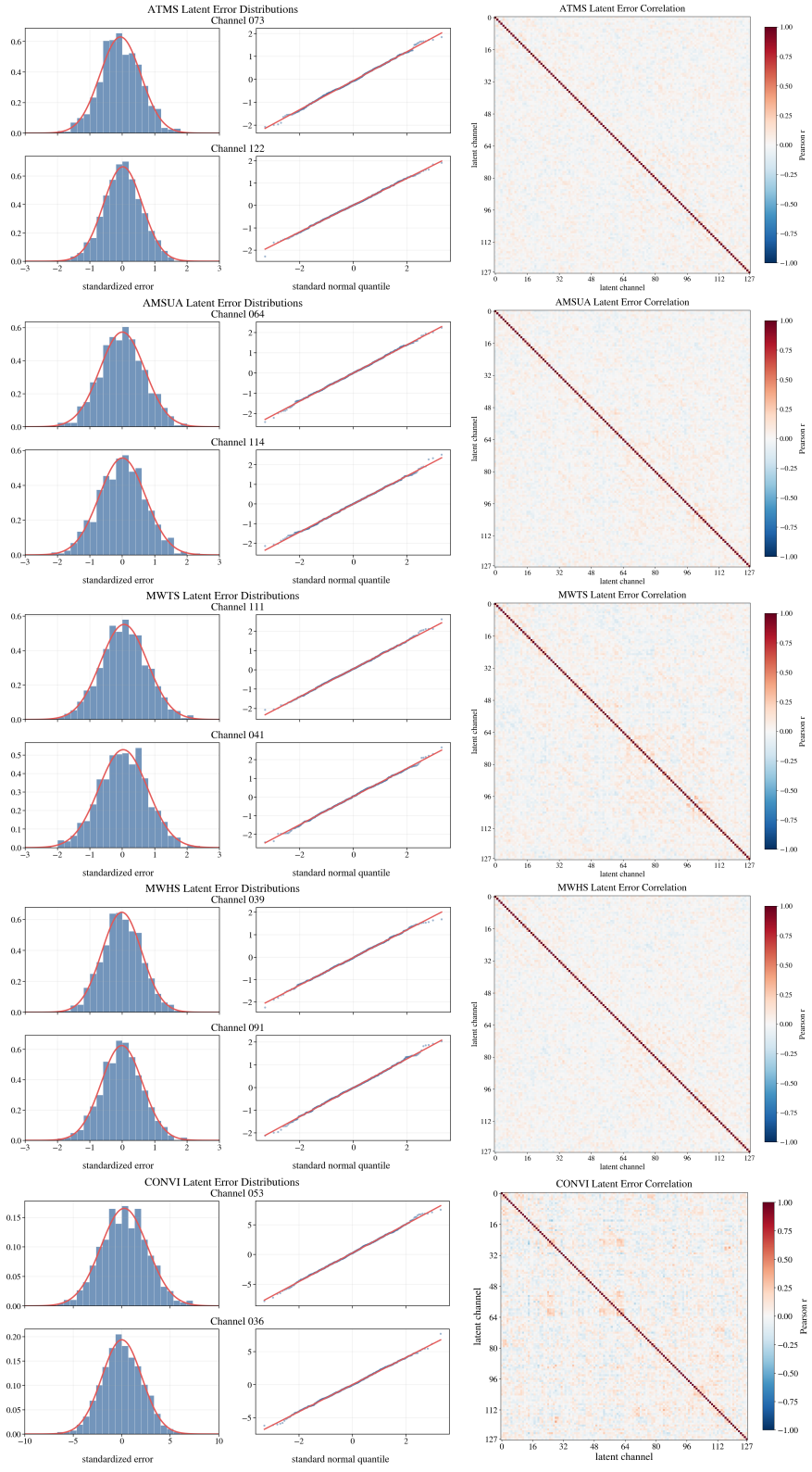


Figure 9: Latent-error diagnostics for ATMS, AMSU-A, MWTS, MWHS, and CONVI (top to bottom). Left: error histograms, Gaussian fits, and Q–Q plots for channels 73/122, 64/114, 111/41, 39/91, and 53/36, respectively. Right: Pearson correlations across all 128 latent error channels.

E CALIBRATION STATISTICS AND SPATIAL SOURCE PRIORITIES

Calibration statistics. We estimate uncertainty and reliability from observation–ERA5 latent errors over January–March 2025, before the April–June evaluation period. For a fixed source and latent channel, let $\mathcal{S}$ be the collection of valid calibration errors and let $N = |\mathcal{S}|$. Channel-wise statistics pool the valid calibration errors across spatial positions; spatially resolved statistics restrict this collection to the corresponding latent position p. In Table 1, AGF+VCLG and PCGrad+VCLG use channel-wise variances v_c^i shared across spatial positions, whereas PPGF^{SD}+VCLG and PPGF+VCLG use spatially resolved variances $v_{p,c}^i$. All settings in the fusion ablation use the same spatially resolved variances. Missing observations are excluded rather than treated as zero errors. Omitting the source, channel, and spatial indices, we assemble the errors following Equation 7 and compute

$$\bar{e} = \frac{1}{N}\sum_{e\in\mathcal{S}} e, \qquad v = \frac{1}{N}\sum_{e\in\mathcal{S}}(e-\bar{e})^2, \qquad q = \frac{1}{N}\sum_{e\in\mathcal{S}} e^2. \tag{18}$$

The centered variance v calibrates the corresponding diagonal entry of the VCLG observation-error covariance $\boldsymbol{R}_{O2L}^i$, whereas the MSE q determines source priority.

Spatial variation in source priorities. Figure 10 and Figure 11 compare the source-wise priorities at two latent-grid locations, $(h, w) = (19, 28)$ and $(1, 2)$, respectively. The matrices visualize source priorities based on residual MSE, with blue indicating higher priority and orange indicating lower priority. The most reliable source varies across both channels and locations. For example, CONVI has the lowest MSE in several channels at $(19, 28)$, whereas satellite sources dominate the rankings at $(1, 2)$. This variation motivates region- and channel-specific ordering rather than a single source order shared across all locations.

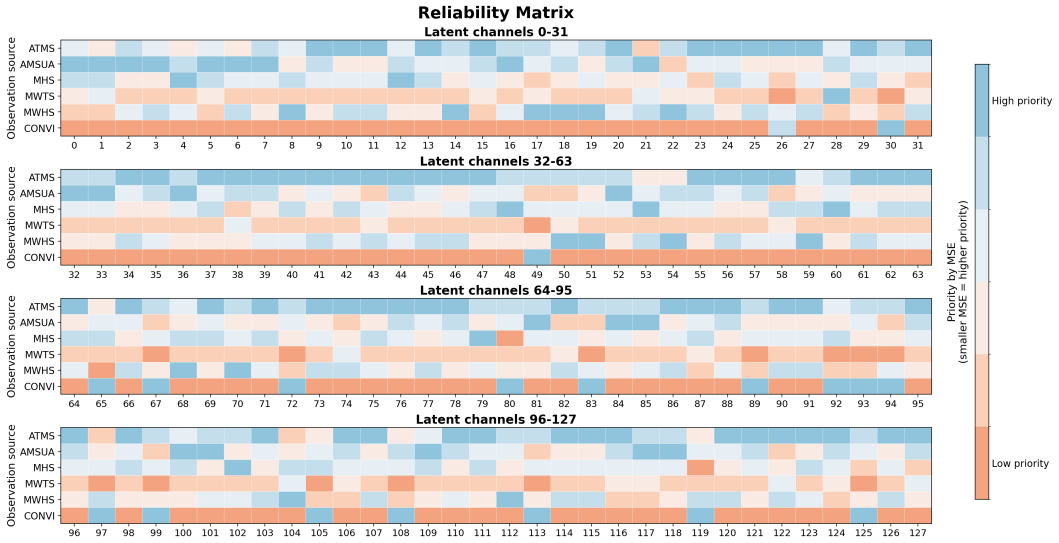


Figure 10: Source-importance matrix at latent-grid position $(h, w) = (19, 28)$, showing source priorities based on observation–ERA5 latent residual MSE. Rows correspond to the six observation sources and columns to the 128 latent channels, displayed in four groups of 32. Blue indicates higher priority (lower MSE), and orange indicates lower priority (higher MSE).

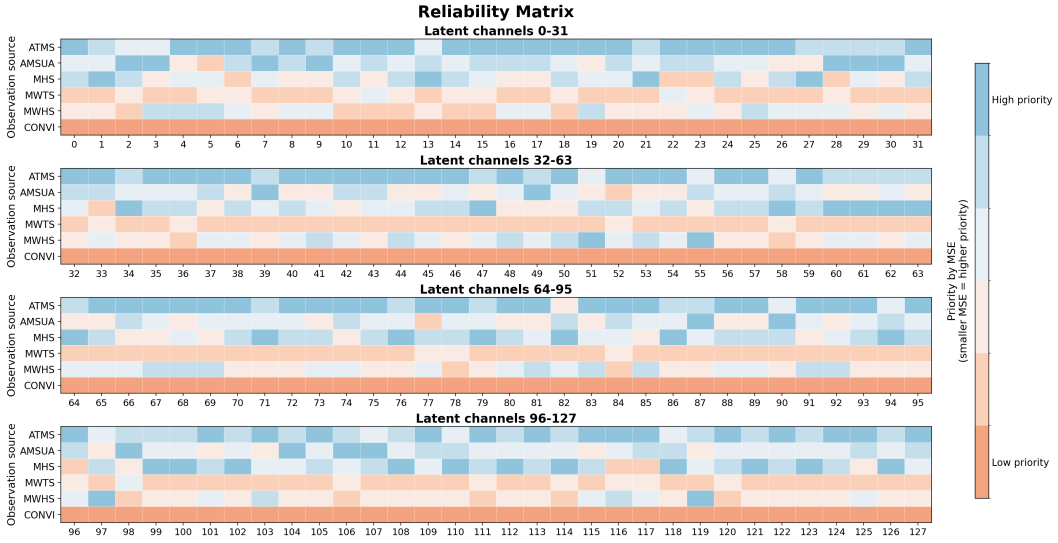


Figure 11: Source-importance matrix at latent-grid position $(h, w) = (1, 2)$, showing source priorities based on observation–ERA5 latent residual MSE. Rows correspond to the six observation sources and columns to the 128 latent channels, displayed in four groups of 32. Blue indicates higher priority (lower MSE), and orange indicates lower priority (higher MSE).