FERA: FORECAST-CONDITIONED ERROR REPRESENTATIONS FOR DATA ASSIMILATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data assimilation is fundamental to weather prediction because forecasts must be continually corrected using observations that constrain only a small fraction of the atmospheric state. The central challenge is not merely fitting these observations, but propagating their information to unobserved locations and variables through a plausible representation of forecast uncertainty. Classical climatological covariances are stable but cannot adapt their correction geometry to the current flow, while small ensembles provide flow dependence only through a rank-limited sample. Neural methods offer more expressive representations, but often perform assimilation directly in learned latent spaces or retain nonlinear neural decoders inside the online optimization, making the connection between observations, physical corrections, and inference less explicit. We introduce FERA (Forecast-conditioned Error Representation for Assimilation), which separates learned error representation from online Bayesian inference. A forecast-conditioned neural transform constructs multivariate physical-space correction patterns, while held-out-variable and sparse-observation objectives train these patterns to recover errors from partial information. During assimilation, the neural transform is frozen: observations determine the correction coefficients through a linear-Gaussian Bayesian update, while the ensemble refines their prior uncertainty and a full-rank historical component preserves directions beyond the member span. For linear observations, this yields a unique, initialization-free analysis with an explicit observation-to-correction map, while retaining forecast-adaptive error structure across cycles. In month-long global cycling experiments with a 69-channel AI weather model, FERA outperforms the strongest matched baseline in 46 of 48 ensemble-size–diagnostic comparisons. In a fixed-input replay, FERA shows lower numerical sensitivity to equivalent observation reorderings than the tested nonlinear neural variational baseline, without optimizer-initialization dependence. Ablation studies further demonstrate the effectiveness of both forecast conditioning and partial-information training for sustained cycling. Together, these results link forecast-adaptive error learning to an explicit Bayesian analysis, providing a basis for reliable and interpretable assimilation.

1 INTRODUCTION

Data assimilation (DA) combines imperfect forecasts with incomplete observations to estimate the evolving state of a dynamical system (Kalnay, 2003; Carrassi et al., 2018). In numerical weather prediction, this problem is particularly challenging: observations cover only a small fraction of the atmosphere, yet the analysis must update a high-dimensional, multivariate state whose unobserved components strongly affect subsequent forecasts. The central question is therefore not merely how to fit available observations, but how observational information should be propagated across space, variables, and scales to produce a plausible full-state correction. This problem remains central as AI weather forecasting models become increasingly accurate (Lam et al., 2023; Chen et al., 2023a; Price et al., 2025), because their future skill still depends on the quality of the state from which each forecast is initialized.

Classical DA addresses this problem through a background-error covariance. Climatological covariances provide statistically stable correction structures but are largely fixed across atmospheric states (Courtier et al., 1998; Bannister, 2008). Ensemble methods instead estimate flow-dependent

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.

uncertainty from forecast members (Evensen, 1994; Houtekamer & Mitchell, 1998; Hunt et al., 2007), but a practical ensemble is necessarily small relative to the atmospheric state dimension. Its covariance is therefore rank limited and must infer both which error directions are relevant and how strongly they should be weighted from only a small set of samples. Hybrid methods alleviate this limitation by combining climatological and ensemble covariances (Hamill & Snyder, 2000; Buehner, 2005; Penny, 2014), but the historical component remains largely state independent. These limitations motivate a richer alternative: learning a correction space from historical forecast errors whose geometry itself adapts to the current forecast, while using the current ensemble only to refine uncertainty within that space.

Neural DA provides new ways to construct such representations, but existing approaches also introduce a second question: *in which space should assimilation itself be performed?* A growing class of methods compresses atmospheric states into a learned latent space and performs inference there (Melinc & Zaplotnik, 2024; Li et al., 2024; Si & Chen, 2025; Fan et al., 2026; Si & Chen, 2026). This can substantially reduce dimensionality, but a compact state representation is not automatically an assimilation-optimal representation. Unless the learned coordinates preserve the posterior geometry relevant to the original state, optimizing a latent-space objective need not be equivalent to optimizing the corresponding physical-space Bayesian objective. Moreover, corrections expressed only through latent coordinates can make it difficult to identify which physical error structures are being activated by individual observations. For data assimilation, where the observation operator and the background-error prior have explicit probabilistic meanings, retaining a transparent connection between coefficients, physical-space corrections, and observations is desirable.

A second family of neural methods models forecast errors themselves. VAE-Var learns a nonlinear decoder of forecast-error increments, while LoRA-EnVar further adapts that decoder using the current ensemble (Xiao et al., 2025b;a). These methods avoid reconstructing the full atmospheric state in latent space, but their nonlinear decoder remains part of the online optimization problem. Consequently, the mapping from latent coefficients to physical corrections changes with the coefficients themselves, and each analysis requires iterative optimization through a high-dimensional neural model. The resulting finite-budget solution can depend on initialization, optimization path, and numerical details. More generally, an expressive neural error representation does not by itself solve the assimilation problem: a representation that reconstructs complete forecast errors accurately may still contain directions that are difficult to infer from the sparse observations available at analysis time.

We introduce **FERA** (**F**orecast-conditioned **E**rror **R**epresentation for **A**ssimilation), which separates these roles. Given the current forecast background, a neural transform constructs a set of multivariate physical-space error patterns. The network therefore determines *which corrections are available* for the current atmospheric state, but it does not determine the online analysis coefficients. Instead, observations select a linear combination of these patterns. We place an explicit Gaussian prior on the coefficients and use historical and current-ensemble statistics to determine their uncertainty. For a fixed background and linear observation operator, the resulting inference problem is a linear-Gaussian Bayesian update: the posterior precision is positive definite, the analysis has a unique solution, and no optimizer initialization or neural-network adaptation is required during coefficient inference. The learned representation remains nonlinear in the background across cycles, while each individual analysis update is linear and directly solvable.

This separation also changes how the representation should be trained. Full-field reconstruction asks whether forecast errors lie near the learned space, but assimilation asks a harder question: whether useful coefficients can be inferred from *partial* evidence. A direction can improve full-field reconstruction while being almost invisible on the observed support and producing large, poorly constrained corrections elsewhere. We therefore train FERA not only to represent historical forecast errors, but also to recover them from restricted information. Held-out-variable objectives test cross-variable transfer, while sparse-observation objectives infer coefficients from a small support set and evaluate reconstruction at disjoint unobserved targets. These objectives explicitly train the observation-to-unobserved-state transfer that the representation must support during assimilation.

FERA therefore separates three quantities that are often coupled in existing methods: the current forecast determines the geometry of the correction space, historical and ensemble statistics determine uncertainty within that space, and observations determine the realized correction. This design preserves the flexibility of learned error representations while retaining an explicit physical-space interpretation and a tractable Bayesian analysis.

Our main contributions are:

1. **A forecast-conditioned physical error representation with explicit Bayesian inference.** We learn background-adaptive multivariate correction patterns rather than a fixed climatological covariance or a state latent space. Given these patterns, the analysis is formulated through an explicit coefficient prior and observation likelihood.
2. **Representation learning targeted at assimilation from partial information.** Beyond full-error reconstruction, we introduce held-out-variable and sparse-support objectives that directly train the learned space for recovering useful corrections from incomplete observations.
3. **A direct Bayesian analysis validated in global cycling.** For linear observations, FERA's fixed-transform solve removes optimizer-initialization dependence and makes the observation-to-correction map explicit. Fixed-input replay supports numerical robustness to equivalent observation reorderings. In global cycling, FERA improves over matched classical and neural baselines; gains persist into subsequent forecasts, and retraining ablations support both representation choices.

2 Problem Setting and Design Requirements

Assimilation as conditional forecast-error inference. Let $x^b, x \in \mathbb{R}^n$ denote the forecast background and unknown state, and let $y \in \mathbb{R}^{n_o}$ contain observations. A linear operator $H \in \mathbb{R}^{n_o \times n}$ maps states to observed quantities. Conditional independence of observations and background information given x yields (Kalnay, 2003; Carrassi et al., 2018) $p(x \mid x^b, y) \propto p(y \mid x)\, p(x \mid x^b)$. Writing the forecast error as $e = x - x^b$ gives $p(e \mid x^b, y) \propto p(y \mid x^b + e)\, p(e \mid x^b)$. This factorization separates two roles. The conditional prior $p(e \mid x^b)$ describes which full-state forecast errors are plausible for the current background, while the likelihood $p(y \mid x^b + e)$ measures how well a candidate error explains the available observations. Data assimilation combines the two to infer a correction. For an observation model $y = H(x^b + e) + \epsilon$ with observation error $\epsilon \in \mathbb{R}^{n_o}$, sparse or partial observations constrain only the projection He. A useful representation must therefore both represent plausible full-state errors and make their relevant degrees of freedom recoverable from observations. Accurate complete-error reconstruction alone does not ensure useful assimilation if distinct corrections have nearly identical observed projections.

What should a learned error representation provide? The conditional formulation above suggests three requirements. *Background adaptivity.* Because the relevant prior is $p(e \mid x^b)$, the geometry of plausible corrections should be allowed to vary with the current forecast rather than remain fixed across backgrounds. *Recoverability from partial information.* Good full-field reconstruction is not sufficient: coefficients inferred from the observed support should imply useful corrections at unobserved locations and variables. *Explicit and tractable Bayesian inference.* The online update should preserve a clear relationship among the error prior, observation operator, and physical-space correction. In particular, when the prior is Gaussian and observations are linear, it is desirable to retain a well-defined linear-Gaussian update rather than introduce a new nonlinear neural optimization problem at every analysis.

Complementary ensemble information. A current forecast ensemble provides additional flow-dependent uncertainty information, but with M members its covariance has rank at most $M - 1$. For atmospheric states whose dimension is vastly larger than practical ensemble sizes, it is therefore useful to distinguish *which correction patterns are available* from *how uncertain those patterns are at the current cycle*. This motivates using an ensemble to refine the current weighting of a richer learned representation, rather than requiring the ensemble itself to span the full correction space.

3 FERA

Given a forecast background x^b, a context map C produces $c = C(x^b)$, and a neural transform $U_\theta(c) \in \mathbb{R}^{n \times d}$ produces d physical-space correction patterns. Here θ denotes the trainable network parameters. With coefficients $\eta \in \mathbb{R}^d$ and prior covariance $B_z \in \mathbb{R}^{d \times d}$, FERA represents conditional forecast errors as

$$e = U_\theta(c)\eta, \qquad \eta \sim \mathcal{N}(0, B_z), \qquad c = C(x^b), \qquad B_x(x^b) \approx U_\theta(c) B_z U_\theta(c)^\top, \tag{1}$$

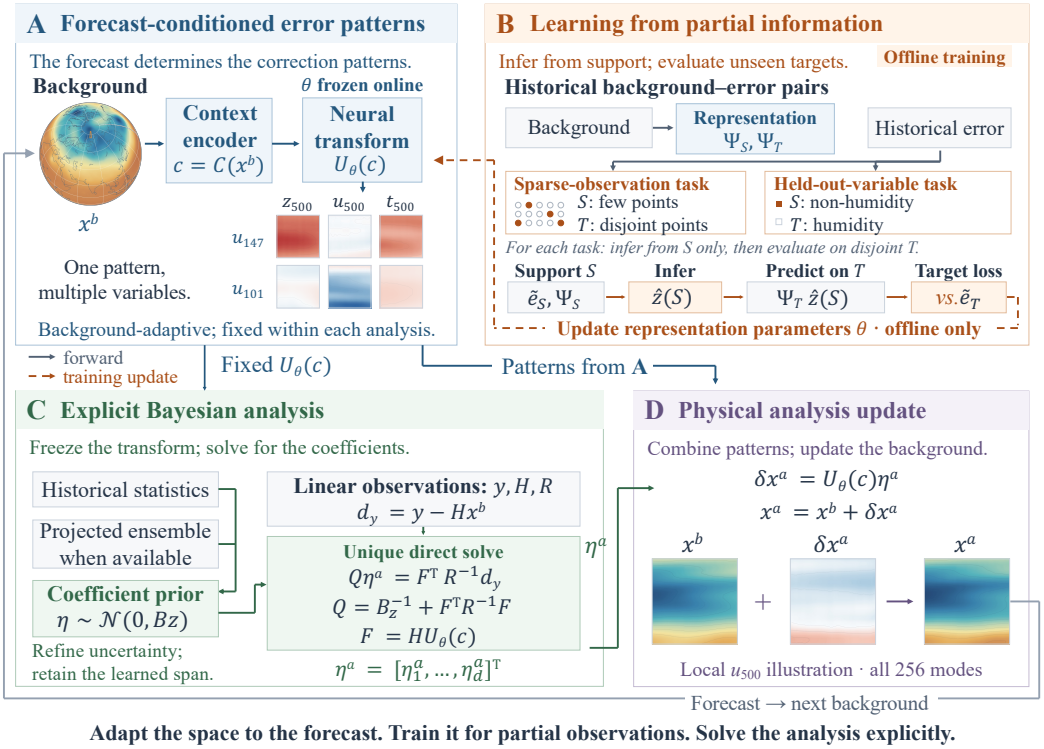


Figure 1: **FERA separates representation learning from the online analysis.** (A) The forecast background conditions multivariate correction patterns, which remain fixed within an analysis. (B) Offline tasks infer from partial support and evaluate on disjoint targets to train the representation. (C) Historical statistics and, when available, a projected ensemble define the coefficient prior; linear observations then yield a direct solve with the transform frozen. (D) The solved coefficients combine all modes into a physical increment and updated state. The field thumbnails illustrate one local diagnostic; the workflow arrows are schematic. Appendix A.7 gives the illustration provenance.

where $B_x(x^b) = \mathrm{Cov}(e \mid x^b)$ is the conditional physical-space error covariance. For fixed c, the coefficients select a correction in $\mathrm{range}(U_\theta(c))$; changing the background can change this space.

The superscript a denotes the analysis. With the transform fixed by the background, observations determine the coefficients and state increment:

$$\delta x^a = \underbrace{U_\theta(c)}_{\text{forecast-conditioned patterns}} \underbrace{\eta^a}_{\text{inferred coefficients}}, \qquad x^a = x^b + \delta x^a. \tag{2}$$

The coefficient prior determines how the available patterns are weighted before observations select their combination. This separation is the central interface of FERA: the forecast determines the correction geometry, historical and ensemble statistics determine uncertainty within that geometry, and observations determine the realized correction. Figure 1 shows how the representation is learned offline and used for the direct online analysis.

3.1 Forecast-Conditioned Physical Error Patterns

A scalar query $q = (\varphi, \lambda, k)$ specifies latitude, longitude, and a variable–level channel. Given context c, the neural output $\psi_\theta(c, q) \in \mathbb{R}^d$ defines one transform row. For a fixed positive channel scale S_k estimated from training errors, $\delta x_k(\varphi, \lambda; \eta) = S_k \psi_\theta(c, q)^\top \eta$, $[U_\theta(c; Q)]_{q,:} = S_{k_q} \psi_\theta(c, q)^\top$. Here Q is a set of scalar queries, k_q its query's channel, and $U_\theta(c; Q) \in \mathbb{R}^{|Q| \times d}$ stacks the physical-scale rows. Its range is the available correction space, of dimension at most d. At fixed context, $\partial \delta x / \partial \eta = U_\theta(c)$ is coefficient independent, and $\eta = 0$ leaves the background unchanged. Thus the representation

can vary nonlinearly with the forecast across cycles while remaining linear in the online coefficients within each analysis.

The context map is $C = E_c \circ \mathcal{N}$, where $\mathcal{N}$ applies fixed channel normalization and E_c is a learned convolutional encoder. Training uses each error sample's paired forecast; online inference uses the current background, not observations. As in local implicit image functions (Chen et al., 2021), queries combine coordinates with local features, but the output here is a transform row rather than a state value. The same transform is used for training, ensemble projection, and field reconstruction.

We share network weights across overlapping regions b, each with local context c_b and coefficients η_b. This bounds each solve and localizes observational influence (Hunt et al., 2007). Weighted overlap-add combines regional corrections. After training, region-specific covariances of historical projected coefficients, $B^{\mathrm{clim}}_{z,b}$, provide a baseline prior. Because the transform depends on the current background, even this historical coefficient covariance induces a background-dependent physical-space covariance: $B^{\mathrm{clim}}_{x,b}(x^b) \approx U_\theta(c_b) B^{\mathrm{clim}}_{z,b} U_\theta(c_b)^\top$.

3.2 Learning Representations Recoverable from Partial Information

A correction space that fits complete forecast errors need not support reliable assimilation from sparse observations. FERA therefore trains both *representability* and *recoverability from limited support* using paired historical backgrounds and NMC-style forecast-error samples (Parrish & Derber, 1992). For sample i and queries Q, let $\widetilde{e}^{(i)}_Q$ contain errors divided by S_k, and let $\Psi^{(i)}_Q$ stack the corresponding unscaled rows $\psi_\theta^\top$. A coefficient $z^{(i)}_Q \in \mathbb{R}^d$ in the same coordinates as η is inferred by

$$z^{(i)}_Q = \arg\min_z |Q|^{-1} \|\widetilde{e}^{(i)}_Q - \Psi^{(i)}_Q z\|_2^2 + \lambda_{\mathrm{ridge}} \|z\|_2^2, \tag{3}$$

with ridge weight $\lambda_{\mathrm{ridge}} > 0$. The reconstruction loss $\mathcal{L}_{\mathrm{rec}}$ averages $|Q|^{-1}\|\widetilde{e}^{(i)}_Q - \Psi^{(i)}_Q z^{(i)}_Q\|_2^2$ over samples and queries, detaching the inferred coefficient for this path. This tests error representation when coefficient fitting and reconstruction use the same support.

Support-to-target inference. Assimilation instead asks whether coefficients inferred from one subset of the state yield useful corrections elsewhere. Suppressing i, let S be observed support and T a disjoint target set. We infer from S alone and score recovery on T:

$$\widehat{z}(S) = \arg\min_z \frac{\|\widetilde{e}_S - \Psi_S z\|_2^2}{|S|\sigma_S^2} + \gamma_S \|z\|_2^2, \qquad \ell(S,T) = \frac{\sum_{q\in T} w_{k_q} (\psi_\theta(c,q)^\top \widehat{z}(S) - \widetilde{e}_q)^2}{\sum_{q\in T} w_{k_q} \widetilde{e}_q^2 + \varepsilon}. \tag{4}$$

Here $\sigma_S^2 > 0$ is the support variance used for training, $\gamma_S > 0$ a coefficient-prior precision, $w_k > 0$ a target-channel weight, and $\varepsilon > 0$ a numerical stabilizer. These training weights are distinct from the online observation weighting.

Two complementary objectives. In held-out-variable tasks, support and targets use different channel groups, such as non-humidity variables as support and humidity as target. In sparse tasks, support contains four sampled points per channel and targets contain 1024 disjoint scalar entries. Denoting these sampling distributions by $\mathcal{D}_{\mathrm{hold}}$ and $\mathcal{D}_{\mathrm{sparse}}$, respectively,

$$\mathcal{L}_{\mathrm{hold}} = \mathbb{E}_{i,(S,T)\sim\mathcal{D}_{\mathrm{hold}}} \ell(S,T), \qquad \mathcal{L}_{\mathrm{sparse}} = \mathbb{E}_{i,(S,T)\sim\mathcal{D}_{\mathrm{sparse}}} \ell(S,T). \tag{5}$$

Both penalize errors made after partial-information coefficient inference, not just errors under a full-field fitted code. The complete objective is

$$\mathcal{L}_{\mathrm{repr}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{diag}}\mathcal{L}_{\mathrm{Gram,diag}} + \lambda_{\mathrm{off}}\mathcal{L}_{\mathrm{Gram,off}} + \lambda_{\mathrm{hold}}\mathcal{L}_{\mathrm{hold}} + \lambda_{\mathrm{sparse}}\mathcal{L}_{\mathrm{sparse}}, \tag{6}$$

where nonnegative λ's weight the terms and the Gram penalties control column norms and redundancy. Appendix A.2 gives their definitions; Appendix B specifies the tasks and weights.

Why partial-information training matters. Writing the solution of Eq. (4) explicitly, the standardized target reconstruction is

$$\widehat{e}_T(S) = \underbrace{\Psi_T \left(\Psi_S^\top \Psi_S + |S|\sigma_S^2 \gamma_S I_d\right)^{-1} \Psi_S^\top}_{\text{regularized support-to-target transfer}} \widetilde{e}_S, \tag{7}$$

where I_d is the $d \times d$ identity. A coefficient direction $v \in \mathbb{R}^d$ with $\Psi_S v \approx 0$ but large $\Psi_T v$ illustrates the problem: nearly identical support fits can imply different unseen corrections. Full-field reconstruction can reward such a direction without testing whether observations constrain it. Partial-information tasks instead favor patterns that predict target corrections from support across sampled masks and channel groups. Recovery still relies on learned relationships and regularization. Gram penalties control column scale and redundancy, not support-to-target transfer; ridge stabilizes coefficients for a fixed transform, whereas partial-information losses train that transform for useful sparse inference.

Variable withholding and sparse support expose cross-variable and spatial recovery, respectively. Both train the same transform used online without supplying its online coefficients or adding a neural update. Training support contains historical errors; assimilation instead uses current innovations and the fitted prior in Section 3.3.

3.3 Ensemble-Informed Prior and Explicit Linear Analysis

FERA uses the forecast ensemble to refine coefficient uncertainty without restricting mean corrections to its member span. Let $x_t^{f,(m)} \in \mathbb{R}^n$ be forecast member m at cycle t, M the ensemble size, and $\bar{x}_t^f = M^{-1} \sum_{m=1}^M x_t^{f,(m)}$ its mean. For $M > 1$, define the conventional normalized anomalies and sample covariance by $A_{x,t}^f = \frac{[x_t^{f,(1)} - \bar{x}_t^f, \ldots, x_t^{f,(M)} - \bar{x}_t^f]}{\sqrt{M-1}}$ and $B_{x,t}^{\mathrm{ens}} = A_{x,t}^f (A_{x,t}^f)^\top$. Its rank is at most $M - 1$. Rather than replacing the learned correction space by this rank-limited span, FERA projects the current member deviations into the learned coordinates and uses them to refine the coefficient prior. In cycling, we use $x_t^b = \bar{x}_t^f$ after the recentering described in Section 3.4.

Prior weighting. The current background fixes $c_t = C(x_t^b)$ and $U_\theta(c_t)$ at cycle t. Suppressing the region index, let $\Phi_t \in \mathbb{R}^{|Q_t| \times d}$ contain unscaled rows at projection queries Q_t, and let $A_{Q,t}^f \in \mathbb{R}^{M \times |Q_t|}$ contain the corresponding unnormalized member deviations $x_t^{f,(m)} - \bar{x}_t^f$, divided channel-wise by S_k. For ridge weight $\lambda_p > 0$, $Z_t^f = \arg\min_{Z \in \mathbb{R}^{M \times d}} |Q_t|^{-1} \|A_{Q,t}^f - Z\Phi_t^\top\|_F^2 + \lambda_p \|Z\|_F^2$. Here $\|\cdot\|_F$ is the Frobenius norm. Centering and processing the projected coefficients gives $L_{e,t} \in \mathbb{R}^{d \times r_t}$, $r_t \le \min(d, M-1)$. With $L_c L_c^\top = B_z^{\mathrm{clim}} \succ 0$, define $T_t = [\sqrt{w_c} L_c,\ \sqrt{w_e} L_{e,t}]$ and $B_{z,t} = T_t T_t^\top = w_c B_z^{\mathrm{clim}} + w_e L_{e,t} L_{e,t}^\top$. The weights satisfy $w_c > 0$, $w_e \ge 0$; $T_t \in \mathbb{R}^{d \times (d + r_t)}$. For $M = 1$ the ensemble block is absent. Because the climatological block remains positive definite, it preserves the full learned range while the ensemble changes its relative weighting.

Bayesian coefficient inference. Let $\mathcal{G}_t : \mathbb{R}^n \to \mathbb{R}^{n_{o,t}}$ map states to the $n_{o,t}$ observations y_t, and let $R_t \in \mathbb{R}^{n_{o,t} \times n_{o,t}} \succ 0$ be their effective error covariance. Writing $\|v\|_W^2 = v^\top W v$, the coefficients solve

$$\eta_t^a = \arg\min_\eta \ \tfrac{1}{2}\|\eta\|_{B_{z,t}^{-1}}^2 + \tfrac{1}{2}\|y_t - \mathcal{G}_t(x_t^b + U_\theta(c_t)\eta)\|_{R_t^{-1}}^2. \tag{8}$$

Under the Gaussian coefficient prior and Gaussian observation model, this is the coefficient-space negative log-posterior up to an additive constant. For the linear observations used here, $\mathcal{G}_t(x) = H_t x$. Define the observed transform $F_t = H_t U_\theta(c_t)$ and innovation $d_{y,t} = y_t - H_t x_t^b$. With posterior precision Q_t,

$$\mathsf{Q}_t \eta_t^a = F_t^\top R_t^{-1} d_{y,t}, \qquad \mathsf{Q}_t = B_{z,t}^{-1} + F_t^\top R_t^{-1} F_t \succ 0. \tag{9}$$

Because Q_t is positive definite, this $d \times d$ system has a unique solution. The neural transform is fixed throughout the solve, so the coefficient-to-increment Jacobian does not change with η and no optimizer initializer or neural-network adaptation is required. The analysis is $\bar{x}_t^a = x_t^b + U_\theta(c_t)\eta_t^a$. Appendix A.5 relates this native formulation to the extended-control implementation, and Appendix D evaluates numerical robustness under fixed-input perturbations. Nonlinear $\mathcal{G}_t$ would retain a linear coefficient-to-increment map but need not yield a quadratic objective.

Propagation to unobserved entries. Suppressing t, let $U_* \in \mathbb{R}^{n_* \times d}$ contain rows for n_* target entries, observed or not. The equivalent observation-space update is $\delta x_*^a = U_* B_z F^\top (F B_z F^\top + R)^{-1} d_y$. For scalar grid observations, the cross-covariance entries are $k_\theta(q, o \mid c) = u_\theta(c, q)^\top B_z u_\theta(c, o)$, where $u_\theta(c, q) = S_{k_q} \psi_\theta(c, q)$ and o is an observed query. Thus forecast conditioning determines the multivariate correction geometry, partial-information training targets

whether its couplings are useful for support-to-target recovery, the prior weights the available directions, and observations select the realized correction. The resulting observation-to-state map is explicit and inspectable for each background.

3.4 Integration into Cycling Assimilation

For $M > 1$, a separate tile-local symmetric ensemble transform Kalman filter (ETKF) updates member anomalies (Bishop et al., 2001; Hunt et al., 2007), distinct from the coefficient prior for the FERA mean. Let $a_t^{a,(m)}$ denote the zero-mean analysis anomalies after overlap-add and recentering, and $\mathcal{M}_{t\to t+1}$ the forecast model. We form

$$x_t^{a,(m)} = \bar{x}_t^a + a_t^{a,(m)}, \qquad x_{t+1}^b = \mathcal{M}_{t\to t+1}(\bar{x}_t^a). \tag{10}$$

Members are advanced independently and recentered so that $\bar{x}_{t+1}^f = x_{t+1}^b$. This hybrid integration (Penny, 2014) supplies the next context and ensemble information for the coefficient prior; the members are not samples from the learned mean-control posterior. Appendix A.4 gives the recurrence. For $M = 1$, only the learned mean is updated and propagated.

4 Experiments

Experimental Setup We evaluate FERA in a global six-hourly observing system simulation experiment (OSSE) using FengWu (Chen et al., 2023a): 69 channels on a 721×1440 grid over 116 cycles (approximately one month). Each cycle observes all ERA5 channels at 1000 freshly sampled horizontal columns unless otherwise stated, without added noise (Hersbach et al., 2020). For single-member runs ($M = 1$), the baselines are 3DVar, using a climatological covariance (Courtier et al., 1998), and VAE-Var, using a learned nonlinear error decoder (Xiao et al., 2025b). For multi-member runs ($M \in \{4, 8, 12\}$), Hybrid-3DEnVar combines climatological and ensemble covariances (Buehner, 2005), while LoRA-EnVar adapts the error decoder to the ensemble (Xiao et al., 2025a). Ensembles use time-lagged initial states and deterministic-control recentering. For diagnostic k and cycle t, we use latitude-weighted spatial WRMSE, $\mathrm{WRMSE}_{k,t} = \left[\frac{\sum_i \cos\varphi_i \,(x^a_{k,t,i} - x^{\mathrm{ERA5}}_{k,t,i})^2}{\sum_i \cos\varphi_i}\right]^{1/2}$, where i indexes grid points, φ_i is latitude, and x^{ERA5} is the reference state. Diagnostic symbols z, q, u, v, t denote geopotential, specific humidity, zonal wind, meridional wind, and temperature; subscripts 500/850 indicate pressure in hPa, $t_{2\mathrm{m}}$ is 2-m temperature, and MSL is mean-sea-level pressure.

Does FERA Improve Cycling Assimilation? FERA consistently improves cycling analyses with and without ensemble assistance. At $M = 1$, where the learned analysis uses no current-ensemble covariance block, Figure 2(a–c) shows time-mean gains of 4.3%, 3.1%, and 1.0% for z_{500}, u_{500}, and t_{500} over the stronger baseline. Across all twelve diagnostics, FERA is best in eleven and nearly tied on t_{850}. The spatial example in Figure 2(d–f) further shows that the improvement corresponds to meaningful full-field corrections: on 13 January, FERA reduces global z_{500} WRMSE from the VAE-Var baseline's 48.32 to $43.41\,\mathrm{m^2\,s^{-2}}$, a 10.2% reduction. The advantage persists when ensemble information is introduced. Across $M \in \{1, 4, 8, 12\}$, FERA obtains the lowest WRMSE in 46 of 48 ensemble-size–diagnostic comparisons; the only exceptions are t_{850} at $M = 1$ and $M = 4$, where the gaps to the best baseline are negligible. At $M = 12$, FERA is best for all twelve diagnostics, with improvements spanning mass, wind, moisture, and temperature variables rather than being concentrated in a single variable family.

How Robust Is the Direct Bayesian Analysis? FERA makes coefficient inference explicit: for fixed inputs and linear observations, a positive-definite posterior precision yields a unique solution without optimizer initialization. The prior and observation-to-state map expose how observations produce physical corrections (Section 3.3). Figure 3 probes numerical stability in one fixed-input case ($M = 12$, 1000 columns), using six equivalent observation reorderings and five random initializations for iterative solvers (Appendix D). Channel-standardized WRMSE differences are measured against method- and budget-matched references. Direct FERA changes by at most 1.47×10^{-7} under reordering and gives identical outputs on unchanged-input repeats; VAE-Var exhibits larger worst-case finite-budget differences. For iterative FERA, the transform and objective remain fixed; larger

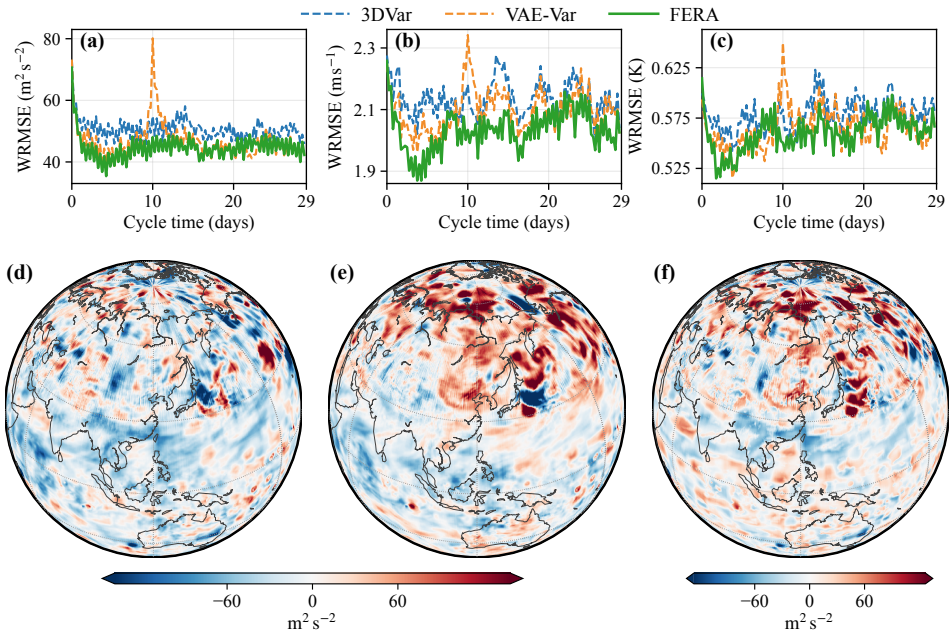


Figure 2: **Cycling skill and spatial analysis errors.** (a–c) Unsmoothed z_{500}, u_{500} and t_{500} analysis WRMSE over 116 cycles ($M = 1$, 1000 columns). (d,e) z_{500} analysis errors $e^a = x^a - x^{\text{ref}}$ against ERA5 (x^{ref}) for FERA and VAE-Var, respectively. (f) $|e^a_{\text{baseline}}| - |e^a_{\text{FERA}}|$: red favors FERA; blue favors the baseline. The Orthographic maps (120°E, 35°N) show 13 January 2024, 18:00 UTC, from a separate cycling run. Panels (d,e) share a symmetric scale. Appendix C.2 gives the field diagnostics and run settings.

Table 1: Time-mean analysis WRMSE over 116 six-hour cycles with 1000 observation columns. Best values within each M block are bold; q is shown as $10^3 q$. Units are $\mathrm{m^2\,s^{-2}}$ for z, $\mathrm{m\,s^{-1}}$ for winds, K for temperatures, and Pa for MSL.

		z		$10^3 q$		u		v		t		Surface	
M	Method	500	850	500	850	500	850	500	850	500	850	MSL	$t_{2\mathrm{m}}$
1	3DVar	50.604	39.761	0.471	1.196	2.134	1.688	2.094	1.675	0.5783	**0.8485**	64.341	0.9510
	VAE-Var	46.648	39.021	0.470	1.183	2.098	1.670	2.028	1.657	0.5657	0.8525	72.174	0.9420
	FERA	**44.661**	**38.758**	**0.465**	**1.135**	**2.033**	**1.664**	**2.020**	**1.641**	**0.5598**	0.8489	**58.955**	**0.9246**
4	Hybrid-3DEnVar	49.936	39.328	0.469	1.190	2.123	1.676	2.083	1.664	0.5711	**0.8402**	64.192	0.9439
	LoRA-EnVar	45.993	38.541	0.473	1.182	2.092	1.661	2.025	1.647	0.5628	0.8498	70.508	0.9332
	FERA	**43.490**	**37.921**	**0.465**	**1.137**	**2.005**	**1.644**	**1.990**	**1.620**	**0.5519**	0.8416	**57.781**	**0.9195**
8	Hybrid-3DEnVar	49.039	39.092	0.468	1.180	2.099	1.657	2.067	1.644	0.5664	0.8351	63.985	0.9377
	LoRA-EnVar	45.782	38.721	0.470	1.181	2.088	1.663	2.025	1.649	0.5622	0.8483	70.816	0.9315
	FERA	**42.488**	**37.104**	**0.464**	**1.136**	**1.980**	**1.624**	**1.969**	**1.602**	**0.5451**	**0.8345**	**56.780**	**0.9153**
12	Hybrid-3DEnVar	49.067	39.247	0.467	1.177	2.095	1.658	2.065	1.645	0.5645	0.8348	64.167	0.9376
	LoRA-EnVar	47.628	39.630	0.471	1.188	2.109	1.678	2.040	1.662	0.5705	0.8576	73.736	0.9449
	FERA	**42.305**	**36.991**	**0.465**	**1.137**	**1.974**	**1.617**	**1.962**	**1.598**	**0.5433**	**0.8320**	**56.550**	**0.9133**

finite-budget differences therefore expose solver dependence within the same representation. This supports reliable numerical inference under the tested perturbations.

Do the Representation Choices Matter? To examine FERA's two representation choices, we conduct retrained forecast-conditioning and partial-information ablations with 1000 and 500 observed columns, respectively. In the first, we remove forecast conditioning by passing a fixed zero stan-

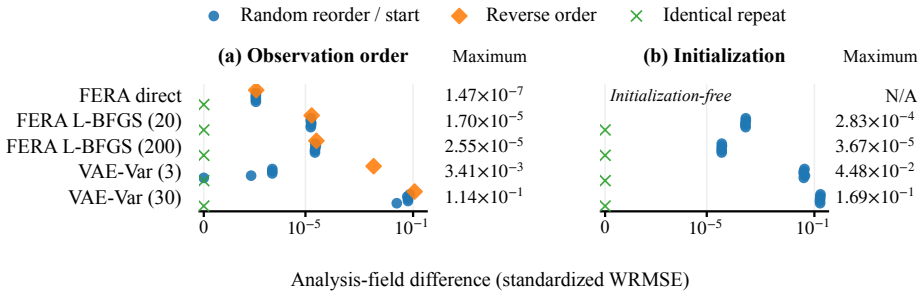


Figure 3: **Numerical robustness of direct analysis.** Analysis-field differences under (a) equivalent observation reorderings and (b) optimization initialization, with all other inputs fixed ($M = 12$, 1000 columns, one weather case). The method labels at left apply to both panels. Circles show five random permutations or starts, diamonds show reverse order, and crosses show identical-input repeats; each panel's adjacent Maximum column reports maxima. Direct FERA needs no initialization.

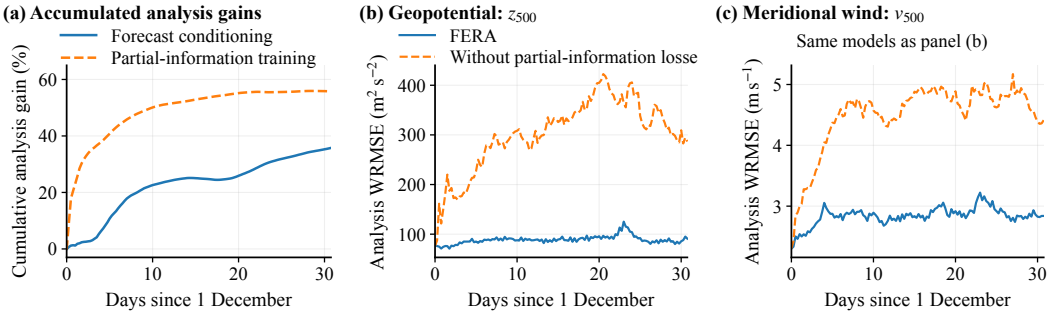


Figure 4: **Representation choices matter through cycling.** (a) Cumulative mean relative analysis gain over the same twelve diagnostics, recomputed from time-mean WRMSE, for the two retrained contrasts. (b,c) Partial-information ablation: z_{500} and v_{500} analysis trajectories.

dardized field through the context encoder during both training and assimilation, while retaining the true background in the innovations and forecasts. In the second, we remove partial-information training by setting both held-out and sparse-posterior loss weights to zero, while retaining forecast conditioning and full-field reconstruction. Across the full 124-cycle December evaluation, as shown in Figure 4, both ablations demonstrate the effectiveness of the corresponding design choice. Forecast conditioning reduces the mean diagnostic-wise analysis WRMSE by 35.76% and the subsequent-background WRMSE by 38.64%, while partial-information training yields corresponding improvements of 55.80% and 58.01%; in both comparisons, FERA improves all twelve diagnostics at both stages. Both contrasts retain direct inference, linking the gains to representation learning. Better subsequent backgrounds show that these benefits persist through forecast propagation. Together, the ablations show that forecast-adaptive patterns and training for recovery from partial observations both contribute to sustained cycling gains.

5 Conclusion

FERA learns forecast-conditioned error patterns for recovery from partial observations and uses a direct Bayesian solve to choose their coefficients. Historical and ensemble statistics refine the prior. Across month-long cycling, FERA leads 46 of 48 ensemble-size–diagnostic comparisons; fixed-input replay supports numerical robustness without optimizer initialization. Retraining ablations support both representation choices. Together, these results connect adaptive error learning with reliable, interpretable Bayesian inference. Extending FERA to heterogeneous noisy observations remains an important next step.

Reproducibility Statement

Appendices A and B describe the representation, training objectives, linear analysis and cycling protocol. Appendix C collects complete cycling and physical-field diagnostics, and Appendix D documents the numerical-consistency replay. Appendices E and F each pair a retraining protocol with its diagnostic scores. The accompanying code package provides FERA's representation, training and cycling implementation with example configurations. End-to-end runs additionally require the external forecast-model and FERA checkpoints, ERA5 fields, and NMC samples.

AI Use Statement

Generative AI tools assisted with conceptual framing, mathematical exposition, experiment design and feedback, code review, diagnostic and plotting-code preparation, interpretation of recorded results, and manuscript editing and formatting. They also assisted with figure design. The reported numerical measurements are taken from the supplied experiment records, rather than generated or imputed by a language model. The authors have checked the analyses, claims, citations, and final manuscript.

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, et al. End-to-end data-driven weather prediction. *Nature*, 641(8065):1172–1179, 2025.

Ross N Bannister. A review of forecast error covariance statistics in atmospheric variational data assimilation. I: Characteristics and measurements of forecast error covariances. *Quarterly Journal of the Royal Meteorological Society*, 134(637):1951–1970, 2008.

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

Craig H Bishop, Brian J Etherton, and Sharanya J Majumdar. Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Monthly Weather Review*, 129(3):420–436, 2001.

Mark Buehner. Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting. *Quarterly Journal of the Royal Meteorological Society*, 131(607):1013–1043, 2005.

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.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. FengWu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023a.

Peter Yichen Chen, Jinxu Xiang, Dong Heon Cho, Yue Chang, G A Pershing, Henrique Teles Maia, Maurizio M Chiaramonte, Kevin Thomas Carlberg, and Eitan Grinspun. CROM: Continuous reduced-order modeling of PDEs using implicit neural representations. In *The Eleventh International Conference on Learning Representations*, 2023b.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8628–8638, 2021.

Philippe Courtier, Erik Andersson, William Heckley, D Vasiljevic, M Hamrud, A Hollingsworth, F Rabier, M Fisher, and J Pailleux. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quarterly Journal of the Royal Meteorological Society*, 124 (550):1783–1807, 1998.

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

Hang Fan, Juan Nathaniel, Yi Xiao, Ce Bian, Fenghua Ling, Ben Fei, Lei Bai, and Pierre Gentine. Accurate and efficient hybrid-ensemble atmospheric data assimilation in latent space with uncertainty quantification. *arXiv preprint arXiv:2603.04395*, 2026.

Thomas M Hamill and Chris Snyder. A hybrid ensemble Kalman filter–3D variational analysis scheme. *Monthly Weather Review*, 128(8):2905–2919, 2000.

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.

Peter L Houtekamer and Herschel L Mitchell. Data assimilation using an ensemble Kalman filter technique. *Monthly Weather Review*, 126(3):796–811, 1998.

Langwen Huang, Lukas Gianinazzi, Yuejiang Yu, Peter Dominik Dueben, and Torsten Hoefler. DiffDA: a diffusion model for weather-scale data assimilation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19798–19815, 2024. URL https://proceedings.mlr.press/v235/huang24h.html.

Brian R Hunt, Eric J Kostelich, and Istvan Szunyogh. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D: Nonlinear Phenomena*, 230(1-2): 112–126, 2007.

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge University Press, 2003.

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

Zhuoyuan Li, Bin Dong, and Pingwen Zhang. Latent assimilation with implicit neural representations for unknown dynamics. *Journal of Computational Physics*, 506:112953, 2024.

Boštjan Melinc and Žiga Zaplotnik. 3D-Var data assimilation using a variational autoencoder. *Quarterly Journal of the Royal Meteorological Society*, 150(761):2273–2295, 2024.

Boštjan Melinc, Uroš Perkan, and Žiga Zaplotnik. A unified neural background-error covariance model for midlatitude and tropical atmospheric data assimilation. *Journal of Advances in Modeling Earth Systems*, 18(1):e2025MS005360, 2026.

David F Parrish and John C Derber. The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Review*, 120(8):1747–1763, 1992.

Stephen G Penny. The hybrid local ensemble transform Kalman filter. *Monthly Weather Review*, 142 (6):2139–2149, 2014.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

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

Pavel Sakov and Peter R Oke. A deterministic formulation of the ensemble Kalman filter: an alternative to ensemble square root filters. *Tellus A*, 60(2):361–371, 2008.

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*, 2025.

Phillip Si and Peng Chen. LEVDA: Latent ensemble variational data assimilation via differentiable dynamics. *arXiv preprint arXiv:2602.19406*, 2026.

Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7537–7547, 2020.

Jeffrey S Whitaker and Thomas M Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, 2002.

Yi Xiao, Hang Fan, Kun Chen, Ye Cao, Ben Fei, Wei Xue, and Lei Bai. LoRA-EnVar: Parameter-efficient hybrid ensemble variational assimilation for weather forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 45626–45661, 2025a.

Yi Xiao, Qilong Jia, Kun Chen, Lei Bai, and Wei Xue. VAE-Var: variational autoencoder-enhanced variational methods for data assimilation in meteorology. In *International Conference on Learning Representations*, 2025b.

Yuan Yin, Matthieu Kirchmeyer, Jean-Yves Franceschi, Alain Rakotomamonjy, and Patrick Gallinari. Continuous PDE dynamics forecasting with implicit neural representations. In *The Eleventh International Conference on Learning Representations*, 2023.

A Additional Method Details

A.1 Additional Context for Conditional Error Modeling

The background-dependent error prior is motivated by linearized forecast-error evolution along a trajectory (Kalnay, 2003; Evensen, 1994; Bannister, 2008):

$$e_t \approx M_t e_{t-1} + q_t, \qquad B_t \approx M_t B_{t-1} M_t^\top + Q_t. \tag{11}$$

Here $e_t \in \mathbb{R}^n$ is the forecast error at time t, $B_t = \mathrm{Cov}(e_t)$, $M_t \in \mathbb{R}^{n\times n}$ is a tangent forecast operator, and q_t is additive model error with covariance $Q_t = \mathrm{Cov}(q_t)$; the covariance approximation neglects cross terms. Here Q_t is a covariance, distinct from the projection-query set in Section 3.3. State-dependent stretching and multivariate coupling motivate using the deployable forecast as context rather than pooling all backgrounds into an unconditional prior.

Neural DA methods learn different parts of the observation-to-state interface. Aardvark Weather learns an observation-to-forecast pipeline (Allen et al., 2025); CROM and DINo learn compact field dynamics (Chen et al., 2023b; Yin et al., 2023). Latent 3D-Var, LAINR, Latent-EnSF, HLOBA and LEVDA bring state representations into assimilation (Melinc & Zaplotnik, 2024; Li et al., 2024; Si & Chen, 2025; Fan et al., 2026; Si & Chen, 2026), and neural covariance models target multivariate flow dependence (Melinc et al., 2026). Latent-EnSF explicitly aligns encodings of complete states and sparse observations; FERA instead learns a forecast-error transform through an explicit linear support solve. A state representation need not preserve the smaller error directions most useful for propagating observational corrections, motivating error-focused rather than full-state reconstruction.

Diffusion and score-based DA provide expressive conditional samples, while ensemble score filters support recursive uncertainty estimation (Rozet & Louppe, 2023; Huang et al., 2024; Bao et al., 2024). These are different inference interfaces from FERA's directly solvable mean-control problem. Deterministic ensemble updates (Whitaker & Hamill, 2002; Sakov & Oke, 2008) remain useful supporting components for propagating members, but are not the learned correction-space contribution.

A.2 Partial-Information Representation Objectives

Section 3.2 defines the two support-to-target objectives. Their support coefficient has the normal equation

$$\left(\frac{\Psi_S^\top \Psi_S}{|S|\sigma_S^2} + \gamma_S I_d\right) \widehat{z}(S) = \frac{\Psi_S^\top \widetilde{e}_S}{|S|\sigma_S^2}, \tag{12}$$

where I_d is the $d \times d$ identity. Numerical stabilization adds a small adaptive ridge to the left-hand matrix. The support contains standardized historical error targets, not the target entries on which the loss is evaluated. Training variances and precisions govern this auxiliary inference task; they are distinct from the effective observation weighting R used online.

Held-out tasks include non-humidity support to humidity targets, mass/surface support to wind targets, mass/wind/surface support to temperature targets, and low-level cross-variable transfer. Sparse tasks use four sampled points per channel for support and 1024 disjoint scalar targets. The held-out objective therefore emphasizes cross-variable inference, while the sparse objective emphasizes recovery from limited spatial support. Neither is an additional online optimizer: both change the frozen transform learned before assimilation.

For $G_Q = |Q|^{-1}\Psi_Q^\top \Psi_Q$, the Gram penalties in Eq. (6) are

$$\mathcal{L}_{\mathrm{Gram,diag}} = \frac{1}{d}\|\operatorname{diag}(G_Q) - \mathbf{1}\|_2^2, \qquad \mathcal{L}_{\mathrm{Gram,off}} = \frac{1}{d(d-1)}\|G_Q - \operatorname{Diag}(\operatorname{diag} G_Q)\|_F^2. \tag{13}$$

Here diag extracts a matrix diagonal, Diag forms a diagonal matrix from a vector, and $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector. Expectations over training examples and query draws are implicit. The main reconstruction code is detached; no explicit hydrostatic or geostrophic balance penalty is imposed. The task definitions and reported objective weights are summarized in Appendix B; support weights in these auxiliary tasks should not be identified with the online observation covariance.

A.3 Construction of the Ensemble-Derived Control Factor

For $M > 1$, let $L_c \in \mathbb{R}^{d\times d}$ be the Cholesky factor of $B^{\mathrm{clim}}_{z,b}$, including prior jitter. Suppressing region and cycle indices, define $P_\perp = I_M - \mathbf{1}\mathbf{1}^\top/M$ and $L_{\mathrm{raw}} = (Z^f)^\top P_\perp/\sqrt{M-1}$ for projected coefficients $Z^f \in \mathbb{R}^{M\times d}$. Here I_M is the $M \times M$ identity and $\mathbf{1} \in \mathbb{R}^M$ the all-ones vector. We whiten the factor as $W = L_c^{-1}L_{\mathrm{raw}}$ and compute a thin SVD $W = U\Sigma V^\top$, where U, V contain left and right singular vectors and Σ contains singular values; this U is distinct from $U_\theta(c)$. Projection-residual, effective-rank, variance, and scale checks reject unreliable regional estimates. For an accepted region, the active singular values are trace matched to the climatological scale and shrunk toward unit whitened variance,

$$\widehat{\lambda}_i = (1-\gamma)s^2\sigma_i^2 + \gamma, \qquad L_e = L_c U_r \,\mathrm{Diag}(\widehat{\lambda}_1^{1/2}, \ldots, \widehat{\lambda}_r^{1/2}). \tag{14}$$

Here r is the retained rank, σ_i are the retained singular values, $U_r \in \mathbb{R}^{d\times r}$ contains their left singular vectors, $s > 0$ is the trace-matching scale, and $\gamma \in [0, 1]$ is the active-variance shrinkage parameter. The analysis-mean transform (denoted T_t in the main text) is

$$T_z = [\sqrt{w_c}L_c,\ \sqrt{w_e}L_e] \in \mathbb{R}^{d\times D}, \qquad D = d + r, \tag{15}$$

with implicit latent covariance

$$T_z T_z^\top = w_c B^{\mathrm{clim}}_{z,b} + w_e L_e L_e^\top \quad \text{(with the same prior jitter).} \tag{16}$$

If any check fails, the region falls back to the climatological control.

This factor enters the mean objective in Eq. (8); Eq. (23) gives the linear-observation system. Weighted overlap-add combines regional analyses. The retained climatological block supplies mean-correction directions not limited to those sampled by the current ensemble.

Structure and magnitude are complementary, not an eigendecomposition. The covariance $U_\theta(c)B_{z,t}U_\theta(c)^\top$ makes the roles of the background and ensemble explicit. The background changes the spatial and cross-variable patterns represented by the transform; the projected ensemble changes their relative variances and correlations. The transform is neither orthonormal nor restricted to unit-amplitude columns, so this is not a unique decomposition into eigenvectors and eigenvalues. In particular, "magnitude" means uncertainty in multiple coefficient directions, not a single global spread multiplier. For $w_c > 0$ and nonsingular L_c, the climatological block ensures $\mathrm{range}(U_\theta(c)T_t) = \mathrm{range}(U_\theta(c))$: the ensemble factor changes weighting within that space, rather than adding physical directions outside it. The separate ETKF branch updates actual member anomalies and is not a sample from this mean-control prior or its posterior.

A.4 Ensemble Anomaly Update and Probabilistic Scope

Updating member anomalies is the second use of the ensemble, distinct from weighting the learned mean control. FERA pairs its learned mean-control update with a separate tile-local symmetric ETKF transform (Bishop et al., 2001; Hunt et al., 2007). For a tile with q_b state entries and p_b observations, let $A^f_x \in \mathbb{R}^{M\times q_b}$ contain unnormalized row-wise forecast deviations and let $H \in \mathbb{R}^{p_b\times q_b}$ be its observation matrix. Thus A^f_x is $\sqrt{M-1}$ times the transpose of the tile restriction of the normalized, column-wise $A^f_{x,t}$ in Section 3.3. Here $R \in \mathbb{R}^{p_b\times p_b}$ is the positive-definite effective observation-error covariance. Their observed anomalies are $Y = A^f_x H^\top \in \mathbb{R}^{M\times p_b}$. The anomaly transform is

$$C_M = I_M + \frac{1}{M-1} Y R^{-1} Y^\top, \qquad T_{\mathrm{ETKF}} = P_\perp C_M^{-1/2} P_\perp + P_0, \tag{17}$$

where $P_0 = \mathbf{1}\mathbf{1}^\top/M$ projects onto the ensemble-mean direction, $P_\perp = I_M - P_0$ centers the members, and $C_M^{-1/2}$ is the symmetric inverse square root. It produces

$$A^a_x = T_{\mathrm{ETKF}} A^f_x. \tag{18}$$

Regions without observations retain their forecast anomalies. After overlap-add, the anomalies are recentered around the FERA mean,

$$x_t^{a,(m)} = \bar{x}_t^a + a_t^{a,(m)}, \tag{19}$$

where $a_t^{a,(m)} \in \mathbb{R}^n$ is member m's assembled, recentered anomaly. Each member is advanced independently to an intermediate forecast,

$$\widetilde{x}_{t+1}^{f,(m)} = \mathcal{M}_{t\to t+1}(x_t^{a,(m)}), \qquad x_{t+1}^b = \mathcal{M}_{t\to t+1}(\bar{x}_t^a). \tag{20}$$

The second expression advances the deterministic control. With $\overline{\widetilde{x}}_{t+1}^f = M^{-1}\sum_m \widetilde{x}_{t+1}^{f,(m)}$, the actual next-cycle members are recentered as

$$x_{t+1}^{f,(m)} = x_{t+1}^b + \widetilde{x}_{t+1}^{f,(m)} - \overline{\widetilde{x}}_{t+1}^f, \qquad \bar{x}_{t+1}^f = x_{t+1}^b. \tag{21}$$

This distinguishes the independently forecast member mean from the recentered background used to construct the next transform. For nonlinear $\mathcal{M}$, the two need not coincide before recentering.

Following the hybrid mean-and-ensemble principle (Penny, 2014), the learned transform allows mean corrections beyond the instantaneous member span, while ETKF retains the member-based covariance update. The nonlinear forecast then evolves these perturbations into the next cycle's uncertainty. The mean and anomaly branches are complementary, not samples from one learned posterior; together they connect background-conditioned structure to sequential uncertainty propagation.

A.5 Uniqueness and Observation-Permutation Consistency

Native and whitened coefficients. The main text and Figure 1 use native coefficients $\eta \in \mathbb{R}^d$, while the extended-control implementation uses $\eta = T\xi$ with whitened control $\xi \in \mathbb{R}^D$ and prior $\xi \sim \mathcal{N}(0, I_D)$. Here $T = T_z \in \mathbb{R}^{d\times D}$, $U = U_\theta(c)$, and region and cycle indices are suppressed. For $B_z = TT^\top \succ 0$, T has full row rank and

$$\min_{\xi: T\xi=\eta} \|\xi\|_2^2 = \eta^\top B_z^{-1}\eta, \qquad \xi_{\min} = T^\top B_z^{-1}\eta. \tag{22}$$

Hence minimizing the native-η objective or the identity-prior ξ objective gives the same physical correction in exact arithmetic, even when T is rectangular. With $G = HUT$ and $d_y = y - H\bar{x}^f$, the latter system is

$$\mathsf{A}\xi^a = G^\top R^{-1} d_y, \qquad \mathsf{A} = I_D + G^\top R^{-1} G \succ 0. \tag{23}$$

Here $G \in \mathbb{R}^{n_o\times D}$ is the observed extended transform and $\mathsf{A} \in \mathbb{R}^{D\times D}$ its posterior precision. Numerical jitter and stabilization must be included consistently when comparing implementations. The local field illustration in Figure 1(D) uses a solve in native η; it is not asserted to be a stored $T\xi$ run.

Uniqueness and reordering. Fix a local background, the forecast ensemble, projection queries, network weights and the ensemble-derived factor. Let $V = U_\theta(c)T$, $G = HV$, $d_y = y - H\bar{x}^f$, and $R \succ 0$. For linear H, the mean objective is

$$J(\xi) = \tfrac{1}{2}\|\xi\|_2^2 + \tfrac{1}{2}(d_y - G\xi)^\top R^{-1}(d_y - G\xi). \tag{24}$$

For any nonzero v, $v^\top \nabla^2 J v = \|v\|_2^2 + (Gv)^\top R^{-1}(Gv) > 0$. Hence the coefficient solution is unique even if G is rank deficient. A fixed positive regularization term used in the numerical system preserves this property. A direct factorization solves the system without selecting an initial control vector; this is distinct from changing the physical background or the initial forecast ensemble.

Let $P \in \mathbb{R}^{n_o\times n_o}$ be a permutation matrix acting on the observation list. A consistent reorder replaces (y, H, R) by $(Py, PH, PRP^\top)$, giving $G' = PG$, $d_y' = Pd_y$, and $R' = PRP^\top$. Since $(PRP^\top)^{-1} = PR^{-1}P^\top$,

$$(G')^\top (R')^{-1} G' = G^\top R^{-1} G, \qquad (G')^\top (R')^{-1} d_y' = G^\top R^{-1} d_y. \tag{25}$$

The normal equation and its solution are therefore unchanged. With fixed regional supports and overlap weights, the same statement holds after overlap-add. It does not cover changing observation times, dropping observations, updating the background sequentially, rebuilding the transform between batches, or changing quality-control decisions.

Permutation invariance of a complete likelihood is not unique to linear methods. The advantage here is that a unique minimizer can be obtained without a nonlinear iterative search or an optimizer

initializer. In finite precision, reordered reductions can change the assembled matrix and right-hand side slightly; the output sensitivity depends on conditioning, precision and solver behavior. Figure 3 therefore measures sensitivity to equivalent reorderings rather than treating uniqueness as a guarantee of numerical stability. For nonlinear observation operators or a nonlinear four-dimensional forecast map, the same control representation can be composed with an iterative variational objective, but uniqueness and quadratic structure are no longer guaranteed by this argument.

A.6 Computational Complexity

Let K be the projection-query count, p the local observation count, q the local output dimension, d the learned control dimension, and $D = d + r \leq 2d$ the extended-control dimension. A dense-algebra cost bound is

$$\begin{aligned} C_{\mathrm{mean}} &= \mathcal{O}(Kd^2 + d^3 + MKd + Md^2 + dM\min(d, M) + pD^2 + D^3 + qdD), \\ C_{\mathrm{ETKF}} &= \mathcal{O}(pM^2 + M^3 + qM^2). \end{aligned} \tag{26}$$

In a separate one-case pilot on an NVIDIA A100 with 1000 observed columns and zero initial controls for iterative solvers, the direct FERA coefficient solve takes 0.91 s, compared with 44.53 s for 20-step L-BFGS on the same FERA objective. The six-step VAE-Var solve stage takes 19.40 s under its own objective. These are solve-stage measurements, not complete analysis or cycling times.

The terms cover projection and whitening, low-rank processing, observation-system assembly, factorization, and explicit transformed output rows. Contracting $\eta = T\xi$ first reduces the last term to $\mathcal{O}(dD + qd)$; this is an algebraic alternative, not a measured implementation speedup. Context encoding and neural row evaluation add costs not represented by these dense-algebra terms. The native solve is $d \times d$ and the extended solve at most $2d \times 2d$, whereas the anomaly branch still factors an $M \times M$ matrix. Eliminating online network adaptation does not by itself establish a lower end-to-end cost.

A.7 Workflow Illustration and Provenance

Figure 1 is exported from an editable PowerPoint. Its boxes, equations, and arrows summarize the method; the field thumbnails illustrate one saved local diagnostic, not a measured comparison among the four workflow stages. Panel A shows the background-conditioned transform and slices of native modes 147 and 101 for z_{500}, u_{500}, and t_{500}; the labels u_{147} and u_{101} denote transform columns, not wind variables. These modes were chosen by unscaled transform-column norm on the common display grid without using observations or truth; they are not covariance eigenvectors. The globe is a coarsened background-only preview.

The illustrative fields come from a frozen-model $M = 1$ case at 00:00 UTC on 3 January 2024, tile 72 (76.25–84°N, 72–79.75°E), on a native 32×32 latitude–longitude grid. Panel D shows the local u_{500} background, the full increment from all 256 coefficient modes, and their sum before overlap-add and postprocessing. The two modes pictured in panel A alone do not reconstruct that increment. The maps use native latitude–longitude coordinates rather than an equal-distance projection; the displayed updated state is not a saved global analysis.

Panel B sketches the two offline partial-information tasks: sparse support points with disjoint targets, and non-humidity support with humidity held out. Panel C sketches the online solve with a prior informed by historical statistics and by projected ensemble information when available. The illustrated $M = 1$ case has no ensemble covariance block, and the neural transform is not retrained during an analysis. The outer arrow indicates that the updated state feeds a later forecast; the thumbnails themselves do not establish cycling skill or ensemble reliability.

B Experimental and Implementation Details

B.1 Cyclic OSSE Protocol

FengWu advances a 69-channel global state every six hours. The upper-air channels are geopotential, specific humidity, zonal wind, meridional wind, and temperature on 13 pressure levels; the four surface channels are 10-m zonal and meridional wind, 2-m temperature, and mean-sea-level pressure.

We evaluate 116 cycles from 3 January 2024, 00:00 UTC through 31 January, 18:00 UTC on a 721×1440 grid. For the main comparison, the FERA representation is trained on NMC samples from February–December 2024, excluding this January evaluation window. At each cycle, the observation generator selects a seeded random set of horizontal columns without replacement and reveals all channels at those columns. Observation counts of 500, 1000, and 2000 are used in the density study; the main baseline setting uses 1000. The separate December ablations use the settings in Appendices E.1 and F.1.

The primary metric is latitude-weighted WRMSE against ERA5. Table 1 averages analysis WRMSE over all 116 cycles. Figure 11 pairs analysis cycle t with the background at cycle $t + 1$ and therefore uses 115 transitions. For its two axes, the baseline is selected independently as the lower-error non-FERA method at the matched ensemble size and evaluation stage.

B.2 Baselines and Ensemble Construction

3DVar solves for an increment under a climatological covariance and an observation penalty (Courtier et al., 1998). Hybrid-3DEnVar adds a covariance estimated from forecast-member deviations to this historical component (Buehner, 2005). VAE-Var instead decodes forecast-error increments from a learned latent representation and optimizes the latent coefficients against observations (Xiao et al., 2025b); LoRA-EnVar additionally adapts the decoder to the current ensemble through low-rank parameter updates (Xiao et al., 2025a). These references describe the algorithmic families; the reported comparisons use their available tuned implementations with the same FengWu forecast system.

The single-member comparison uses 3DVar, VAE-Var and FERA. Multi-member comparisons use Hybrid-3DEnVar, LoRA-EnVar and FERA at $M \in \{4, 8, 12\}$. Forecast ensembles are initialized from time-lagged states and recentered on a deterministic control forecast. The $M = 1$ FERA run omits both the ensemble-derived coefficient block and the ETKF anomaly update.

Each method retains its tuned observation-error scaling. In particular, VAE-Var and LoRA-EnVar use channel standard-deviation factors different from those of the classical baselines and FERA. Thus equal ensemble size and observing density do not imply a common likelihood weight or equal tuning budget. Tables 2–4 give the available nominal FERA settings; the two retraining protocols in Appendices E.1 and F.1 state their separate training horizons and evaluation settings.

B.3 Neural Control-Transform Parameterization

Coordinate embeddings use Fourier features (Tancik et al., 2020). Local context sampling follows the coordinate-plus-feature interface of implicit representations (Chen et al., 2021), while the output remains linear in the online control coefficients.

Table 2: Architecture and query parameterization used by the reported model.

Component	Setting
State grid and tiling	69 channels on a 721×1440 latitude–longitude grid; overlapping 32×32 tiles with eight-cell overlap.
Context source	Normalized matched background during training and normalized forecast-ensemble mean online. Observations and individual member identities are not context inputs.
Context encoder	Three convolutional layers with widths $69 \to 128 \to 128 \to 64$ and SiLU activations. The resulting feature map is sampled bilinearly at each query.
Query representation	Normalized latitude–longitude, variable–level identity, normalized tile-local coordinates, global and local Fourier features, a 64-dimensional channel embedding, and a 64-dimensional sampled context feature.
Neural transform	Two multiscale MLP branches, each with six hidden layers of width 512 and SiLU activations. Each branch returns 128 coefficients, giving $d = 256$.
Scope	Grid-aligned horizontal queries and discrete model channels; continuous pressure coordinates and observation-conditioned context are not used.

B.4 Training and Climatological-Prior Settings

Table 3: Representation-learning and climatological-prior settings.

Component	Setting
Coefficient inference	Per-sample, per-tile normalized ridge projection with absolute and relative coefficients 10^{-4}; coefficients are detached for the main reconstruction path.
Objective weights	Reconstruction plus Gram diagonal/off-diagonal penalties with weights 0.05/0.005, held-out cross-variable projection with weight 0.12, and sparse-observation reconstruction with weight 0.14.
Held-out tasks	Non-humidity variables to humidity; mass/surface variables to wind; mass/wind/surface variables to temperature; and low-level cross-variable transfer.
Sparse task	Four sampled points per channel infer coefficients, which are evaluated on 1024 disjoint scalar targets.
Sampling and optimization	Eight tiles and 2048 scalar queries per training sample; AdamW for 20 epochs, learning rate 10^{-4}, batch size 1.
Climatological statistics	Tile-specific covariance of historical projected coefficients; 0.2 shrinkage toward the diagonal and $10^{-4}I$ jitter.
Excluded from this study	Explicit hydrostatic/geostrophic balance loss, tile-overlap consistency loss, tile-graph context, and variable-group masking.

B.5 Online Mean Analysis and Cycling

Table 4: Online mean-analysis and cycling-ensemble settings.

Component	Setting
Ensemble projection	512 all-channel scalar queries per tile, ridge coefficient 10^{-4}, fixed projection seed, and joint projection of all members.
Low-rank processing	Whitening by the climatological Cholesky factor, thin SVD, projection-residual/effective-rank/variance/scale gates, trace matching, active-variance shrinkage 0.3, and unit inflation.
Extended mean control	Climatological-block weight $w_c = 1.0$, ensemble-block weight $w_e = 0.3$, accepted rank $r \leq \min(d, M-1)$, and climatological fallback when a gate fails.
Mean analysis	Tile-local $D \times D$ normal equation under grid-aligned identity observations, followed by weighted overlap-add.
Cycling anomalies	Tile-local symmetric ETKF in float64 with eigenvalue floor 10^{-8} and spread inflation 1.0; tiles without observations retain forecast anomalies.
Observation sampling	Fresh seeded random horizontal columns each cycle; all 69 channels at each selected column; values copied from ERA5 truth.
Scope	Three-dimensional, grid-aligned identity observations; nonlinear operators and four-dimensional assimilation are not evaluated.

The base diagonal variance in the noise-free OSSE is

$$R_{k,\mathrm{base}} = \left(0.01\,\sigma_k^{\mathrm{model}} s_k^{\mathrm{err}}\right)^2 . \tag{27}$$

Here $\sigma_k^{\mathrm{model}}$ is the forecast model's fixed channel-normalization standard deviation in physical units, and s_k^{err} is the dimensionless channel error-scale factor in those normalized coordinates. Their product is the physical error scale S_k used by the reported transform. $R_{k,\mathrm{base}}$ is the base variance for an observation of channel k, before online reweighting. The FERA solver divides this variance by an observation weight of 6×10^{-5} and applies group multipliers of 1.0 for surface wind and temperature, 2.0 for mean-sea-level pressure, 1.5 for geopotential, 4.0 for humidity, and 0.75 for upper-air wind and temperature. These values are algorithmic weights for this OSSE rather than instrument-calibrated error statistics.

C Additional Results

C.1 Complete Single-Member Upper-Air Results

Figure 5 retains all ten combinations of z, q, u, v, t and 500/850 hPa, using the same 116 cycles and the same single-member configurations as Figure 2. Figure 2 retains $z_{500}, u_{500}, t_{500}$ to pair temporal skill with spatial diagnostics. Table 1 reports the corresponding time-mean scores for all twelve diagnostics.

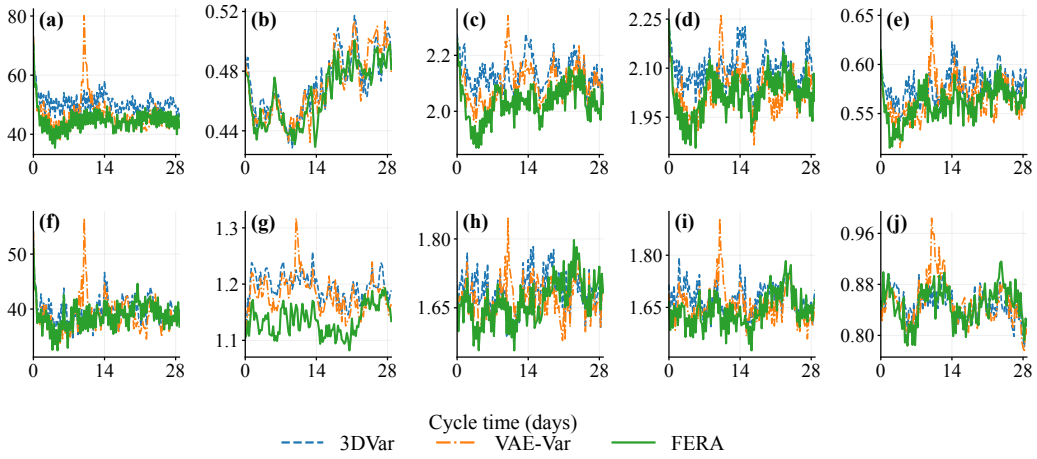


Figure 5: Complete single-member cycling results at 1000 observed columns. Columns correspond to z, q, u, v, t from left to right; the upper and lower rows are 500 and 850 hPa. Panels (a)–(e) are $z_{500}, q_{500}, u_{500}, v_{500}, t_{500}$; (f)–(j) are their 850-hPa counterparts. All curves use unfiltered per-cycle data. Vertical axes show WRMSE in $\mathrm{m^2\,s^{-2}}$ for z, $\mathrm{g\,kg^{-1}}$ for q, $\mathrm{m\,s^{-1}}$ for winds, and K for temperature.

C.2 Physical-Space Diagnostics

Sources and selection. The maps use cycles 043 and 085 of a separate field-saving cycling run (3 January–1 March 2024, $M = 1$, 1000 columns, 721×1440). This run is distinct from the main 116-cycle comparison. ERA5 supplies the reference; z_{500} is geopotential, not height. We select cycle-043 z_{500} for its strongest increment–required-correction correlation among six candidates: $z_{500}, u_{500}, t_{500}$ at the two times. We use the fixed 120°E, 35°N hemisphere, and show both cases globally. These are selected spatial examples, not a statistical validation sample.

Table 5: Physical-field checks on the original grid. WRMSE is cosine-latitude weighted and expressed in $\mathrm{m^2\,s^{-2}}$. Unobserved-grid metrics exclude only the columns observed in the current cycle. Correlation is between $x^{\mathrm{ref}} - x^b$ and $x^a - x^b$.

	Global WRMSE		Unobserved WRMSE		
Case	Background	Analysis	Background	Analysis	Corr.
13 Jan., 18 UTC	53.94	43.41	53.94	43.43	0.562
24 Jan., 06 UTC	49.36	40.70	49.36	40.72	0.549

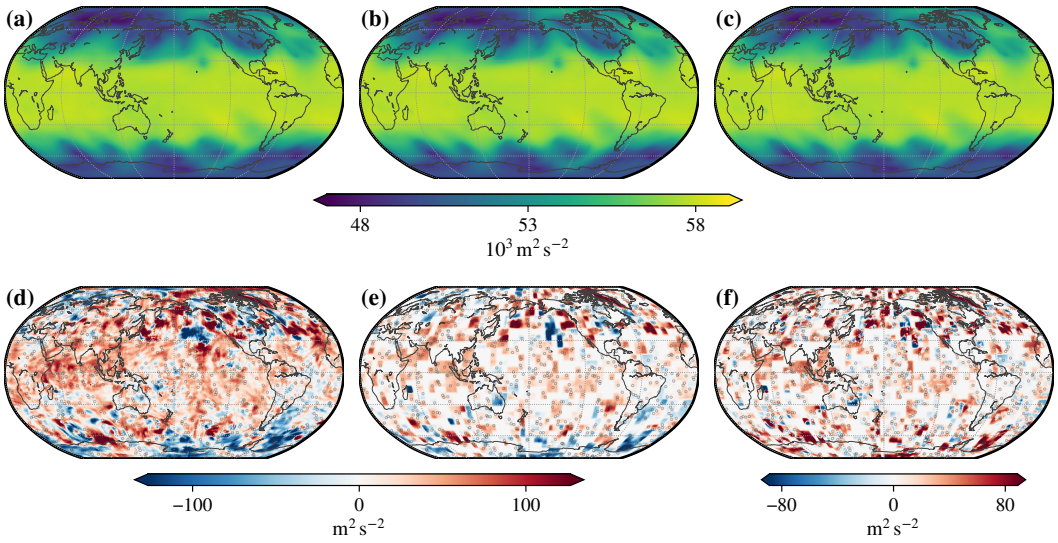


Figure 6: **Global context for the main spatial case.** Cycle 043, 13 January 2024 at 18:00 UTC: (a) ERA5 reference, (b) FERA background, (c) FERA analysis, (d) required correction, (e) analysis increment and (f) absolute-error reduction. The first row shares a geopotential scale in $10^3\,\mathrm{m^2\,s^{-2}}$; (d,e) share a signed scale in $\mathrm{m^2\,s^{-2}}$. In (f), positive values indicate improvement. Hollow points mark observed columns on the diagnostic maps. The Robinson projection also displays regions outside the hemisphere in Figure 2(d–f).

Shared display scales. Required correction and actual increment share a zero-centered scale. For Figure 8, limits are their pooled 98th percentile ($133.99\,\mathrm{m^2\,s^{-2}}$) and the reduction's 98th percentile (97.11). The six-panel global views pool both cases, using $126.72/88.88$ for correction/reduction and $47–59 \times 10^3$ for states. Colorbar extensions indicate saturation; neither sign is discarded. Basemap supplies true Orthographic and Robinson projections; bilinear resampling affects display only. Statistics use native float32 arrays with float64 reductions, not projected pixels. Main-text analysis-error/difference limits are $138.91/126.47$ (98th percentiles). The diagnostic records retain limits, saturation fractions and array hashes.

Blockwise localization and cycling performance. Block-like patterns result from separate local coefficient estimates; overlap-add does not eliminate every seam. Localization restricts distant observational influence and helps suppress spurious long-range correlations (Hunt et al., 2007). The boundaries do not prevent global WRMSE reductions (19.5% and 17.5%) or comparable unobserved-grid gains (Table 5), sustained cycling accuracy (Table 1), and next-forecast improvements (Figure 11). These system-level results do not measure localization's isolated benefit, which would require a matched ablation.

Baseline comparison and provenance. The baseline file supplies only its analysis and an ERA5 reference recorded as identical to FERA's. Figure 2(d–f) therefore compares independently cycled analyses, not shared-background updates. The VAE-Var baseline has global WRMSE 48.32 versus FERA's $43.41\,\mathrm{m^2\,s^{-2}}$. Cosine-latitude-weighted mean absolute error is also lower, decreasing from 32.95 to $30.52\,\mathrm{m^2\,s^{-2}}$ on the same native fields. Thus the spatial claim concerns the aggregate magnitude of improvements, not uniform improvement at every location.

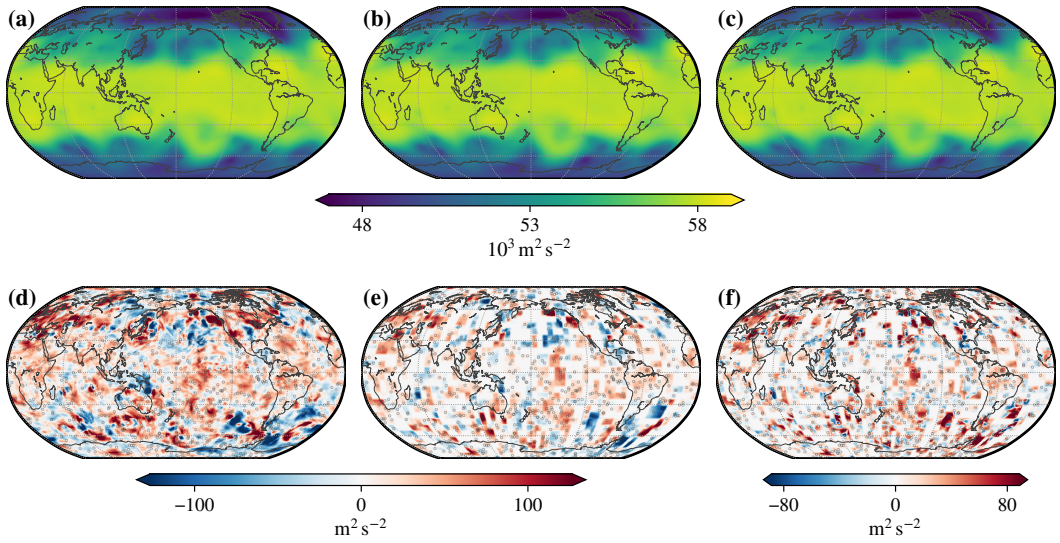


Figure 7: **A second saved analysis under the same visual conventions.** Cycle 085, 24 January 2024 at 06:00 UTC. Panel definitions, geographic view and color scales match Figure 6; both positive and negative local error changes are retained. Table 5 reports the corresponding global and unobserved-grid scores.

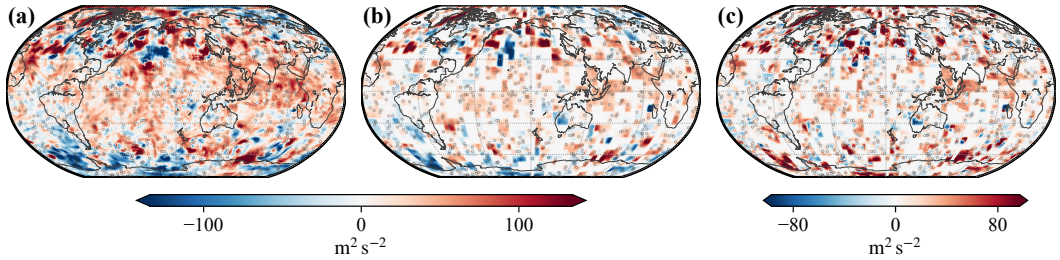


Figure 8: **Global increment structure under blockwise analysis.** At 18:00 UTC on 13 January 2024: (a) required z_{500} correction $x^{\mathrm{ref}} - x^b$, (b) FERA increment $x^a - x^b$, and (c) absolute-error reduction $|x^b - x^{\mathrm{ref}}| - |x^a - x^{\mathrm{ref}}|$. Hollow points mark current observations. The Robinson projection is centered at 180°; (a,b) share a scale. Red/blue in (c) denotes improvement/degradation. Block-boundary patterns remain visible; no additional smoothing is applied to these diagnostic fields.

C.3 Frequency of Cycle-Wise Accuracy Gains at 1000 Observations

A win is a cycle in which FERA has lower analysis WRMSE than both matched-M baselines. The variable-averaged win rate increases from 67.3% at $M = 4$ to 74.6% at $M = 8$ and 77.4% at $M = 12$ (Figure 9). The heatmap reports every diagnostic and ensemble size, retaining the variation across variables and cycles.

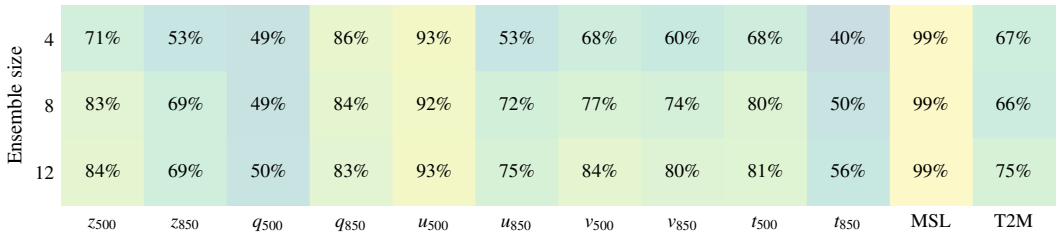


Figure 9: Cycle-wise win rate of FERA with 1000 observation columns. A win is a cycle for which FERA has lower analysis WRMSE than both matched-M baselines. Columns denote diagnostics and rows denote ensemble sizes. Each cell reports the fraction of 116 cycles won.

C.4 Observation Density and Ensemble Size

From $M = 1$ to $M = 12$, the largest FERA WRMSE reductions are for z_{500} (5.3%), z_{850} (4.6%), MSL (4.1%), and 500-hPa wind and temperature (2.8–2.9%). These changes measure end-to-end cycling, since M jointly changes backgrounds, analysis members and subsequent forecasts; Table 1 retains every diagnostic.

Figure 10 compares 500, 1000, and 2000 observed columns. For each setting, we compute FERA's relative WRMSE reduction against the best matched baseline separately for all twelve diagnostics, then average the percentages equally. This avoids mixing physical units.

The observed end-to-end benefit of larger ensembles is greater with sparse observations in these runs. At 500 columns, the balanced reduction grows from 1.8% at $M = 1$ to 5.4% at $M = 12$, with all twelve diagnostics won in every multi-member configuration. At 1000 columns it grows from 2.2% to 4.5%, reaching 12/12 wins for $M \geq 8$. At 2000 columns it levels off near 3.2–3.3%, with 9/12 multi-member wins. This trend is consistent with observations increasingly constraining the analysis directly, but changing M also changes the cycling trajectory and does not isolate the covariance block.

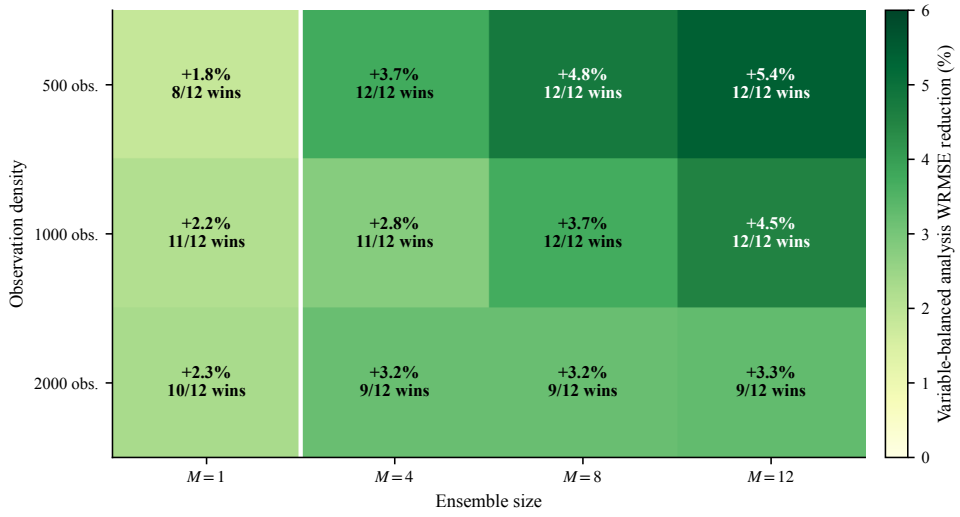


Figure 10: Interaction between observation density and ensemble size. Each cell reports the variable-balanced mean analysis-WRMSE reduction of FERA relative to the best matched baseline, followed by the number of diagnostics won out of twelve. For $M = 1$, the references are 3DVar and VAE-Var; for $M > 1$, they are Hybrid-3DEnVar and LoRA-EnVar.

C.5 Forecast Carryover of Analysis Improvements

Although forecast carryover is not a separate design objective of FERA, it provides a useful check that lower analysis error is not obtained only by fitting the current observations more aggressively. We pair each analysis at cycle t with the six-hour background at $t + 1$, before assimilating the next observations. Figure 11 reports FERA's relative WRMSE reduction against the lower-error matched-M baseline at each stage over 115 consecutive transitions.

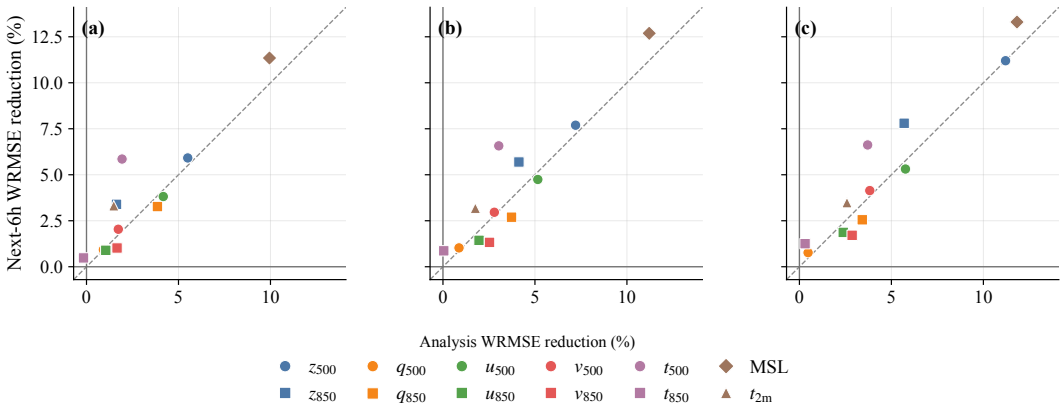


Figure 11: **Analysis improvements carry into the next background.** Relative WRMSE reductions at 1000 observed columns, paired over 115 transitions: (a) $M = 4$, (b) $M = 8$, and (c) $M = 12$. Each point is a diagnostic; each axis uses the stronger matched-M non-FERA baseline at that stage. Upper-right points improve both stages: 11/12 diagnostics for $M = 4$, and 12/12 for $M = 8, 12$.

At $M = 4$, eleven of twelve diagnostics improve at both stages; at $M = 8$ and $M = 12$, all twelve lie in the upper-right quadrant. For $M = 12$, the analysis/background reductions are $11.2\%/11.2\%$ for z_{500}, $11.8\%/13.3\%$ for MSL, and $3.7\%/6.6\%$ for t_{500}. These results show that the improved analyses remain useful after one FengWu integration. Because the two axes use stage-specific baselines and independently cycled trajectories, the percentages measure retained skill rather than a perturbation-amplification factor.

D Numerical-Consistency Replay

Section 4 uses fixed-input replay to assess the numerical robustness of direct coefficient inference. This appendix documents the replay protocol, extended iterative budgets, precision settings, and interpretation limits.

D.1 Replay Protocol and Interpretation

Data and reference definition. The replay contains one saved weather state at 00:00 UTC on 3 January 2024, a twelve-member forecast ensemble and 1000 observed columns on the 721×1440 grid. No new training or forecast integration is part of this experiment. There are 68 records in total, including method-specific reference runs and identical-input repeats. For each method and budget, the reference analysis uses the original observation order and zero control coefficients; direct FERA does not accept an optimizer initializer. Observation positions, values, channel identities and weights are reordered together. Backgrounds, networks, projection queries and priors remain fixed throughout each group.

The plotted quantity is the mean of the 69 channel-standardized spatial WRMSE values for $x^a_{\mathrm{run}} - x^a_{\mathrm{ref}}$, using the normalization and latitude weighting of the replay. Here x^a_{run} is the perturbed-run analysis and x^a_{ref} its method- and budget-matched reference analysis. This reference analysis is distinct from the ERA5 state x^{ERA5} (also written x^{ref} in the physical-field figures). The metric is not $|\mathrm{WRMSE}(x^a_{\mathrm{run}}, x^{\mathrm{ERA5}}) - \mathrm{WRMSE}(x^a_{\mathrm{ref}}, x^{\mathrm{ERA5}})|$. Per-channel measurements are retained in the experiment records. No error bars over independent weather cases are available.

Table 6: Maximum analysis-field difference under order or initialization changes. Each entry summarizes repetitions of the same weather case; direct FERA needs no optimizer initializer.

Method and budget	Reordering	Initialization
FERA direct	1.47×10^{-7}	N/A
FERA L-BFGS (20)	1.70×10^{-5}	2.83×10^{-4}
FERA L-BFGS (200)	2.55×10^{-5}	3.67×10^{-5}
VAE-Var (3)	3.41×10^{-3}	4.48×10^{-2}
VAE-Var (30)	1.14×10^{-1}	1.69×10^{-1}

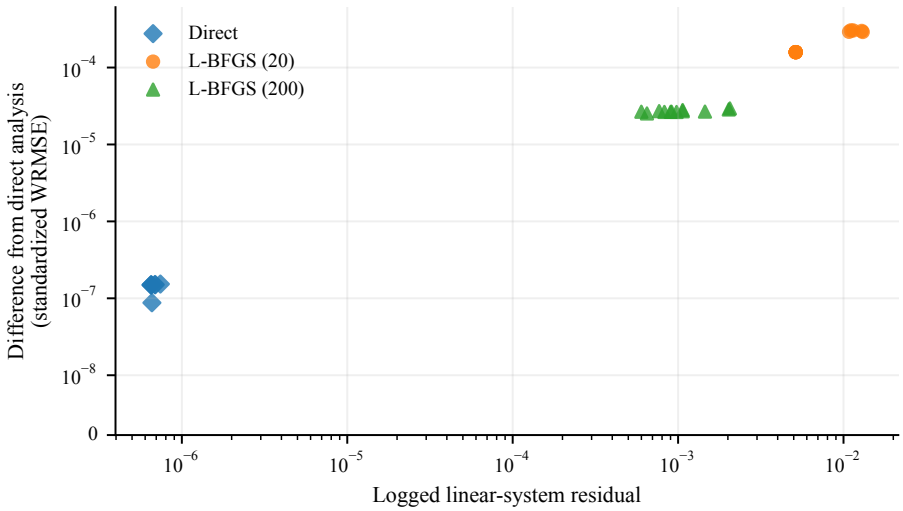


Figure 12: Same-objective FERA diagnostic. The horizontal axis is the residual supplied by the replay logger; the vertical axis compares analyses with the direct reference. The points come from one replay case. An extended iteration cap alone does not establish convergence; residuals are compared only within the same FERA objective.

Perturbations and stopping. The order test uses the reverse list and five random permutations. The initialization test uses seeds 101, 202, 303, 404 and 505, scaling each random control so that its decoded, channel-standardized physical increment has amplitude 0.05. Zero controls retain their actual method-specific physical output. The FERA iterative comparator uses the same frozen quadratic objective as the direct analysis, with local iteration caps of 20 or 200; VAE-Var uses 3 or at most 30 outer L-BFGS steps. These are method-specific budgets, not matched runtimes. All iterative outputs fail the recorded convergence check. A larger cap need not improve stationarity if the optimizer stops early. Figure 12 shows the logged same-objective FERA residuals alongside analysis differences.

Precision and configuration scope. The replay uses deterministic CUDA algorithms on an NVIDIA A100-PCIE-40GB GPU with PyTorch 2.6.0+cu124. Fields, neural operations and direct/iterative solves use float32; final objective and residual checks use float64. TF32 is disabled. Identical-input repeats have zero reported analysis-field difference. The replay configuration uses $w_c = 1$, $w_e = 12$ and 1024 projection queries per tile, distinct from the nominal settings in Table 4. Both FERA solvers use this same configuration. The VAE comparator uses an epoch-5 checkpoint with its production damping and prior settings; it is a separate solver diagnostic rather than the selected baseline used in the main accuracy comparison. Thus this diagnostic measures solver consistency within each fixed method, not a common-prior or common-likelihood accuracy ranking across methods.

Interpretation limits. At the standard VAE budget, the five random reorders have differences between zero and 5.87×10^{-7}; the reverse order gives the larger 3.41×10^{-3} maximum. At the extended budget, output differences increase even though a lower variational objective need not give a lower error against truth. This does not establish multiple converged optima or an inherent order dependence of the likelihood. The random repetitions are numerical perturbations, not samples of meteorological uncertainty. Cached preparation and batched field generation in this replay are not used as speed benchmarks. Broader, converged multi-case comparisons remain necessary to generalize the replay observation.

E FORECAST-CONDITIONING ABLATION

E.1 PROTOCOL FOR THE FORECAST-CONDITIONING ABLATION

What is held fixed. The comparison in Section 4 uses two separately trained representations with identical architecture and parameter count (3,187,200), control dimension 256, objective weights, data order, query sampling and training seed 42. FERA receives the normalized current-background tile; the control passes a fixed zero reference field in those same standardized coordinates through the *same trained context-encoder architecture*. Its true forecast remains in the background, innovations and subsequent FengWu integration. The control is therefore not an inference-only perturbation of a conditioned checkpoint. Both methods use the partial-information losses and the same linear coefficient solve.

Training and prior estimation. Both models use the first 1200 lexicographically ordered 2024 NMC error/background pairs, ending in late October, with eight tiles and 2048 scalar queries per tile. AdamW uses learning rate 10^{-4} and batch size one. The matched schedule comprises five epochs followed by five more; the optimizer is reset for both continuations because the checkpoint format does not retain optimizer state. Each representation fits its own tile-local 256×256 coefficient covariance from 256 training samples, with diagonal shrinkage 0.2 and jitter 10^{-4}. This ten-epoch experiment is distinct from the nominal twenty-epoch baseline-training configuration in Table 3. It tests the conditioned representation *with its consistently estimated prior*, not the basis alone at a common coefficient covariance.

Validation and configuration selection. The configuration search varied training horizon (five or ten epochs) and observation density (500 or 1000 columns) on the same eight-cycle window, 1 November 2024 at 00:00 UTC to 3 November at 00:00 UTC, exclusive of the endpoint. Candidates were ranked by the mean relative analysis gain over twelve fixed diagnostics, with the next-six-hour score as a tie-breaker. The reported conditioning comparison uses the selected ten-epoch, 1000-column setting.

Evaluation and temporal dependence. The December run contains all 124 analyses from 1 December 2024, 00:00 UTC through 31 December, 18:00 UTC. Both variants use $M = 1$, cycling/observation seed 0, the same initial-state configuration and noise-free random-column observations of all 69 channels. Their initial 69-channel background-WRMSE vectors agree exactly; thereafter they cycle independently. Only the checkpoint and context mode differ in the online model configuration. Observation weighting is 6×10^{-5} with identical channel-wise factors, localization and numerical safeguards. All 124 cycles are included; each model forecasts its own analysis forward without replaying a shared background trajectory.

For diagnostic j, let $\bar{r}_{\mathrm{FERA},j}$ and $\bar{r}_{\mathrm{control},j}$ denote temporal means of spatial WRMSE within the same configuration. For either representation ablation, the primary score is

$$G = \frac{100}{12} \sum_{j=1}^{12} \left(1 - \frac{\bar{r}_{\mathrm{FERA},j}}{\bar{r}_{\mathrm{control},j}} \right). \tag{28}$$

Analysis uses all 124 cycles. The secondary score uses 123 subsequent backgrounds, excluding the common initial one; these are the forecasts from the first 123 analyses. The primary and secondary scores therefore have deliberately different endpoints. We do not pool errors across physical units or average cycle-wise percentages to obtain the reported monthly score.

A paired calendar-day block bootstrap resamples 31 day blocks with replacement, retaining the within-day six-hour scores and all twelve diagnostics together. We recompute Eq. (28) for 10,000 draws, with seeds 20260921 and 20260922 for analysis and forecast. The 2.5th–97.5th percentile intervals are $[30.82, 39.95]\%$ and $[33.64, 42.93]\%$, respectively. The first forecast block contains three rather than four scores because the initial background is excluded. The intervals describe sensitivity to resampling the observed weather days with the fitted models fixed; the block length does not capture dependence beyond a day.

Implementation. The fixed-context control uses the same encoder path in training and assimilation: a zero standardized reference is passed through the encoder rather than bypassing it. This preserves the architecture and parameter count while removing state-dependent context.

E.2 Complete Scores and Primary Trajectories

The main text uses a compact two-contrast summary rather than repeating two full tables. Table 7 retains all twelve fixed forecast-conditioning diagnostics, including their physical units, absolute errors and next-forecast scores. Figure 13 retains the original geopotential, zonal-wind and temperature trajectories; the additional cases below broaden the diagnostic coverage.

Table 7: **Forecast-conditioning ablation.** December 2024 time-mean WRMSE ($M = 1$, 1000 columns). "No cond." is separately retrained without dynamic context, with its own fitted prior. Gains are relative to this control; all twelve fixed diagnostics are retained.

		Analysis (124 cycles)			Next six hours (123 forecasts)		
Diagnostic	Unit	FERA	No cond.	Gain (%)	FERA	No cond.	Gain (%)
z_{500}	$\mathrm{m^2\,s^{-2}}$	**47.59**	106.05	55.13	**56.31**	146.23	61.49
z_{850}	$\mathrm{m^2\,s^{-2}}$	**41.65**	79.14	47.37	**49.03**	104.84	53.24
q_{500}	$\mathrm{g\,kg^{-1}}$	**0.458**	0.705	35.06	**0.472**	0.747	36.79
q_{850}	$\mathrm{g\,kg^{-1}}$	**1.169**	2.044	42.80	**1.199**	2.156	44.39
u_{500}	$\mathrm{m\,s^{-1}}$	**2.154**	2.698	20.17	**2.218**	2.863	22.53
u_{850}	$\mathrm{m\,s^{-1}}$	**1.764**	2.351	24.97	**1.799**	2.456	26.75
v_{500}	$\mathrm{m\,s^{-1}}$	**2.112**	2.627	19.59	**2.189**	2.808	22.06
v_{850}	$\mathrm{m\,s^{-1}}$	**1.717**	2.235	23.15	**1.761**	2.353	25.17
t_{500}	K	**0.585**	0.921	36.49	**0.614**	1.020	39.80
t_{850}	K	**0.869**	1.263	31.23	**0.910**	1.391	34.60
MSL	Pa	**62.13**	131.05	52.59	**69.28**	142.09	51.24
t_{2m}	K	**0.958**	1.611	40.51	**0.973**	1.788	45.59
Mean relative gain		–	–	**35.76**	–	–	**38.64**

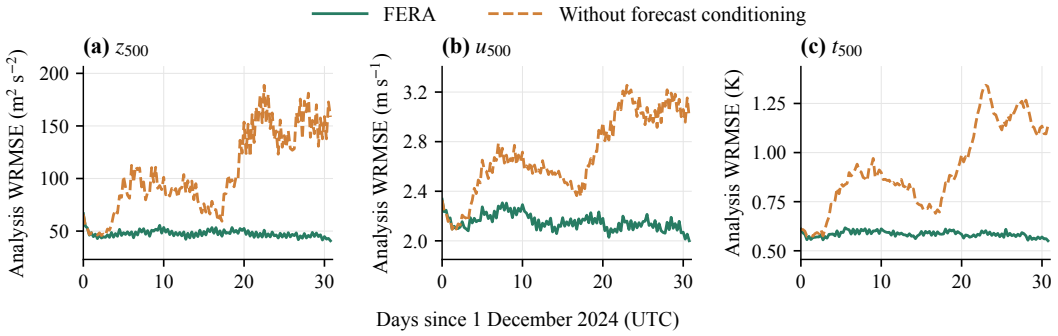


Figure 13: **Original primary forecast-conditioning trajectories.** All 124 unsmoothed December analyses for z_{500}, u_{500} and t_{500}, with ten-epoch models, $M = 1$ and 1000 columns. FERA and the retrained fixed-context control share their recorded initial background scores and then cycle independently. These trajectories are distinct from the five-epoch, 500-column loss-removal comparison in Figure 4(b,c).

Both trajectories remain finite. FERA and the fixed-context control improve their own background scores in 99.40% and 98.92% of cycle–diagnostic pairs, respectively. The full-window gap therefore does not merely reflect a control that fails to reduce its immediate background errors.

E.3 Coverage Across Variables and Levels

Figure 14 extends the fixed twelve-diagnostic comparison to every state channel. Each cell is computed from native per-cycle WRMSE, not a spatial average of local percentage improvements. All 69 channels improve in their time-mean analysis and next-six-hour scores, with reductions spanning 17.61–64.86% and 20.20–68.96%, respectively. This expanded diagnostic coverage complements the fixed twelve-diagnostic primary score.



Figure 14: **Forecast-conditioning gains across all 69 channels.** Time-mean WRMSE reductions relative to the retrained representation without dynamic forecast context. Upper-air panels contain five variables at thirteen pressure levels; the four surface channels are shown separately below. (a) All 124 analysis cycles. (b) The 123 subsequent backgrounds. Each panel uses a common 0–70% color scale, and every value is printed.

E.4 Additional Trajectories and Accumulated Gains

Figure 15 adds surface pressure, lower-tropospheric humidity, meridional wind and near-surface temperature to Figure 13, broadening the plotted variable families.

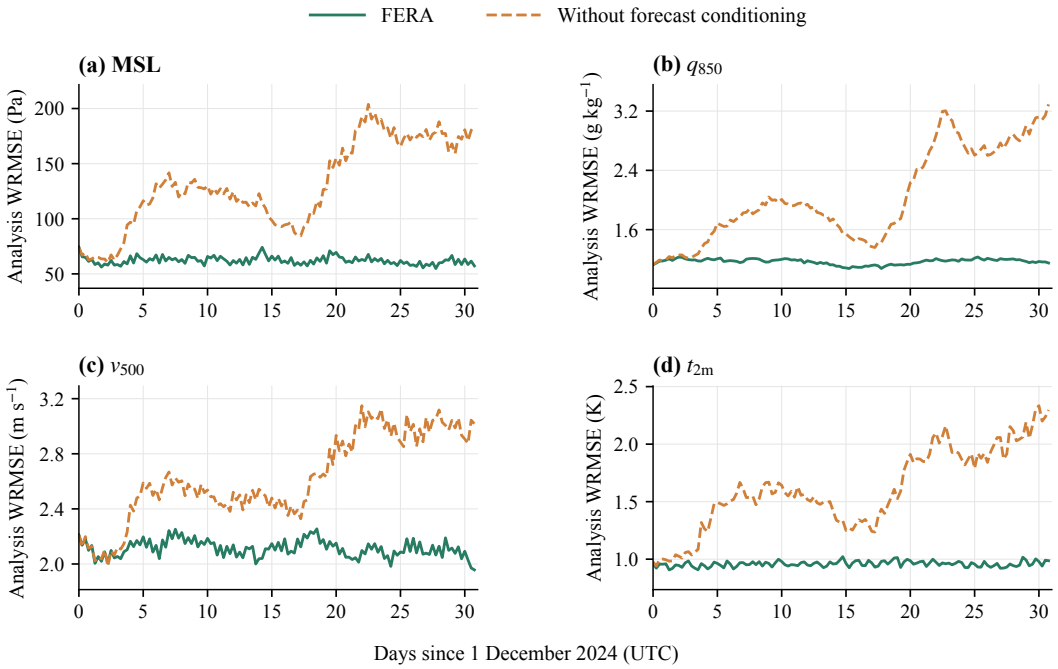


Figure 15: **Additional representative analysis trajectories.** December 2024, $M = 1$, 1000 columns, all 124 cycles, without temporal smoothing. MSL is in Pa; q_{850} is converted from kg kg^{-1} to g kg^{-1} for display. The training budget, observation sequence and online settings are matched. Independent cycling after the initial state permits accumulated feedback; these are not frozen-background interventions.

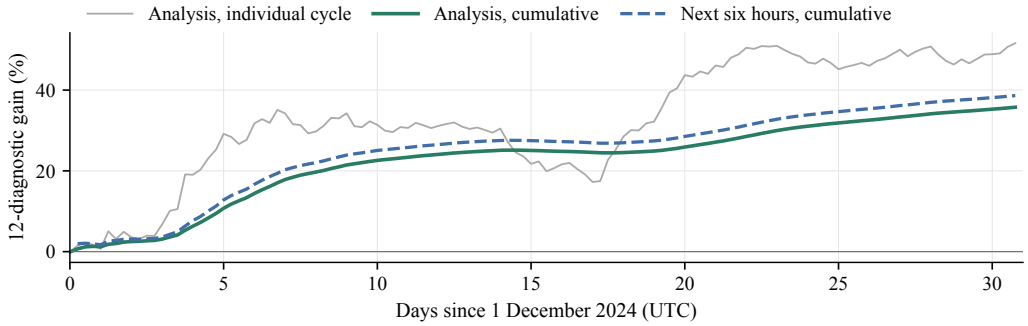


Figure 16: **Growth of the twelve-diagnostic cycling advantage.** Gray is the mean relative analysis gain at an individual cycle. The colored curves recompute Eq. (28) using all scores available up to each time, ending at 35.76% for analysis and 38.64% for next-six-hour forecasts. The cumulative forecast curve excludes the shared initial background. The gray curve and cumulative curves use different aggregation orders and should not be conflated.

F Partial-Information Training Ablation

F.1 Protocol for Partial-Information Training

Intervention and architecture. The retrained control sets both held-out and sparse-observation loss weights to zero, retaining forecast conditioning, full-field least-squares reconstruction and Gram weights $0.05/0.005$. FERA uses $\lambda_{\mathrm{hold}} = 0.12$ and $\lambda_{\mathrm{sparse}} = 0.14$. Both models use a 256-dimensional control and 1800 overlapping 32×32 tiles with overlap 8. Each model uses its own tile-local coefficient priors of shape $1800 \times 256 \times 256$, so the comparison evaluates the learned representation together with its associated prior. Both execute the mean-only branch ($M = 1$), without ensemble-derived covariance or ETKF updates.

Training and validation. The comparison uses five-epoch models and 500 observed columns. The control training schedule has 1200 steps per epoch. Validation comprises eight analyses on 1–2 November 2024 and seven subsequent backgrounds. The mean diagnostic-wise gains are 9.063% in analysis and 10.841% in the next background, with all twelve diagnostics improved. These validation scores correspond to the five-epoch, 500-column setting used in the cycling comparison.

Cycling evaluation. Both models assimilate noise-free observations of all 69 channels at 500 sampled horizontal columns per cycle. Each trajectory contains 124 six-hour analyses from 1 December 2024, 00:00 UTC through 31 December, 18:00 UTC. The initial background-WRMSE vectors agree across all twelve diagnostics; subsequent cycling is independent. Thus later differences include the feedback of each method's corrections through its own forecasts, rather than updates applied repeatedly to a shared background.

Scoring and reporting. Analysis scores average all 124 cycles. Next-forecast scores average the 123 subsequent backgrounds, excluding the initial background; the final forecast beyond December is outside this scored window. Relative gains are computed from unrounded scores using Eq. (28), with the same fixed twelve diagnostics used for validation and evaluation. Table 8 reports every diagnostic. This joint loss-removal experiment evaluates sparse spatial inference; it does not separately estimate the effect of each loss or test fully unobserved channels.

F.2 Complete Diagnostics

Table 8 contains all twelve evaluation diagnostics. The analysis reduction ranges from 36.73% for v_{500} to 73.15% for MSL, and the next-background reduction from 38.51% to 75.40%. The main figure pairs geopotential with meridional wind to show distinct variable families. All values compare the two trajectories within the five-epoch, 500-column setting; the ten-epoch conditioning experiment is not their shared reference.

Table 8: **Complete partial-information ablation scores.** December 2024, $M = 1$, 500 columns, reported five-epoch models. *No partial* removes both held-out and sparse-posterior losses; forecast conditioning is retained. Values are time-mean spatial WRMSE; bold marks the lower value within each pair. Humidity is converted to $\mathrm{g\,kg^{-1}}$ for display.

		Analysis (124 cycles)			Next six hours (123 forecasts)		
Diagnostic	Unit	FERA	No partial	Gain (%)	FERA	No partial	Gain (%)
z_{500}	$\mathrm{m^2\,s^{-2}}$	**88.83**	301.79	70.57	**99.19**	357.29	72.24
z_{850}	$\mathrm{m^2\,s^{-2}}$	**72.28**	251.45	71.25	**80.95**	302.19	73.21
q_{500}	$\mathrm{g\,kg^{-1}}$	**0.553**	1.069	48.29	**0.566**	1.195	52.61
q_{850}	$\mathrm{g\,kg^{-1}}$	**1.411**	3.408	58.60	**1.438**	3.700	61.14
u_{500}	$\mathrm{m\,s^{-1}}$	**2.883**	4.764	39.48	**2.961**	5.034	41.19
u_{850}	$\mathrm{m\,s^{-1}}$	**2.306**	4.358	47.09	**2.350**	4.666	49.64
v_{500}	$\mathrm{m\,s^{-1}}$	**2.840**	4.488	36.73	**2.917**	4.744	38.51
v_{850}	$\mathrm{m\,s^{-1}}$	**2.254**	3.720	39.42	**2.298**	3.803	39.57
t_{500}	K	**0.826**	2.440	66.15	**0.863**	2.859	69.83
t_{850}	K	**1.120**	2.261	50.46	**1.165**	2.472	52.88
MSL	Pa	**102.65**	382.36	73.15	**112.17**	456.02	75.40
$t_{2\mathrm{m}}$	K	**1.179**	3.732	68.41	**1.203**	4.000	69.93
Mean relative gain		–	–	**55.80**	–	–	**58.01**

The cumulative curve in Figure 4(a) recomputes ratios of temporal means at each horizon, rather than averaging cycle-wise percentages. Its forecast score excludes the common initial background.

F.3 Additional Representative Cycling Trajectories

The additional cases in Figure 17 span surface mass, lower-tropospheric moisture, mid-tropospheric temperature and lower-level wind. MSL provides the largest diagnostic response, but is reported alongside variables selected for distinct physical roles, not as the sole representative case. These are complete trajectories with identical plotted endpoints and no temporal smoothing; favorable and unfavorable individual cycles are retained.

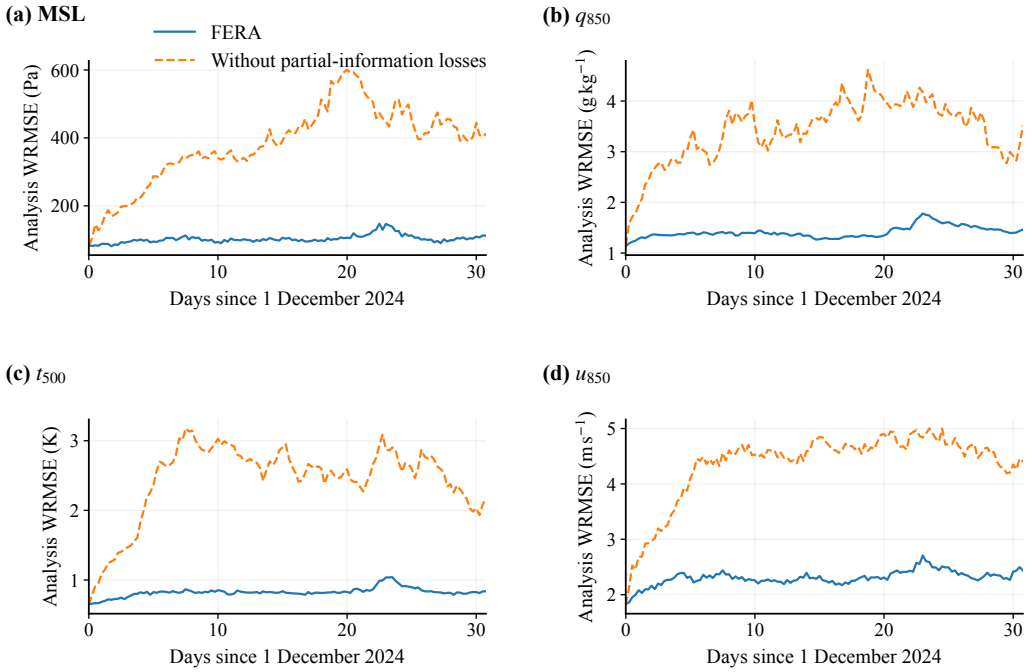


Figure 17: **Partial-information training across complementary variable families.** Time-resolved analysis WRMSE for all 124 December cycles. FERA is compared with the retrained model without held-out and sparse-posterior losses; both retain forecast context. MSL is in Pa, q_{850} in $\mathrm{g\,kg^{-1}}$, t_{500} in K, and u_{850} in $\mathrm{m\,s^{-1}}$. These independently cycled trajectories include analysis–forecast feedback and do not identify a shared-background single-update effect.

F.4 Early Propagation and Logged Coefficient Activity

For diagnostic j and transition t, let $r^a_{j,t}$ be analysis WRMSE at cycle t and $r^b_{j,t+1}$ the next background's WRMSE, both against ERA5. Define the descriptive ratio

$$\rho_{j,t} = \frac{r^b_{j,t+1}}{r^a_{j,t}}. \tag{29}$$

The first-eight-transition mean averages these ratios over eight transitions and twelve diagnostics, giving 1.052 for FERA and 1.177 without partial-information losses. It is not the ratio of pooled cross-variable errors, not the monthly gain in Eq. (28), and not a dynamical stability certificate. The numerator and denominator refer to different valid times and different reference fields, so the ratio describes recorded error growth rather than an isolated perturbation-amplification factor.

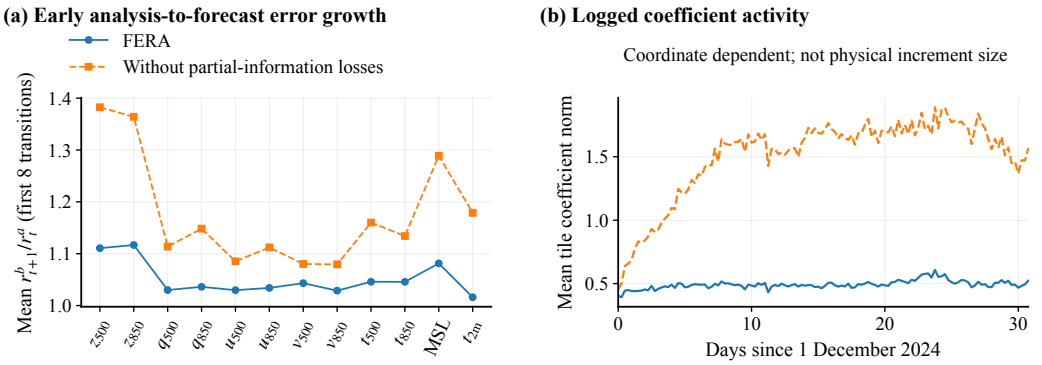


Figure 18: **Auxiliary evidence on propagation and coefficient activity.** (a) First-eight-transition means of Eq. (29), retaining all twelve diagnostics. (b) Unsmoothed mean tile coefficient norms extracted from the 124-cycle console logs. The latter live in separately learned coordinate systems and cannot be read as directly comparable physical increment sizes or prior-normalized distances. The mean-only setting has no active ensemble quality-gate attempts.

FERA's logged mean coefficient norm ranges from 0.3923 to 0.6079. The control begins at 0.4543, reaches a maximum of 1.891, and ends at 1.568; it does not grow monotonically. This co-occurs with larger accumulated physical-space errors, but does not by itself prove that large coefficients cause those errors. Latent coordinate scaling and independently fitted priors also affect coefficient norms.

The fraction of analysis scores below each method's own background score is 99.19% for FERA and 98.92% for the control. The weaker trajectory therefore usually still benefits from its local analysis step. The monthly gap is compatible with different forecast consequences of the corrections, while leaving the basis/prior decomposition and single-step mechanism unresolved.

Table 9: **Complete diagnostic-wise transition ratios.** Temporal means of $r^b_{j,t+1}/r^a_{j,t}$ for the partial-information comparison. The ratio is dimensionless; averaging ratios is distinct from the relative gain of time-mean WRMSE in Table 8. *No partial* denotes removal of both partial-information losses.

	First 8 transitions		All 123 transitions	
Diagnostic	FERA	No partial	FERA	No partial
z_{500}	1.111	1.382	1.118	1.191
z_{850}	1.117	1.364	1.122	1.207
q_{500}	1.030	1.114	1.025	1.118
q_{850}	1.036	1.148	1.020	1.088
u_{500}	1.030	1.085	1.027	1.058
u_{850}	1.034	1.112	1.020	1.073
v_{500}	1.043	1.080	1.028	1.058
v_{850}	1.029	1.079	1.020	1.024
t_{500}	1.046	1.160	1.045	1.170
t_{850}	1.046	1.134	1.040	1.094
MSL	1.081	1.289	1.094	1.195
t_{2m}	1.016	1.179	1.021	1.081
Mean over diagnostics	1.052	1.177	1.048	1.113