ENHANCING DIFFUSION WEATHER MODEL'S PHYSICAL CONSISTENCY WITH TEST-TIME GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion-based data-driven weather models now surpass numerical weather prediction in forecast skill, but the meteorology community still does not fully trust their output, because the predictions can violate basic physical principles. Violations such as hydrostatic imbalance and dry-air mass drift not only make a single forecast step look questionable, they also compound over the autoregressive rollouts that operational forecasting relies on. Physics-informed training or fine-tuning can reduce these violations, but the computational cost can be prohibitive. Recent advances in test-time steering offer another opportunity to enforce physical principles, by guiding a frozen diffusion model during sampling. Existing methods such as Manifold Preserving Guided Diffusion, however, measure the violation on a coarse estimate of the final state, which limits how much they can correct. We therefore propose Intermediate Gradient Guidance (IGG), a training-free, plug-and-play method that measures the violation on the intermediate state itself and steers each sampling step of a frozen diffusion model toward states that better satisfy a target physical constraint. Our experiments show that IGG improves physical consistency by up to 72.4%, with no loss in forecast skill and negligible extra computation.

1 INTRODUCTION

Generative weather models now issue medium-range ensemble forecasts that rival operational numerical systems. Trained on decades of weather data, probablistic models such as GenCast (Price et al., 2025) and stochastic-interpolant models such as Atlas (Albergo et al., 2025) draw each forecast step from a learned conditional distribution, and GenCast outperforms ECMWF's operational ensemble numerical model on most of the weather variables. However, matching the data distribution does not, make a forecast physically consistent. Nothing in the training objective requires the output variables to satisfy the physical laws between them. Data-driven forecasts depart from hydrostatic and geostrophic balance and drift in dry-air mass (Bonavita, 2024; Kasteleyn et al., 2026), and so do both models we study (Sec. 4). Building these laws into training (Sec. 2.1) means retraining large models, which is expensive for their developers and out of reach for most users of a released checkpoint.

Models with sampler offer a cheaper remedy in one of two ways: drawing several samples and keeping the most consistent, or steering each sample as it is generated. Selection can only choose among the errors the model's samples make, and on GenCast the hydrostatic imbalance is shared by all of them. Even the best of eight members, chosen with knowledge of ERA5, is at most **2**% closer to ERA5's balance than the average member at any lead time (Fig. 1).

Existing gradient-based guidance methods (He et al., 2024; Chung et al., 2023; Rozet et al., 2024) steer samples routinely for inverse problems in imaging. In weather, prior work (Aich et al., 2026) used Moment Matching Posterior Sampling (MMPS) (Rozet et al., 2024) to hold global dry-air mass at its initial value through a multi-year rollout. However, it guides a diffusion model that its authors trained as a prior for inverse problems, and it constrains a single global scalar. Forecasters in practical use are frozen checkpoints built on different samplers, and the atmosphere's balance laws constrain every grid column.

This raises the question: *is there a test-time guidance method that works across, architectures and physical constraints at negligible cost?* Cost is the reason to guide rather than retrain, and it rules out

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.

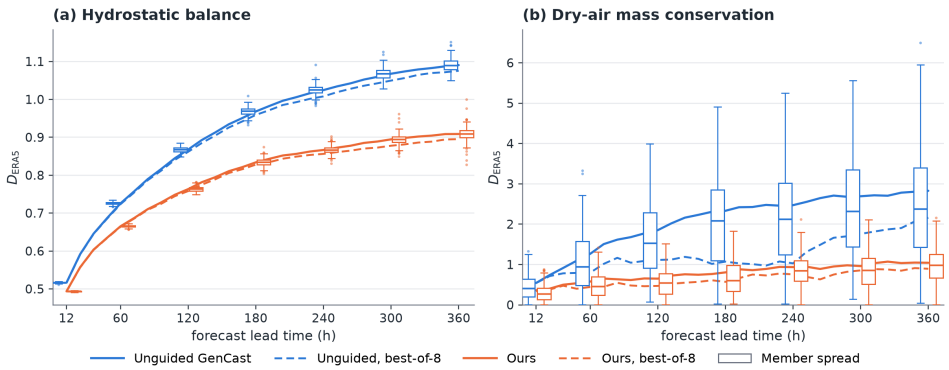

Figure 1: **Guidance lowers D_{ERA5} at every lead time, below even the best unguided member.** D_{ERA5} (lower is better) on the held-out test set on GenCast for (a) hydrostatic balance and (b) dry-air mass conservation . Solid lines show the RMS over dates and members. Dashed lines show the best-of-8 oracle: for each date, the member with the lowest D_{ERA5} over 12–360 h. It is selected using ERA5, so it is an upper bound on member selection. Boxes show the spread across members, after rescaling each date to the all-date RMS.

guidance methods requiring back-propagation through the network at sampling step, or derived for the Gaussian denoising posterior of a diffusion model, so they do not transfer to a stochastic interpolant model, without re-deriving the method. Manifold Preserving Guided Diffusion (MPGD) (He et al., 2024) avoids both problems: it differentiates the objective at the sampler's clean estimate rather than through the network, so it costs one gradient of the physics operator per step and applies to any sampler that produces a clean estimate. However, that the gradient is taken at the clean estimate, is not the best proxy for the state the sampler actually carries forward in high-dimension guidance objective.

We propose Intermediate Gradient Guidance (IGG), which inherits MPGD's strengths and improve its guidance performance with simple change. IGG evaluates the objective on the sampler's intermediate state, so it needs no posterior mean, network Jacobian, and applies to any sampler that exposes its state and update. Next, it sets each correction to a fixed fraction λ of the norm of the sampler's update. We evaluate our IGG on two state-of-the-art probabilistic weather forecasting backbones with different physics guidance objectives, and compare it with several remarkable gradient-based steering methods, including MPGD itself.

Our contributions are:

- We propose IGG, a training-free method that needs only a differentiable residual and runs with one setting on both a diffusion model and a stochastic interpolant.
- We identify two different shorthands on current gradient-based steering methods: clean estimation and unmatched gradient. And through these two modification on MPGD, we improved D_{ERA5} up-to 28.4% more than original MPGD.
- Guiding hydrostatic balance cuts its error against ERA5 by 16.7% (GenCast) and 30.3% (Atlas) on day-15, at a cost of 1.09% in geopotential at 500 hPa ($z500$) CRPS at 12 h. Guiding dry-air mass cuts its error by 63.1% (GenCast), 67.3% (Atlas) on day-15.
- IGG in hydrostatic guidance recovers 94 % of the excess D_{ERA5}, 87 % of the z_{500} CRPS lost at 12 h due to down-precision from TF32 to BF16.

2 Related Work

2.1 Physical consistency in deep learning weather forecasting

Deep learning forecasts are not constrained by the governing equations, and their violations grow along autoregressive rollouts (Kasteleyn et al., 2026). Existing remedies differ in how much access to the model they require. Architectural approaches build physics into the forward pass through

a differentiable dynamical core (Kochkov et al., 2024), interleaved PDE kernels (Xu et al., 2024) or conservation layers (Sha et al., 2025); optimisation approaches fine-tune a pretrained forecaster with a physics-informed (Luo et al., 2025) or error-tolerant hydrostatic loss (Subramaniam et al., 2025). Both require training. The training-free alternative is post-hoc correction, such as global mass fixers (**?**) or re-diagnosing geopotential from temperature, which overwrites part of the forecast with a quantity diagnosed from the rest. No existing method improves the physical consistency of a frozen generative forecaster within its sampling process.

2.2 Test-time guidance of generative samplers

Training-free guidance(Ye et al., 2024) steers a pretrained sampler toward an objective in one of two ways. *Particle-based* methods, such as Feynman–Kac steering (Singhal et al., 2025) and best-of-k selection, resample trajectories by reward. They need no gradient but multiply sampling cost by the number of particles, which is prohibitive for a global pretrained model. *Gradient-based* methods instead add the gradient of a differentiable objective to each sampler update, without extra trajectories.

Posterior sampling methods evaluate the constraint on the Tweedie estimate $\hat{x}$ and backpropagate it through the denoiser (Chung et al., 2023). MMPS (Rozet et al., 2024) additionally propagates the denoiser's posterior covariance through a conjugate-gradient solve that takes a vector–Jacobian product through the network at every iteration. WIND (Aich et al., 2026) applies MMPS to a video-diffusion atmospheric prior and holds global dry-air mass at its initial value over a four-year rollout without degrading 14-day CRPS, but slows sampling by 5.7–$10.5\times$ (Aich et al., 2026, Table B.1). Its success criterion, whether the scalar holds its target, is one that post-hoc projection meets by construction; we instead score against the ERA5 residual (Sec. 2.1). Flow-model adaptations (Pokle et al., 2024; Feng et al., 2025) assume a Gaussian probability path, which the interpolant used by Atlas does not provide.

MPGD (He et al., 2024) removes the network from the gradient by differentiating the objective with respect to $\hat{x}$ rather than the intermediate state. This keeps the sample near the data manifold at the cost of one backward pass through the objective alone, which is why we build on it (Sec. A.2). And our method further improved it by using intermediate guidance and gradient norm matching.

3 Methodology

3.1 Problem formulation and notation.

We define the atmospheric state at time k as $X^k \in \mathbb{R}^{C\times H\times W}$, where C is the number of variables, and $H,\ W$ denote the spatial resolution. A forecast model maps X^k to $X^{k+\Delta k}$ for a fixed physical time stride Δk. For autoregressive rollouts K steps, the model f takes its own previous prediction as input for the next forecast. Both backbones condition on the two most recent states:

$$X^{k+1} = f(X^{k-1}, X^k) \qquad k = 0, ..., K-1$$

f is probabilistic model that generate X^{k+1} by sampling iteratively under a learned network. Our aim is to steer that sampling so that X^{k+1} is more consistent with a set of physical constraints, without retraining. We implement our method on two pretrained Deep learning weather medium-range models, GenCast (Price et al., 2025) and Atlas (Kossaifi et al., 2026); configuration (C, stride, resolution) are given in Sec 4.

3.2 Intermediate gradient guidance

Both samplers iterate a state $\boldsymbol{z}$ over a progress variable $s \in [0, 1]$ and reach the clean end at $s = 1$. For GenCast, $\boldsymbol{z}$ is the standardised residual on the 1° grid and starts from noise; for Atlas, it is the 1° latent state, which the interpolant carries from the standardised current state at $s = 0$ to the standardised residual at $s = 1$. Each network evaluation yields a clean estimate $\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}, s; \boldsymbol{c})$ with $\boldsymbol{c} = (X_{k-1}, X_k)$: the denoiser output for GenCast, and $\hat{\boldsymbol{x}}_\theta = \boldsymbol{z} + (1 - s)\, \boldsymbol{b}_\theta$ for Atlas's linear interpolant with drift $\boldsymbol{b}_\theta$. $\mathrm{Decode}(\boldsymbol{z}; X_k)$ de-standardises the residual and adds it to the current state X_k; it is affine and contains no network (Atlas's decoder to 0.25° acts only after sampling).

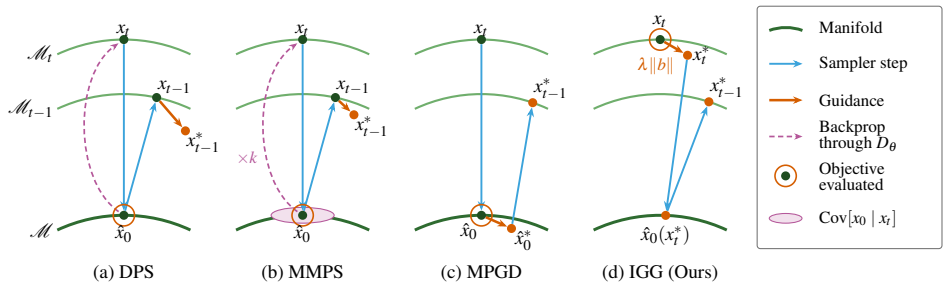

Figure 2: Where each method evaluates the physics objective (ring) and applies its correction (orange). DPS and MMPS differentiate through the network, MMPS with k vector–Jacobian products per step. MPGD corrects the clean estimate $\hat{\boldsymbol{x}}$ directly. Ours evaluates the objective on the intermediate state $\boldsymbol{z}$, uses $\hat{\boldsymbol{x}}$ only to set the step length, and scales the correction to a fraction λ of the sampler's own update. Illustration inspired by He et al. (2024).

Guidance adds to the output of each sampler iteration a step against the gradient of the objective

$$P(\boldsymbol{z}) = d\big(\,\mathrm{Decode}(\boldsymbol{z}; X_k),\, X_0\big), \tag{1}$$

the residual distance of Eq. 2 between the intermediate state and the initial condition (Algorithm 1). Two choices distinguish it from MPGD, and Sec. 4.4 ablates each.

State (noisy operand). P is evaluated on the state $\boldsymbol{z}$ that enters the network, not on the clean estimate $\hat{\boldsymbol{x}}_\theta$ as in MPGD; the clean estimate enters only through the step length. The motivation is consistency: the correction is applied to the sampler's intermediate state, so we take the gradient on that state rather its clean estimation (Fig. 2). Sec. 4.4 shows this choice is what turns MPGD's short-lead loss into a gain. The gradient $\nabla_{\boldsymbol{z}} P$ passes through Decode and $\mathcal{R}$ only, never through the network, so the overhead over unguided sampling is one evaluation of $\mathcal{R}$ and one backward pass through it per iteration.

Magnitude (norm matching). MPGD steps by a constant c_s in data units. We instead move the iteration's output by $\lambda \,\|\hat{\boldsymbol{x}}_\theta - \boldsymbol{z}\|$ along $-\nabla_{\boldsymbol{z}} P / \|\nabla_{\boldsymbol{z}} P\|$, norms over the full state of each member. λ is dimensionless, the correction in units of the member's remaining distance to its clean estimate. That distance vanishes as $s \to 1$, so the correction anneals without a schedule and λ is constant; the scale selected at TF32 is used unchanged at BF16 (Sec. 4.5). The step is applied once per iteration, at its last network evaluation, rescaled so that the iteration's output moves by exactly this amount (Appendix D).[1]

Algorithm 1 Intermediate gradient guidance, one sampler iteration at forecast step k

Require: state $\boldsymbol{z}$ at progress s entering the iteration's last network evaluation; conditioning $\boldsymbol{c} = (X_{k-1}, X_k)$; clean estimate $\hat{\boldsymbol{x}} \leftarrow \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}, s; \boldsymbol{c})$; weight w of this evaluation in the iteration's output; ratio λ

$\boldsymbol{x} \leftarrow \mathrm{Decode}(\boldsymbol{z}; X_k)$ ▷ physical units; affine, no network
$P \leftarrow d(\boldsymbol{x}, X_0)$ ▷ Eq. 2, with $y = \mathcal{R}(X_0)$ frozen
$\boldsymbol{g} \leftarrow \nabla_{\boldsymbol{z}} P$ ▷ through Decode and $\mathcal{R}$ only
$\boldsymbol{\delta} \leftarrow -\lambda \,\|\hat{\boldsymbol{x}} - \boldsymbol{z}\|\, \boldsymbol{g} / \|\boldsymbol{g}\|$ ▷ per member
return unguided output $+ \boldsymbol{\delta}$ ▷ realised as $\boldsymbol{\delta}/w$ added to $\hat{\boldsymbol{x}}$ (GenCast) or to $\boldsymbol{b}_\theta$ (Atlas)

4 Experiments

We evaluate intermediate gradient guidance on two pretrained backbones, GenCast and Atlas, without retraining or fine-tuning either. The experiments address four questions:

[1] The implementation differentiates $\sqrt{P^2 + \epsilon}$ so the gradient is finite at zero residual; only its direction is used, so P and P^2 give the same step.

- Does guidance make forecasts more consistent with the constraint it is given, and does that consistency reflect the atmosphere rather than only the objective (Sec. 4.2)?
- Which component of the guidance carries the effect (Sec. 4.4)?
- Does our method transfer to a second backbone, including reduced numerical precision (Sec. 4.5)?

4.1 Experimental setup

Backbones and protocol. We guide two released checkpoints without retraining: GenCast (Price et al., 2025) and Atlas (Kossaifi et al., 2026). Every run is an 8-member, 15-day rollout from a 00 UTC ERA5Hersbach et al. (2020) initial condition, verified against ERA5 at the valid time. Guidance scales are chosen on a development set of six March 2023 dates (days 1, 6, ..., 26) and evaluated once on a held-out test set: the same days of June, September and December 2023 and March 2024 (24 dates) for GenCast, and the six March 2024 dates for Atlas. Checkpoints and sampler settings are in Appendix D.

Constraints. The hydrostatic target is the discrete residual of hydrostatic equation B.1 on the ten adjacent pressure pairs between 50 and 850 hPa; the dry-air-mass target is the global dry-air mass M_d, reconstructed from mean sea-level pressure and specific humidity on both GenCast and Atlas B.2. Both targets are taken relative to the initial condition, $r(X; X_0) = \mathcal{R}(X) - \mathcal{R}(X_0)$, and standardized by scales measured on ERA5 (Appendix C).

Methods. All methods solve the same inverse problem, $y = \mathcal{R}(X_0)$ in the standardised coordinates of Appendix E, and differ only in where the gradient is taken and how the step is scaled.

- **Unguided.** The released sampler; the reference for every number reported.
- **Projection.** No guidance; the completed state is projected exactly onto $r(X; X_0) = 0$ after every autoregressive step. Both backbones; the limiting case that motivates scoring against ERA5 rather than the initial condition.
- **DPS** (Chung et al., 2023). Gradient of the objective at the denoised estimate, back-propagated through the network; step normalised by the residual norm. GenCast only, since it assumes a Gaussian denoising posterior.
- **MMPS** (Rozet et al., 2024). Posterior sampling with the covariance of the denoising posterior retained; $2n + 2$ network backward passes per denoiser evaluation for n solver iterations. The method WIND (Aich et al., 2026) uses for conservation laws; we run FA-APF's GenCast port. GenCast, mass task only.
- **MPGD** (He et al., 2024). Gradient at the denoised estimate without the network Jacobian; constant step in data units. The method ours is derived from; both backbones.
- **Ours.** MPGD with the gradient taken at the noisy state and the step set to a fraction λ_0 of the sampler's own update (Sec. 3.2); both backbones.

Metrics. Forecast skill is the fair CRPSFerro (2014); Gneiting & Raftery (2007) of the 8-member ensemble, cosine-latitude weighted on each model's native grid; we show z_{500} and MSLP at 12 and 120 h and list the other variables in Appendix H. Physical consistency is scored by a residual distance in units of the calibration scales,

$$d^2(X, Y) = \frac{1}{L} \sum_{\ell=1}^{L} \sum_{i} w_i \Big[\frac{r_\ell(X; Y)_i}{\sigma_\ell} \Big]^2, \; D_{\mathrm{ERA5}}(t) = \Big[\tfrac{1}{KM} \textstyle\sum_{k,m} d^2(X_t^{k,m}, X_t^{k,\mathrm{ERA5}}) \Big]^{1/2} \quad (2)$$

with $L = 10$ pairs for the hydrostatic target and $L = 1$ for mass, pooled over K initializations and M members. D_{IC} is the same distance to the initial condition, i.e. the guidance objective itself. We rank on D_{ERA5}, integrated over 12–360 h, and not on D_{IC}: the atmosphere drifts away from its initial residual structure, so $r(X; X_0) = 0$ is not correct, and projection, which enforces it exactly, is the limiting case (Sec. 4.2). Our hydrostatic residual evaluation is extended from the PhysMetrics.Weather hydrostatic score (Kasteleyn et al., 2026).

Table 1: Hydrostatic balance guidance on the test set. Entries are % change relative to the unguided baseline, $100\,(1 - \text{method}/\text{baseline})$, so higher is better in every column.

		D_{ERA5}				CRPS			
						12 h		120 h	
Model	Method	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
GenCast	Projection	-6.1	-49.1	-12.8	$+7.6$	-15.59	$+0.00$	-8.63	-6.34
	DPS	$+11.5$	-6.2	$+9.6$	$+16.0$	-2.45	-0.03	-0.72	-0.55
	MPGD	$+10.6$	-4.0	$+9.0$	$+14.6$	-1.97	-0.00	-0.58	-0.41
	Ours	**+13.6**	**+4.4**	**+12.0**	**+16.7**	**-1.09**	**+0.01**	**-0.50**	**-0.37**
Atlas	Projection	-66.1	-50.8	$+0.6$	-196.2	-15.96	-5.42	-110.23	-79.19
	MPGD	$+20.9$	-7.0	$+18.5$	$+26.5$	-2.23	-0.01	-0.39	-0.07
	Ours	**+26.9**	**+21.4**	**+24.8**	**+30.3**	**+0.71**	**+0.34**	**-0.30**	**+1.20**

Table 2: Dry-air-mass guidance on the test set. Entries are % change relative to the unguided baseline, $100\,(1 - \text{method}/\text{baseline})$, so higher is better in every column. The dry-air mass is computed with the same derived operator on both backbones (surface pressure reconstructed from mean sea-level pressure, moisture integrated from specific humidity; Appendix B.2).

		D_{ERA5}				CRPS			
						12 h		120 h	
Model	Method	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
GenCast	Projection	$+58.6$	$+1.9$	$+63.8$	$+63.5$	$+0.00$	-1.23	$+0.20$	$+0.73$
	DPS	$+60.7$	$+17.1$	$+63.7$	$+63.7$	**+0.51**	-0.47	**+0.34**	$+0.74$
	MPGD	$+60.7$	$+10.7$	**+64.9**	**+64.1**	$+0.27$	-0.45	$+0.28$	**+0.80**
	MMPS	$+56.0$	-9.4	$+53.9$	$+59.8$	$+0.37$	-0.16	-0.86	$+0.41$
	Ours	**+62.5**	**+33.4**	**+64.9**	$+63.1$	$+0.15$	**-0.04**	$+0.27$	**+0.80**
Atlas	Projection	$+52.8$	$+28.5$	$+65.3$	$+58.2$	**+0.10**	-1.51	$+0.01$	-0.00
	MPGD	$+54.1$	$+29.5$	$+66.7$	$+61.1$	$+0.04$	-1.40	$+0.06$	**+0.02**
	Ours	**+67.3**	**+67.8**	**+72.4**	**+67.3**	-0.45	**-0.20**	**+0.11**	-0.01

4.2 Physical consistency under guided generation

Hydrostatic balance. *Guidance reduces the imbalance at every lead, and the gain grows with lead.* Unguided GenCast drifts away from ERA5's balance structure along the rollout (Fig. 1): D_{ERA5} grows from $0.52\,\sigma$ at 12 h to $1.09\,\sigma$ at day 15. Ours holds it to 0.49, 0.76 and $0.91\,\sigma$ at 12 h, day 5 and day 15, a reduction of 4.4, 12.0 and $16.7\,\%$ (Table 1). On Atlas the same ordering holds with about twice the gain ($+26.9\,\%$ on the integral, $+21.4\,\%$ at 12 h). The gain grows with lead because the guidance acts at every autoregressive step and removes a drift that accumulates rather than a fixed offset.

Every gradient method helps at long lead; only ours helps at 12 h. DPS, MPGD and ours all reduce the integral by 11–14 % on GenCast and are within 3 points of each other at day 15. They separate at short lead: DPS and MPGD worsen D_{ERA5} at 12 h (-6.2 and $-4.0\,\%$); they satisfy the stated objective by pulling the forecast toward the initial condition's residual structure faster than the atmosphere moves away from it. Ours is the only method that improves D_{ERA5} at 12 h ($+4.4\,\%$ GenCast, $+21.4\,\%$ Atlas).

Projection is the limiting case. Setting $r(X; X_0) = 0$ exactly after every step pins $r(X; X_0)$ equal to ERA5's own drift from the initial condition: $-49\,\%$ at 12 h, $+7.6\,\%$ at day 15 on GenCast, at a $15.6\,\%$ z_{500} CRPS cost at 12 h because the correction is spent in geopotential. On Atlas the projected state diverges ($-196\,\%$ at day 15). Best-of-k recovers at most $1.1\,\%$ (GenCast) and $0.5\,\%$ (Atlas).

Dry-air mass conservation. Unguided GenCast loses dry-air mass along the rollout: D_{ERA5} grows from $0.54\,\sigma_M$ at 12 h to $2.83\,\sigma_M$ at day 15, and, unlike the hydrostatic residual, the drift differs strongly between members, so the best of eight unguided members already removes $41.5\,\%$

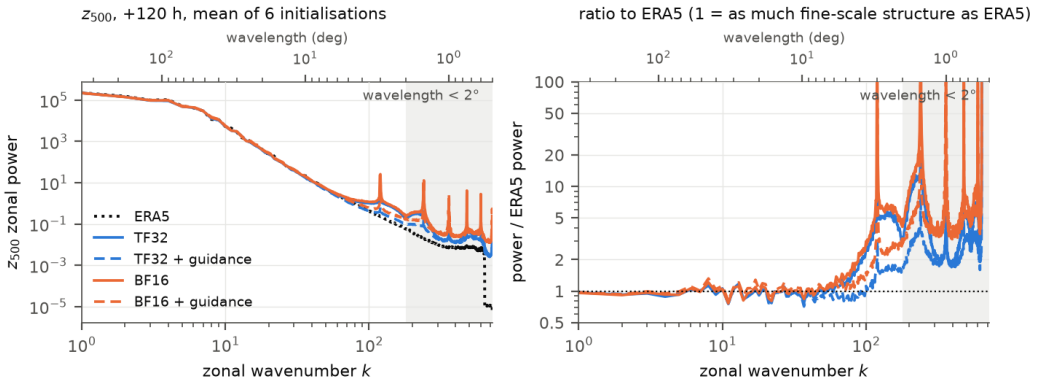

Figure 3: (Atlas with hydrostatic guidance) Zonal power spectrum of z_{500} at +120 h on the test dates (mean over dates of the eight-member mean spectrum), Left: power against zonal wavenumber k; right: ratio to the ERA5 spectrum at the same valid time. The shaded band marks wavelengths below 2°.

of the integral (GenCast) and 54.7 % (Atlas). Guidance removes 56–63 % on GenCast and 53–67 % on Atlas (Table 2). On GenCast every guided method beats selection; on Atlas, where the member spread is larger, only the two methods with the noisy operand do, and ours by a wide margin.

On GenCast the methods tie on the integral (DPS, MPGD and ours are within the bootstrap interval of each other, at +60.7 to +62.5 %), because the integral saturates at ERA5's own mass wobble. They separate at short lead and in skill. At 12 h ours removes 33.4 % of the distance against 17.1 % for DPS and 10.7 % for MPGD, and ours is the only guided method whose MSLP CRPS at 12 h is indistinguishable from the baseline (−0.04 % vs. −0.47 and −0.45 % for DPS and MPGD). Projection removes 58.6 % at a 1.2 % MSLP cost; it also removes the real semi-diurnal cycle of the reconstructed mass, which is why pinning is not the D_{ERA5} optimum.

On Atlas the same ordering holds with wider gaps. Ours removes 67.3 % of the integral. At 12 h ours removes 67.8 % of the distance against 28–30 % for MPGD and projection. The skill cost is again smallest for ours: MPGD and projection lose 1.4–1.5 % of MSLP CRPS at 12 h, whereas ours loses 0.2 %, at the price of a 0.45 % loss in z_{500} CRPS at 12 h that vanishes by 120 h. Atlas also predicts surface pressure and column water vapour, so its mass can be constrained through a second, *native* operator; that formulation reverses the ranking (MPGD best, ours weakest), and we report it separately in Appendix I, where we also show that guidance on either operator does not transfer to the other.

4.3 Spectral fidelity

Deep-learning forecasts carry excess grid-scale power that has no meteorological counterpart. On the Atlas test dates the z_{500} power between wavenumbers 60 and 180 is 3.7× ERA5's at day 5 (median over the band), and 4.6× under BF16 inference (Fig. 3). Hydrostatic guidance brings it to 1.5× and 2.3× while leaving wavenumbers below 20 unchanged, at a small cost in the $k = 20$–60 band (0.97 → 0.90× at TF32). GenCast shows the same behaviour with a smaller excess: above $k = 60$ its power is 1.44× ERA5's at day 5 and 1.73× at day 15, which guidance reduces to 1.33× and 1.60× (Appendix I.1). The effect is specific to the constraint: dry-air-mass guidance, which acts on the surface pressure and moisture columns, leaves the GenCast z_{500} spectrum unchanged (1.73 → 1.72×).

4.4 Ablation: from MPGD to intermediate gradient guidance

Each component of Sec. 3.2 is added in turn to MPGD, holding everything else fixed (Table tables 3 and 4). The noisy operand is the ingredient that moves the long-lead numbers: on its own it lifts the hydrostatic integral from +10.6 to +12.5 % (GenCast), +20.9 to +26.0 % (Atlas), and turns

Table 3: Hydrostatic balance guidance on the test set. Entries are % change relative to the unguided baseline, $100\,(1 - \text{method}/\text{baseline})$, so higher is better in every column.

		D_{ERA5}				CRPS			
						12 h		120 h	
Model	Method	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
GenCast	MPGD	$+10.6$	-4.0	$+9.0$	$+14.6$	-1.97	-0.00	-0.58	-0.41
	MPGD + norm match	$+9.8$	-5.7	$+7.8$	$+14.1$	-1.47	-0.01	**-0.50**	**-0.29**
	MPGD + noisy	$+12.5$	$+1.0$	$+10.6$	$+16.6$	-1.41	-0.00	-0.82	-0.68
	Ours	**+13.6**	**+4.4**	**+12.0**	**+16.7**	**-1.09**	**+0.01**	**-0.50**	-0.37
Atlas	MPGD	$+20.9$	-7.0	$+18.5$	$+26.5$	-2.23	-0.01	-0.39	-0.07
	MPGD + norm match	$+19.7$	-2.9	$+17.2$	$+24.8$	-1.03	$+0.01$	**+0.05**	$+0.29$
	MPGD + noisy	$+26.0$	$+9.4$	$+24.3$	$+29.7$	-1.47	$+0.07$	-0.44	-0.07
	Ours	**+26.9**	**+21.4**	**+24.8**	**+30.3**	**+0.71**	**+0.34**	-0.30	**+1.20**

Table 4: Dry-air-mass guidance on the test set, ablation. Entries are % change relative to the unguided baseline, $100\,(1 - \text{method}/\text{baseline})$, so higher is better in every column. Best per model in bold.

		D_{ERA5}				CRPS			
						12 h		120 h	
Model	Method	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
GenCast	MPGD	$+60.7$	$+10.7$	$\mathbf{+64.9}$	$\mathbf{+64.1}$	$+0.27$	-0.45	$+0.28$	$+0.80$
	MPGD + norm match	$+57.3$	-12.0	$+59.3$	$+61.6$	$\mathbf{+0.61}$	-0.90	$\mathbf{+0.32}$	$+0.72$
	MPGD + noisy	$+60.8$	$+22.9$	$+63.2$	$+60.1$	$+0.17$	-0.12	$+0.25$	$\mathbf{+0.82}$
	Ours	$\mathbf{+62.5}$	$\mathbf{+33.4}$	$\mathbf{+64.9}$	$+63.1$	$+0.15$	$\mathbf{-0.04}$	$+0.27$	$+0.80$
Atlas	MPGD	$+54.1$	$+29.5$	$+66.7$	$+61.1$	$\mathbf{+0.04}$	-1.40	$+0.06$	$+0.02$
	MPGD + norm match	$+53.3$	$+46.6$	$+61.1$	$+54.8$	-0.05	-0.81	$+0.07$	$\mathbf{+0.11}$
	MPGD + noisy	$+61.6$	$+28.4$	$+71.5$	$\mathbf{+69.4}$	-0.10	-1.47	$\mathbf{+0.12}$	$+0.03$
	Ours	$\mathbf{+67.3}$	$\mathbf{+67.8}$	$\mathbf{+72.4}$	$+67.3$	-0.45	$\mathbf{-0.20}$	$+0.11$	-0.01

the 12 h loss into a gain. Norm matching on its own does not improve D_{ERA5}, but it halves the 12 h CRPS cost ($-1.97 \to -1.47\,\%$ GenCast, $-2.23 \to -1.03\,\%$ Atlas). Combined, the two are better than either: the 12 h D_{ERA5} gain rises from $+1.0$ to $+4.4\,\%$ (GenCast) and $+9.4$ to $+21.4\,\%$ (Atlas), and the z_{500} CRPS cost at 12 h falls to $-1.09\,\%$ and to a $+0.71\,\%$ gain. The dry-air-mass task shows the same pattern with larger short-lead effects: $+22.9 \to +33.4\,\%$ at 12 h, MSLP CRPS cost $-0.45 \to -0.04\,\%$.

4.5 Guidance under reduced precision model

Memory or throughput may force a forecaster to run a model below the precision it was validated at. We run Atlas on the March 2024 test dates in BF16 (weights, activations and autocast; sampler state and physical objective in FP32) instead of the TF32 of every other run, which cuts peak memory by $38\,\%$ and the step time by $16\,\%$ (Appendix 12), and apply the guidance scales selected at TF32 without change.

Reduced precision breaks consistency more than skill. Unguided BF16 raises the hydrostatic D_{ERA5} integral to $3.95\times$ its TF32 value ($4.4\times$ at day 5) and raises the z_{500} CRPS by $54\,\%$ at 12 h and $45\,\%$ at day 5 (Table 5).

Guidance repairs it. With the TF32 scales unchanged, ours removes $94\,\%$ of the excess D_{ERA5} (MPGD: $89\,\%$), leaving the guided BF16 forecast at $1.16\times$ the TF32 unguided value, and recovers $87\,\%$ of the z_{500} CRPS lost at 12 h (MPGD: $73\,\%$). Guidance adds about $4\,\%$ to the step time (Appendix 12), so a guided BF16 forecast costs a third less memory than TF32 and is closer to it in balance and in z_{500} skill than unguided BF16 is.

Table 5: Atlas at reduced precision, hydrostatic guidance, March 2024 test dates, 8 members, $0.25°$ grid. Entries are % change relative to unguided TF32, $100\,(1 - \text{run}/\text{TF32})$, higher is better; the guidance scales are those selected at TF32.

	D_{ERA5}				CRPS			
					12 h		120 h	
Mode	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
TF32	0	0	0	0	0	0	0	0
TF32 + ours	+26.9	+21.4	+24.8	+30.3	+0.71	+0.34	−0.30	+1.20
BF16	−295	−164	−342	−247	−54.5	−9.8	−45.1	−11.6
BF16 + MPGD	−32.0	−32.7	−33.5	−32.5	−14.5	−9.5	−12.6	−15.0
BF16 + ours	−16.3	−21.7	−13.2	−23.3	−6.9	−9.5	−14.6	−14.3

5 Limitations

First, we evaluate each guidance objective in isolation. We do not test whether guiding one physical constraint also improves unguided ones (e.g., whether hydrostatic-balance guidance reduces dry air mass drift, or vice versa), nor whether it degrades them. Second, we do not study multi-objective guidance. IGG extends naturally to several constraints by summing their losses, but this leaves open how to weight objectives under gradient norm matching and how to handle conflicting gradients. Third, IGG is sensitive to how a constraint is expressed in terms of model outputs. For dry air mass conservation, computing the constraint from mean sea-level pressure and layer-wise specific humidity yields substantially better results than computing it directly from surface pressure and total column water vapour. This holds even though the two formulations are physically equivalent, since total column water vapour is the vertical integral of specific humidity. We hypothesize that the derived formulation spreads the guidance gradient over the full three-dimensional humidity field, whereas the direct formulation concentrates it on two surface variables.

6 Conclusion

Probabilistic deep learning weather models now surpassed numerical models in forecast skill, but produce physical inconsistencies in their forecasting rollouts. In this work, we introduces Intermediate Gradient Guidance, a test-time guidance method that pushes each sampling step towards states that better satisfy a physical constraint, generating more physically realistic forecast. Across two SOTA models, IGG consistently help them to generate more physically consistent forecasts with near-zero extra computation costs, regardless model precision. Also, spectral fidelity of model is more consistent to ground truth under guidance. Future works including discover multi-objective guidance and guidance transferability.

Reproducibility statement

IGG modifies only the sampling loop of two publicly released checkpoints, so every result can be regenerated from public weights and public data. Algorithm 1 gives the complete update; Appendix D.1 states, for each backbone, which network evaluation is corrected, in which coordinates the gradient is taken, and how the step is rescaled. Appendix D records checkpoint identifiers, code commits, grids, variable lists, and every sampler setting we inherit (Tables 7 and 8), including the numerical precision of each run. The two constraint operators are defined in Appendix B, and the calibration scales that standardize them are derived in Appendix C from 31 ERA5 initializations of March 2022 (Table 6); these scales are fixed before any model is run and are shared by the guidance objective and the evaluation metric (Eq. 2). The development and test dates, the ensemble size and rollout length, and the selection rule for the guidance scale (the D_{ERA5} integral over 12–360 h on the development set, applied once to the held-out test set) are stated in Sec. 4.1; the selected scales for every method, task and backbone are listed in Appendix E, together with the implementation of each baseline and the random-seed convention that gives every method the same noise realization per (initialization, member) pair. ERA5 initial conditions and verification fields are read from the public ARCO-ERA5 archive. Appendix G reports the per-step cost of each method and the measured wall-

clock overhead of IGG on the hardware used. Our implementation, including the guidance code for both backbones, the residual operators, the evaluation scripts, and the configuration files of every reported run, is provided as anonymous supplementary material and will be released publicly on acceptance; the MMPS baseline uses the FA-APF GenCast port at commit `9abba5a`.

LLM DISCLOSURE STATEMENT

In this work, we used generative AI tools to implement methods and experiment code, and to assist in interpreting experimental results. We have not used generative AI tools to generate synthetic data, develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or give feedback on the research methodology or experiments, clean or reformat datasets, or translate text; writing proofs and qualitative/thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to improve the readability of the manuscript, to draw conceptual illustrations 2, and to transcribe equations from cited prior work into the notation used in this paper.

We have reviewed all AI-assisted work. LLM-generated code was inspected and tested by the authors, including reproducing reported baseline results. LLM-assisted interpretations of results were independently verified by the authors against the experimental outputs, and all conclusions drawn in the paper are the authors' own. Transcribed equations were checked against the original sources, which are cited where each equation appears. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Michael Aich, Andreas Fürst, Florian Sestak, Carlos Ruiz-Gonzalez, Niklas Boers, and Johannes Brandstetter. WIND: Weather inverse diffusion for zero-shot atmospheric modeling, 2026. URL https://arxiv.org/abs/2602.03924.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025. URL http://jmlr.org/papers/v26/23-1605.html.

Massimo Bonavita. On some limitations of current machine learning weather prediction models. *Geophysical Research Letters*, 51(12):e2023GL107377, 2024. doi: 10.1029/2023GL107377.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 2024.

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

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16993–17029. PMLR, 2025.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140:1917–1923, 2014. doi: 10.1002/qj.2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=o3BxOLoxm1.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022.

Emma Kasteleyn, Timo Maier, Axel Lauer, Veronika Eyring, Pierre Gentine, and Ana Lucic. PhysMetrics.Weather: An evaluation framework for physical consistency in ML weather models, 2026. URL https://arxiv.org/abs/2606.10642.

Dmitrii Kochkov, Janni Yuval, Ian Langmore, Peter Norgaard, Jamie Smith, Griffin Mooers, Milan Klöwer, James Lottes, Stephan Rasp, Peter Düben, Sam Hatfield, Peter Battaglia, Alvaro Sanchez-Gonzalez, Matthew Willson, Michael P. Brenner, and Stephan Hoyer. Neural general circulation models for weather and climate. *Nature*, 632(8027):1060–1066, 2024. doi: 10.1038/s41586-024-07744-y.

Jean Kossaifi, Nikola Kovachki, Morteza Mardani, Daniel Leibovici, Suman Ravuri, Ira Shokar, Edoardo Calvello, Mohammad Shoaib Abbas, Peter Harrington, Ashay Subramaniam, Noah Brenowitz, Boris Bonev, Wonmin Byeon, Karsten Kreis, Dale Durran, Arash Vahdat, Mike Pritchard, and Jan Kautz. Demystifying data-driven probabilistic medium-range weather forecasting, 2026. URL https://arxiv.org/abs/2601.18111.

Yingtao Luo, Shikai Fang, Binqing Wu, Qingsong Wen, and Liang Sun. Physics-guided learning of meteorological dynamics for weather downscaling and forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 2010–2020, New York, NY, USA, 2025. ACM. doi: 10.1145/3711896.3737081.

Ashwini Pokle, Matthew J. Muckley, Ricky T. Q. Chen, and Brian Karrer. Training-free linear image inverses via flows. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=PLIt3a4yTm.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, 2025. doi: 10.1038/s41586-024-08252-9.

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

Yingkai Sha, John S. Schreck, William Chapman, and David John Gagne, II. Improving AI weather prediction models using global mass and energy conservation schemes. *Journal of Advances in Modeling Earth Systems*, 17(11):e2025MS005138, 2025. doi: 10.1029/2025MS005138.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827. PMLR, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Akshay Subramaniam, Dale Durran, David Pruitt, Nathaniel Cresswell-Clay, and William Yik. Imposing the fundamental dynamical constraint of hydrostatic balance to improve global ML weather prediction, 2025. URL https://arxiv.org/abs/2506.08285.

U.S. Committee on Extension to the Standard Atmosphere. *U.S. Standard Atmosphere, 1976*. National Oceanic and Atmospheric Administration, National Aeronautics and Space Administration, and United States Air Force, Washington, D.C., 1976. NOAA-S/T 76-1562.

Wanghan Xu, Fenghua Ling, Wenlong Zhang, Tao Han, Hao Chen, Wanli Ouyang, and Lei Bai. Generalizing weather forecast to fine-grained temporal scales via physics-AI hybrid modeling. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0734.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0704.

A Background

This appendix fixes the sampler abstraction that IGG operates on (Appendix A.1) and reviews gradient-based test-time guidance, in particular MPGD, from which IGG is derived (Appendix A.2). The two atmospheric constraints we guide toward, hydrostatic balance and dry-air mass conservation, are defined in Appendix B.

Notation. Two time variables appear throughout. Physical time is the forecast lead t: X_t is the atmospheric state at lead t, X_0 is the initial condition, and Δt is the model time step (12 h for GenCast, 6 h for Atlas). Sampling time is the progress variable $s \in [0, 1]$ within a single forecast step, and z is the sampler state. Quantities that the diffusion literature indexes by a diffusion time are written here in terms of s, with the noisy end at $s = 0$ and the clean end at $s = 1$. When several forecasts are compared, a superscript k indexes the initialization.

A.1 Diffusion models and stochastic interpolants

Diffusion models. A diffusion model defines a forward process that gradually corrupts a clean sample $x \sim p_{\mathrm{data}}$ into noise, and generates by reversing it. In the variance-exploding parameterization of Karras et al. (2022), the forward process is the perturbation kernel $p(z \mid x) = \mathcal{N}(z; x, \sigma^2 I)$ indexed by a noise level σ, and generation follows the score $\nabla_z \log p_\sigma(z)$ as σ decreases from $\sigma_{\max}$ to 0. The score is parameterized through a denoiser $\hat{x}_\theta(z; \sigma) \approx \mathbb{E}[x \mid z]$ trained by denoising score matching; the two are related by Tweedie's formula (Efron, 2011),

$$\nabla_z \log p_\sigma(z) = \frac{\hat{x}_\theta(z; \sigma) - z}{\sigma^2}. \tag{3}$$

Sampling integrates the corresponding probability-flow ODE, or its SDE counterpart, from $\sigma_{\max}$ to 0,

$$\frac{\mathrm{d}z}{\mathrm{d}\sigma} = \frac{z - \hat{x}_\theta(z; \sigma)}{\sigma}, \tag{4}$$

so every step moves the state toward the current denoised estimate. GenCast (Price et al., 2025) instantiates this construction conditioned on the two most recent atmospheric states, and denoises the residual to the latest state rather than the state itself.

Stochastic interpolants. Stochastic interpolants (**?**Chen et al., 2024) bridge two arbitrary densities p_0 and p_1 in finite time. For $x_0 \sim p_0$, $x_1 \sim p_1$ and $\varepsilon \sim \mathcal{N}(0, I)$, the interpolant

$$z_s = \alpha(s)\, x_0 + \beta(s)\, x_1 + \gamma(s)\, \varepsilon, \qquad \alpha(0) = \beta(1) = 1, \quad \alpha(1) = \beta(0) = 0, \quad \gamma(0) = \gamma(1) = 0, \tag{5}$$

defines a path of densities from p_0 to p_1. Its transport is generated by the drift $b_\theta(z, s) \approx \mathbb{E}[\dot{z}_s \mid z_s = z]$, learned by regression and integrated as $\mathrm{d}z/\mathrm{d}s = b_\theta(z, s)$. When the noise term is generated by a Wiener process, $\gamma(s)\varepsilon \overset{d}{=} \sigma(s)W_s$ with $\gamma(s) = \sigma(s)\sqrt{s}$, and one integrates the SDE $\mathrm{d}z =$

$b_\theta(z, s)\,\mathrm{d}s + \sigma(s)\,\mathrm{d}W_s$ instead (**?**). Unlike in diffusion, p_0 need not be Gaussian. For forecasting it is natural to take p_0 to be the distribution of the current atmospheric state (**?**). Atlas (Kossaifi et al., 2026) makes this choice, with the next-step residual at the $s = 1$ end, so the sampling trajectory starts from a physical weather state rather than from noise.

A shared sampler abstraction. Although the two formulations differ in their forward construction, they generate in the same way: a state z is iterated over $s \in [0, 1]$, from the tractable reference at $s = 0$ to the clean end at $s = 1$, along a network-produced update direction $b_\theta(z, s; c)$ conditioned on $c = (X_{t-\Delta t}, X_t)$,

$$z \;\leftarrow\; z + \Delta_s\, b_\theta(z, s; c) \;+\; \text{(optional noise)}. \tag{6}$$

For GenCast, $b_\theta = \hat{x}_\theta(z; \sigma(s)) - z$ with $\sigma(s)$ decreasing in s, so that an Euler step of Eq. 4 from σ to σ' has $\Delta_s = 1 - \sigma'/\sigma$. For Atlas, b_θ is the interpolant drift. Equation 6 is the only interface IGG requires: Section 3.2 modifies b_θ and changes nothing else in either sampler.

A.2 Gradient-based test-time guidance

Test-time guidance samples from a tilted distribution $p(x \mid y) \propto p(x) \exp(-\mathcal{L}(x; y))$ for a differentiable objective $\mathcal{L}$ that expresses the desired behavior. Because the score of the tilted path decomposes as

$$\nabla_z \log p_s(z \mid y) = \nabla_z \log p_s(z) + \nabla_z \log p_s(y \mid z), \tag{7}$$

guidance adds a gradient term to the update of Eq. 6 and leaves the pretrained network untouched (Song et al., 2021).

The practical difficulty is that $\mathcal{L}$ is defined on clean data, whereas the sampler only ever holds a noisy intermediate state z. DPS (Chung et al., 2023) resolves this by evaluating $\mathcal{L}$ at the Tweedie estimate(Efron, 2011) and differentiating through it,

$$z \;\leftarrow\; z - \zeta_s\, \nabla_z \mathcal{L}\big(\hat{x}_\theta(z; \sigma); y\big), \tag{8}$$

which requires a backward pass through the network at every sampling step. MMPS (Rozet et al., 2024) sharpens the likelihood term by also matching the second moment of $p(x \mid z)$, at the cost of additional vector–Jacobian products through the denoiser. It obtains that second moment from Tweedie's covariance formula, which requires a denoiser $\mathbb{E}[x \mid z]$ and its Jacobian; the interpolant sampler of Atlas exposes only a drift, so MMPS does not apply to it directly.

Manifold preserving guided diffusion. MPGD (He et al., 2024) starts from the observation that the update of Eq. 8 searches a neighborhood in the ambient space $\mathbb{R}^d$, whereas the data concentrate on a low-dimensional manifold $\mathcal{M}$ and, at noise level σ, noisy states concentrate on an associated shell $\mathcal{M}_\sigma$. An ambient-space gradient step can therefore push the state off $\mathcal{M}_\sigma$, into a region where the network was never trained, which is why DPS-style methods need carefully tuned step-size schedules or many sampling steps to remain stable. MPGD instead restricts the step to the tangent space of $\mathcal{M}$, and shows that it suffices to differentiate with respect to the clean estimate rather than with respect to z,

$$\hat{x} \;\leftarrow\; \hat{x} - c_s\, \nabla_{\hat{x}} \mathcal{L}(\hat{x}; y), \qquad \hat{x} = \hat{x}_\theta(z; \sigma), \tag{9}$$

after which sampling continues with the modified $\hat{x}$ in place of $\hat{x}_\theta$. The two updates are related by the chain rule,

$$\nabla_z \mathcal{L}\big(\hat{x}_\theta(z; \sigma); y\big) = \Big(\frac{\partial \hat{x}_\theta}{\partial z}\Big)^\top \nabla_{\hat{x}} \mathcal{L}(\hat{x}; y), \tag{10}$$

so Eq. 9 is Eq. 8 with the denoiser Jacobian dropped. By the second-order Tweedie identity this Jacobian equals $\mathrm{Cov}[x \mid z]/\sigma^2$. Under the linear manifold hypothesis it acts as a scaled projection onto the tangent space, so dropping it costs only a scale factor, provided $\nabla_{\hat{x}} \mathcal{L}$ already lies in that tangent space. He et al. (2024) bound the resulting discrepancy: with matched step sizes, $\|z^{\mathrm{DPS}} - z^{\mathrm{MPGD}}\| \le \kappa\, \zeta_s\, \sigma$, where κ bounds the product of the loss gradient and the network Jacobian. Two consequences matter here. The guidance costs a backward pass through $\mathcal{L}$ only, never through the network; and the approximation is tightest at the clean end of sampling and loosest at the noisy end, since the bound scales with σ. MPGD's manifold-projection variants, MPGD-AE and MPGD-Z, enforce the tangent-space condition with a pretrained autoencoder. No such autoencoder exists for global atmospheric states, so we use MPGD without projection, and write MPGD for that variant throughout.

B PHYSICAL CONSTRAINTS

Both constraints are evaluated in SI units on a forecast state X_t and compared with the initial condition X_0. Throughout, g is the gravitational acceleration and R_d the gas constant of dry air.

B.1 HYDROSTATIC BALANCE

Hydrostatic balance is the approximate balance between gravity and the vertical pressure-gradient force. With geopotential Φ ($\mathrm{m^2\,s^{-2}}$), virtual temperature T_v and specific humidity q,

$$\frac{\partial \Phi}{\partial \ln p} = -R_d\, T_v, \qquad T_v \approx T\,(1 + 0.608\, q). \tag{11}$$

On pressure levels $p_1 < \cdots < p_L$ we define the discrete hydrostatic residual at every grid point,

$$\mathcal{R}_\ell(X) = \frac{\Phi_{\ell+1} - \Phi_\ell}{\ln(p_{\ell+1}/p_\ell)} + R_d\, \overline{T}_{v,\ell}, \qquad \overline{T}_{v,\ell} = \tfrac{1}{2}\big(T_{v,\ell} + T_{v,\ell+1}\big), \qquad \ell = 1, \ldots, L-1, \tag{12}$$

using the $L = 11$ levels from 50 to 850 hPa, that is, the ten adjacent pairs of Table 6.

Because the layer-mean virtual temperature is approximated by the mean of its two endpoint values, $\mathcal{R}_\ell$ does not vanish even for a balanced atmosphere. We therefore guide toward the residual of the initial condition rather than toward zero,

$$r_\ell(X_t; X_0) = \mathcal{R}_\ell(X_t) - \mathcal{R}_\ell(X_0), \tag{13}$$

and reduce its magnitude during sampling. The reference X_0 is held fixed over the whole forecast. It absorbs the discretization error already present in the initial condition, but it does not make the discrete balance exact, and ERA5 itself drifts away from it with lead time (Appendix C).

B.2 DRY-AIR MASS CONSERVATION

The global dry-air mass is the total atmospheric mass minus its water vapor,

$$M_d = \int_\Omega \Big[\frac{p_s}{g} - W_v\Big]\, \mathrm{d}A, \qquad W_v = \frac{1}{g}\int_0^{p_s} q\, \mathrm{d}p, \tag{14}$$

where Ω is the sphere, p_s the surface pressure and W_v the column water vapor. We evaluate the area integral as a sum over grid cells weighted by their exact areas, and define the guidance target

$$r_M(X_t; X_0) = M_d(X_t) - M_d(X_0), \tag{15}$$

forming the difference cell by cell before the area sum.

The two backbones supply p_s and W_v differently, and we call the resulting maps $X \mapsto M_d(X)$ operators. The *native* operator (Atlas) reads the predicted surface-pressure and total-column-water-vapor channels directly. The *derived* operator (GenCast, which predicts neither quantity) reconstructs surface pressure from mean sea-level pressure and surface geopotential Φ_s with the U.S. Standard Atmosphere 1976 (U.S. Committee on Extension to the Standard Atmosphere, 1976),

$$\hat{p}_s = p_{\mathrm{msl}}\Big(1 - \frac{\Gamma H}{T_0}\Big)^{g/(R_d\Gamma)}, \qquad H = \frac{\Phi_s}{g}, \quad T_0 = 288.15\,\mathrm{K}, \quad \Gamma = 0.0065\,\mathrm{K\,m^{-1}}, \tag{16}$$

and integrates q over the 13 pressure levels up to $\hat{p}_s$ with the trapezoidal rule. The reconstruction makes the derived mass noisier than the native one, which is why the two operators receive separate scales (Appendix C).

C CALIBRATION OF THE GUIDANCE SCALES

Both guidance targets are standardized by fixed scales measured on ERA5 before any model is run: one scale σ_ℓ per pressure pair for the hydrostatic target and a single scale σ_M for global dry-air mass. The same scales standardize the guidance objective (Eq. 19) and the evaluation distances (Eq. 2), so a distance of one means the same thing in both.

Data. We use the 31 daily 00 UTC initializations of 1–31 March 2022, which are disjoint from the tuning month (March 2023) and from every test date. For each initialization k we read ERA5 at the initial time and at the 30 leads $t \in \mathcal{T} = \{12, 24, \ldots, 360\}$ h from the ARCO-ERA5 archive at $0.25°$, and keep every fourth grid point. This yields exactly the $1°$ (181×360) grid on which GenCast runs and on which Atlas applies guidance, and gives $31 \times 30 = 930$ initialization–lead pairs. We write E_t^k for the ERA5 state of initialization k at lead t (denoted $X_t^{k,\mathrm{ERA5}}$ in Section 3.2).

Estimator. Guidance pulls a forecast toward its own initial condition, so the natural tolerance is how far the real atmosphere moves from its initial condition in the same quantity. We therefore calibrate on ERA5's own increment relative to the initial state, and use its zero-mean root mean square rather than its standard deviation: the target is *no* change, so a systematic drift of ERA5 over the rollout counts toward the tolerance.

Hydrostatic scale σ_ℓ. With r_ℓ the target of Eq. 13 evaluated for the ten pairs, and ω_i the normalized cosine-latitude weights of Eq. 19,

$$\sigma_\ell^2 = \frac{1}{K\,|\mathcal{T}|\,|I|} \sum_{k=1}^{K} \sum_{t\in\mathcal{T}} \sum_{i\in I} \omega_i \left[r_\ell(E_t^k; E_0^k)_i \right]^2, \qquad K = 31, \quad |\mathcal{T}| = 30, \tag{17}$$

pooled over all leads. Table 6 lists the result. The increment grows with lead but saturates early: in every pair the pooled scale lies 0.6–4.7% below the 360 h value, whereas the 12 h value lies 16–34% below it. Across pairs the scales differ by a factor of 3.5, which is why a single shared scale would let the stratospheric and lower-tropospheric pairs dominate the objective. The weighted mean of each increment is below 2.1% of its RMS, so here the zero-mean RMS and the standard deviation coincide. GenCast and Atlas use the same σ_ℓ.

Dry-air-mass scale σ_M. With r_M the target of Eq. 15,

$$\sigma_M^2 = \frac{1}{K} \sum_{k=1}^{K} r_M\big(E_{t^\star}^k; E_0^k\big)^2, \qquad t^\star = 360\,\mathrm{h}, \tag{18}$$

so a forecast whose mass has drifted by $1\,\sigma_M$ has moved as far as ERA5 typically does over 15 days. σ_M depends on the operator that produces p_s and W_v (Appendix B.2), so we calibrate one per operator. The derived operator gives a scale about four times larger than the native one ($\sigma_M = 4.89 \times 10^{14}$ vs. 1.16×10^{14} kg, or 0.0096% vs. 0.0023% of M_d). Most of the derived value is a systematic decrease over the 15-day window (mean increment -4.21×10^{14} kg at 360 h), which the zero-mean RMS keeps and a standard deviation (2.53×10^{14} kg) would discard.

Pooled versus fixed-lead scales. The two targets use different lead conventions; neither choice changes which configuration is selected. σ_M is a single scalar: changing it reparameterizes the guidance strength and the unit of the mass metric, and leaves the norm-matched arms, which are invariant to the scale of the objective, unchanged. The pooled alternative would differ from the 360 h value by $+0.7\%$ for the derived operator and by -25% for the native one. For σ_ℓ only the relative weights of the ten pairs matter, and pooling changes them by less than 5% relative to the 360 h scales while using all 930 increments per pair instead of 31. As a consequence, a distance of one in Eq. 2 corresponds to ERA5's own residual drift averaged over leads of 12–360 h.

Reference level for D_{IC}. Write $D_{\mathrm{IC}}(t)$ for the distance of Eq. 2 with the initial state X_0^k in place of $X_t^{k,\mathrm{ERA5}}$, that is, the pooled RMS of the standardized guidance target $r(X_t; X_0)$ itself. On the calibration dates, ERA5's own D_{IC} rises from 0.78 at 12 h to 0.87 at 24 h, 0.98 at 72 h and 1.00 at 120 h, and its mean square over leads is one by construction. A guided forecast whose D_{IC} falls far below this curve is over-constrained rather than accurate, so we never use D_{IC} for selection or ranking and quote it only against this reference.

Transfer checks. All checks below use evaluation dates and served only as diagnostics, never to adjust a scale. *Year:* recomputing σ_ℓ on March 2023 gives 0.93–1.08 times the 2022 values across pairs, and the derived σ_M at 360 h changes by 0.6%. *Season:* over the 24 test initializations (June, September and December 2023, and March 2024) the per-pair ratio is 0.93–1.05; within a single test

month it ranges from 0.84 (December 2023) to 1.14 (June 2023). *Resolution:* Atlas is evaluated on its native 0.25° grid with the 1° scales; ERA5's own D_{IC} at 72 h is 0.969 on the full grid and 0.968 on the 1° subsample.

Standardized residuals. The guidance objective and all baselines operate on residuals in standardized coordinates,

$$\mathcal{S}(d)_{\ell i} = \sqrt{\omega_i}\,\frac{d_{\ell i}}{\sigma_\ell}, \qquad \omega_i = \frac{\cos\varphi_i}{\frac{1}{|I|}\sum_{j\in I}\cos\varphi_j}, \tag{19}$$

where φ_i is the latitude of grid point $i \in I$. For the scalar mass target, $\mathcal{S}(d) = d/\sigma_M$.

Table 6: Guidance scales calibrated on ERA5 (31 initializations, March 2022). Hydrostatic scales in $\mathrm{m^2\,s^{-2}}$ per pressure pair; the pooled column is used. Dry-air-mass scales in kg; the 360 h column is used.

	12 h	360 h	pooled 12–360 h
Hydrostatic σ_ℓ (GenCast and Atlas)			
50–100 hPa	316.5	441.6	**421.0**
100–150 hPa	212.2	280.0	**272.7**
150–200 hPa	178.4	234.4	**231.8**
200–250 hPa	156.0	198.4	**196.8**
250–300 hPa	102.3	122.1	**121.4**
300–400 hPa	148.1	188.7	**185.4**
400–500 hPa	110.3	131.6	**130.3**
500–600 hPa	99.6	121.7	**119.9**
600–700 hPa	101.8	146.5	**139.8**
700–850 hPa	202.4	306.3	**298.6**
Dry-air mass σ_M			
derived operator (GenCast)	3.23×10^{14}	$\mathbf{4.89 \times 10^{14}}$	4.92×10^{14}
native operator (Atlas)	5.63×10^{13}	$\mathbf{1.16 \times 10^{14}}$	8.71×10^{13}

D Backbone checkpoints and inference settings

Both backbones are used exactly as released; IGG changes only the sampling loop. We therefore do not restate their training recipes (optimizer, schedule, compute), which are documented in the original papers (Price et al., 2025; Kossaifi et al., 2026). Table 7 records provenance, architecture and predicted variables, and Table 8 gives the sampler settings, which we keep at their released defaults unless stated otherwise.

D.1 Where guidance acts

On both backbones the constraint is evaluated in physical units, while the correction is applied in the sampler's own standardized coordinates. At each guided network evaluation, the current noisy sampler state z is mapped to a physical forecast state $\mathcal{P}(z; X_t)$ with the model's own normalization statistics: for GenCast, the last input state plus the noisy residual rescaled by the per-level residual standard deviation; for Atlas, the current coarse state plus the de-standardized residual. The residual (hydrostatic balance for the ten pairs from 50 to 850 hPa, or global dry-air mass) is computed in SI units, differenced against its value at the initial condition, and standardized with Eq. 19, so the objective is dimensionless. Its gradient is taken with respect to the standardized sampler state. Norm matching (Section 3.2) then sets each member's step length relative to that member's remaining denoising displacement $\|\hat{x} - z\|$ in the same space, so the physical units of the residual never enter the step size. The correction is added to the denoised estimate (GenCast) or to the drift (Atlas), and scaled so that the state leaving the corrected stage moves by the nominal amount.

For Atlas, the guided state is the 1° (181×360) latent state that the stochastic interpolant evolves and that Atlas carries from one forecast step to the next, not the decoded 0.25° field. The decoder runs once, after the final sampler step, and is never differentiated. Guidance therefore constrains the

Table 7: Frozen backbones: provenance, architecture and predicted variables.

	GenCast (1°)	Atlas-SI
Checkpoint	`GenCast 1p0deg <2019.npz`	`nvidia/atlas-era5` (Hugging Face), revision `fdce048`: `genmodel/` + `autoencoders/autoencoder_0/`
Code / commit	`google-deepmind/graphcast` @ `f2f2c51` (package `weathernext`)	`NVIDIA/earth2studio` v0.17.0 (`d027d61`)
Grid ($H \times W$)	181×360	721×1440
Time step Δt	12 h	6 h
Conditioning	two most recent states ($t - 12$ h, t) and forcings	two most recent states ($t - 6$ h, t)
Pressure levels	13 (both): 50, 100, 150, 200, 250, 300, 400, 500, 600, 700, 850, 925, 1000 hPa	
Surface variables	2 m temperature, 10 m u/v wind, mean sea-level pressure, sea-surface temperature, total precipitation (12 h)	10 m and 100 m u/v wind, 2 m temperature, surface pressure, mean sea-level pressure, total column water vapor, sea-surface temperature, total precipitation (6 h)
Pressure-level variables	geopotential, temperature, u, v, vertical velocity, specific humidity	geopotential, temperature, u, v, specific humidity
Channels C	84 (6 surface + 6×13 pressure-level)	75 (10 surface + 5×13 pressure-level)
Training data	ERA5, 1979–2018	ERA5, 1980–2019 (train 1980–2016, validation 2017–2019)
Generative model	EDM diffusion	stochastic interpolant (linear α, β, σ)
Network	GNN encoder/decoder + sparse graph-transformer processor (16 layers, width 512) on a 5×-refined icosahedral mesh	global-attention DiT (2.4B) on a 1° (181×360) bilinearly coarsened state; local (neighborhood)-attention DiT decoder (1.8B) to $0.25°$
Prediction target	residual $X_{t+\Delta t} - X_t$ scaled by the per-level std of 12 h differences (precipitation, not an input, predicted as a standardized value)	coarsened residual $\mathcal{B}(X_{t+\Delta t} - X_t)$, standardized; $\mathcal{B}$ = bilinear resampling to 181×360
Normalization	per-variable, per-level standardization (state and residual separately)	per-variable standardization (state and residual separately)
Parameters	57.5M	2.4B (latent DiT) + 1.8B (decoder DiT)

coarse state, and the $0.25°$ product inherits the constraint only through the decoder; all metrics are nevertheless evaluated on the $0.25°$ output.

Guidance is active at every solver step with a constant scale: all $N{=}20$ steps for GenCast and all $N{=}100$ steps for Atlas. Within each step it corrects exactly one of the two network evaluations, the one that produces the step's output: the second DPM-Solver++2S stage for GenCast (the first stage on the final Euler step, which has no second stage), and the second drift evaluation of the Roberts step for Atlas. This amounts to 20 of GenCast's 39 denoiser evaluations and 100 of Atlas's 200 generator evaluations per forecast step. Because the objective is evaluated on the detached state, no gradient passes through either network.

E Baseline implementations

E.1 Common formulation

All guidance baselines solve the same inverse problem. Let $\mathcal{R}$ be a residual operator of Appendix B (the stacked $\mathcal{R}_\ell$, or M_d), $y = \mathcal{R}(X_0)$ the frozen observation taken from the initial condition, and

Table 8: Sampler settings inherited from the released checkpoints. IGG is inserted into these loops without changing any value below; the only departure from the released configuration is GenCast's attention kernel, replaced by its GPU equivalent.

GenCast (EDM)	
Solver	DPM-Solver++2S, N=20 noise levels (39 denoiser evaluations per 12 h step)
Noise schedule	$\sigma_{\min}$=0.03, $\sigma_{\max}$=80, ρ=7
Stochastic churn	S_{churn}=2.5, $S_{t_{\min}}$=0.75, $S_{t_{\max}}$=∞ (equivalent to the value 80 listed by Price et al. (2025), since $\sigma \le \sigma_{\max}$), S_{noise}=1.05
Noise distribution	spherical-harmonic white noise (isotropic Gaussian on the sphere)
Attention kernel	block-diagonal attention with a full mask (`triblockdiag_mha`), the GPU replacement for the released TPU splash attention
Precision	FP32 weights and activations; FP32 matrix products run as TF32 (JAX/XLA GPU default)
Atlas-SI (stochastic interpolant)	
Interpolant	$\alpha(s)$=$1-s$, $\beta(s)$=s, $\sigma(s)$=$g(s)$=$5(1-s)$ (so $\gamma(s)$=$\sigma(s)\sqrt{s}$ in Eq. 5), from the current coarse state (s=0) to the coarse next-step residual (s=1)
Solver	stochastic Runge–Kutta (Roberts), 100 uniform steps on $s \in [0,1]$ (200 generator evaluations per 6 h step)
Noise distribution	spherical-harmonic white noise
Decoding	coarsened residual decoded once after the final step, conditioned on X_t at 0.25°
Precision	TF32 mode of Table 12 (FP32 weights and state, TF32 matrix products) for all reported runs; BF16 only in the precision study (Section 4.5)

$\mathcal{A} = \mathcal{R} \circ \mathcal{P}(\,\cdot\,; X_t)$ the measurement operator that maps a sampler state to a residual field, with $\mathcal{P}$ the map to physical units of Appendix D.1. We work throughout in the standardized coordinates of Eq. 19, so that a diagonal observation covariance $\Sigma_y = \tau^2 I$ in these coordinates corresponds to per-pair scaling by σ_ℓ and cosine-latitude weighting on the native grid. The methods differ only in how they approximate the likelihood $p(y \mid z)$ and in how they scale the resulting correction. Unless stated otherwise, every baseline uses the same sampler, the same number of sampling steps, and the same random seed per (initialization, member) pair as unguided sampling. DPS and MMPS are evaluated on GenCast only: MMPS requires a denoiser that Atlas does not expose (Appendix A.2).

E.2 Diffusion posterior sampling (DPS)

DPS (Chung et al., 2023) approximates the intractable likelihood by collapsing $p(x \mid z)$ to a point mass at its mean,

$$p(y \mid z) = \int p(y \mid x)\, p(x \mid z)\, \mathrm{d}x \;\approx\; \mathcal{N}\big(y \mid \mathcal{A}(\hat{x}_\theta(z)),\, \Sigma_y\big), \tag{20}$$

where $\hat{x}_\theta(z) = \mathbb{E}[x \mid z]$ is the denoiser output. The sampler update becomes

$$z \;\leftarrow\; z + \Delta_s b_\theta(z, s; c) - \zeta_s \nabla_z \big\| \mathcal{S}\big(\mathcal{A}(\hat{x}_\theta(z)) - y\big) \big\|_2, \tag{21}$$

which requires one backward pass through the network per sampling step. Following the released implementation, we differentiate the norm rather than the squared norm. Since $\nabla\|r\| = \nabla\|r\|^2/(2\|r\|)$, this is equivalent, up to a factor of two absorbed into ζ', to the step size $\zeta_s = \zeta'/\|\mathcal{S}(\mathcal{A}(\hat{x}_\theta(z)) - y)\|_2$ of their Algorithm 1, and it makes the correction magnitude independent of the current size of the residual. This normalization differs from the norm matching of Section 3.2: DPS normalizes by the *measurement residual*, which carries no information about the sampler, whereas IGG normalizes by the sampler's own update, which is what makes λ_0 dimensionless and transferable across backbones and numerical precisions (Section 4.5).

Eq. 20 is accurate only when σ is small, since $p(x \mid z)$ concentrates around its mean only at the clean end of sampling. This is the same σ-dependence that bounds the MPGD shortcut (Appendix A.2), and it is consistent with the late onset of the guidance effect we observe for both methods.

E.3 MOMENT MATCHING POSTERIOR SAMPLING (MMPS)

MMPS (Rozet et al., 2024) replaces the point estimate of Eq. 20 with the Gaussian approximation of $p(x \mid z)$ that is closest in KL divergence, $q(x \mid z) = \mathcal{N}\big(x \mid \mathbb{E}[x \mid z],\, \mathbb{V}[x \mid z]\big)$. For a linear Gaussian forward process this marginalizes in closed form,

$$q(y \mid z) = \mathcal{N}\big(y \mid A\,\mathbb{E}[x \mid z],\ \Sigma_y + A\,\mathbb{V}[x \mid z]\,A^\top\big), \tag{22}$$

giving the likelihood score

$$\nabla_z \log q(y \mid z) = \nabla_z \mathbb{E}[x \mid z]^\top A^\top \big(\Sigma_y + A\,\mathbb{V}[x \mid z]\,A^\top\big)^{-1}\big(y - A\,\mathbb{E}[x \mid z]\big), \tag{23}$$

with the posterior covariance taken from Tweedie's second-order formula, $\mathbb{V}[x \mid z] = \Sigma_\sigma \nabla_z^\top \hat{x}_\theta(z)$. Neither $\mathbb{V}[x \mid z]$ nor the inverse in Eq. 23 is materialized; the linear system is solved iteratively, each iteration requiring one vector–Jacobian product through the denoiser.

Linearization. Eq. 22 assumes a linear measurement matrix A, which our two constraints satisfy to different degrees. Dry-air mass is effectively linear in the guided variables. The surface-pressure proxy of Eq. 16 is linear in p_{msl} because the orographic height $H = \Phi_s/g$ is a static field, so the exponential factor is a fixed spatial multiplier, and the water-vapor term is linear in q up to the dependence of the integration limit on p_s. Hydrostatic balance is not linear: the virtual temperature $T_v \approx T(1 + 0.608\,q)$ makes $\mathcal{R}_\ell$ bilinear in (T, q). We therefore linearize $\mathcal{A}$ about the current denoised estimate, holding q fixed at its value in $\hat{x}_\theta(z)$, which renders $\mathcal{R}_\ell$ linear in (Φ, T) and supplies the matrix A used in Eq. 23. We re-linearize at every sampling step.

This configuration, with $\mathcal{A}$ linearized about the denoised estimate and the linear system solved iteratively, is the guidance procedure WIND (Aich et al., 2026) applies to its frozen pretrained backbone to conserve dry-air mass, and we label it as such in Table 4.

E.4 MANIFOLD PRESERVING GUIDED DIFFUSION (MPGD)

MPGD (He et al., 2024) removes the network gradient by differentiating the objective with respect to the clean estimate rather than the state,

$$\hat{x} \ \leftarrow\ \hat{x} - c_s \nabla_{\hat{x}} \big\| \mathcal{S}\big(\mathcal{R}(\hat{x}) - y\big) \big\|_2^2, \tag{24}$$

after which sampling continues with the modified $\hat{x}$. We use MPGD without projection, since its MPGD-AE and MPGD-Z variants require an autoencoder pretrained on the data manifold and no such model exists for global atmospheric states. Under GenCast the clean estimate is the denoiser output. Under Atlas's interpolant sampler there is no denoiser, so we take the one-step extrapolation along the drift, $\hat{z} = z + (1 - s)\, b_\theta(z, s; c)$, as the clean estimate; this estimate is crude at small s, which is one of the two limitations IGG removes. The step size c_s is constant in ambient data units.

E.5 PROJECTION

As a limiting case, we project the completed unguided forecast onto the constraint after each autoregressive step, using the minimum-norm Gauss–Newton correction

$$X \ \leftarrow\ X - J^\top \big(J J^\top\big)^{-1}\big(\mathcal{R}(X) - y\big), \qquad J = \left.\frac{\partial \mathcal{R}}{\partial X}\right|_X, \tag{25}$$

iterated **[n]** times. We apply it at every autoregressive step rather than only at the final lead, so that the corrected state conditions the next step, matching the conditions under which the guided methods operate. Projection attains $D_{\mathrm{IC}} \approx 0$ by construction. For the hydrostatic target it is therefore the degenerate case that motivates the D_{ERA5} integral as the primary metric: it satisfies the stated objective exactly, while the forecast's residual field no longer tracks the evolving atmosphere. For dry-air mass the constraint is a single scalar per forecast, so the correction moves the state along one direction only and this degeneracy does not arise; there, projection is a strong baseline (Table 2).

E.6 BEST-OF-k SELECTION

Our premise is that the learned distribution already contains states that satisfy the constraint better than the ones sampling happens to reach. Best-of-k tests that premise directly: we score each of the $k = M = 8$ unguided members of an initialization by the guidance objective evaluated on its final state (360 h), and retain the single best member. This requires one forward evaluation of $\mathcal{R}$ per member and no additional sampling. If it recovered most of the gain, guidance would be performing selection that could be obtained for free; if it does not, guidance reaches states the sampler does not reach unaided. Because the procedure collapses the ensemble to one member, no probabilistic score is defined for it, and we report it on the consistency metrics only.

F SELECTED GUIDANCE SCALES

Table 10 lists, for every method, task and backbone, the guidance scale used for all test-set results, together with the ladder it was chosen from. Every ladder was run on the development set of Sec. 4.1 (six March 2023 initializations, 8 members, 15-day rollouts), and the pick is the plain argmax of the D_{ERA5} integral over 12–360 h (Eq. 2); CRPS was recorded but never used for selection. Each method is therefore evaluated at its own best dose on the primary metric, as in the validation-set practice of He et al. (2024) and Chung et al. (2023). The scales of the constant-step methods (DPS, MPGD, MPGD + noisy) are in the standardized coordinates of Eq. 19 and are not comparable across methods; the norm-matched scales λ_0 are dimensionless fractions of the sampler's own update. Projection and best-of-k have no scale ($k = 8$). All runs use base random seed 1126, from which the per-(initialization, member) noise is derived identically for every method.

On Atlas the dry-air-mass constraint can be evaluated with two operators (Appendix B.2): the derived operator shared with GenCast, which reconstructs surface pressure from mean sea-level pressure and integrates specific humidity, and the native operator, which reads Atlas's own surface pressure and total column water vapour. Each operator was tuned separately, with its own development ladders and its own D_{ERA5} (computed with the same operator), and the table lists both sets of picks; the main text reports the derived operator, and the native picks belong to the native results of Appendix I. Every pick is interior to its ladder with three exceptions: on the Atlas derived-operator mass task, MPGD + noisy (16) and MPGD + norm match (5×10^{-4}) sit at the top and bottom edge of their ladders respectively, and on GenCast the MMPS pick ($\tau = 0.05$) is the documented stability edge of that method (smaller τ diverges). We report these as they are rather than extending the ladders after seeing test results.

G COMPUTATIONAL COST

Table 11 counts the per-step cost of each method on GenCast, and Table 12 measures the overhead of IGG on Atlas across numerical precisions. For reference, WIND reports that MMPS dry-air-mass guidance with one to three CG iterations makes sampling 5.7–10.5× slower than unguided sampling on its own backbone (Aich et al., 2026).

H ADDITIONAL RESULTS: FORECAST SKILL ON OTHER VARIABLES

The main text reports fair CRPS for z_{500} and MSLP only. Tables 13 and 14 extend the same test-set comparison to three variables that neither constraint acts on directly: temperature at 850 hPa (t_{850}), specific humidity at 700 hPa (q_{700}) and 10 m zonal wind (u_{10}), at 12 and 120 h. Entries are the % change in fair CRPS relative to the unguided baseline, computed exactly as in the main tables: mean over initializations of the 8-member fair CRPS, cosine-latitude weighted on the native grid, with a paired date bootstrap over initializations for the 95 % interval (24 dates for GenCast, six for Atlas). Both backbones are scored with the same code.

Hydrostatic guidance costs a small amount of skill in every variable, most of it at 120 h and most of it in q_{700}, the humidity level closest to the guided pressure pairs; the loss is below $0.6\,\%$ on GenCast and below $2.3\,\%$ on Atlas for every gradient method. The pattern follows the ablation of Sec. 4.4: the noisy operand alone costs the most, norm matching roughly halves the loss, and the combination sits at the norm-matched level while delivering the larger D_{ERA5} gain. Projection, which enforces

Table 9: Guidance scales used on the test set and the development-set ladders they were selected from (bold: selected). ζ': DPS step (Eq. 21); c: constant MPGD step in standardized units (Eq. 24); λ_0: norm-matched fraction (Algorithm 1); τ: MMPS observation noise with $n = 2$ solver iterations. Dagger: pick at a ladder edge.

Backbone / task	Method	Scale	Development ladder
GenCast, hydrostatic	DPS (ζ')	750	300, 450, 600, **750**, 1200
	MPGD (c)	150	40, 75, 100, **150**, 200, 300, 600
	MPGD + norm match (λ_0)	0.0125	0.006, 0.009, **0.0125**, 0.018, 0.025, 0.05, 0.1
	MPGD + noisy (c)	375	100, 150, 200, 250, 300, **375**, 450 (600 diverged)
	Ours (λ_0)	0.07	0.025, 0.035, 0.05, **0.07**, 0.1
GenCast, dry-air mass	DPS (ζ')	22	4, 8, 11, 15, **22**, 30, 60
	MMPS (τ, $n = 2$)	0.05†	0.2, 0.1, 0.07, 0.06, **0.05**
	MPGD (c)	8	2, 4, 6, **8**, 11, 15, 30
	MPGD + norm match (λ_0)	0.0025	0.0006, 0.00125, **0.0025**, 0.005, 0.01, 0.025
	MPGD + noisy (c)	8	2, 4, 6, **8**, 11, 15, 30
	Ours (λ_0)	0.025	0.0025, 0.005, 0.01, **0.025**, 0.05
Atlas, hydrostatic	MPGD (c)	150	50, 100, **150**, 200, 400, 800
	MPGD + norm match (λ_0)	0.001	0.00025, 0.0005, **0.001**, 0.002, 0.004
	MPGD + noisy (c)	300	50, 100, 200, **300**, 400
	Ours (λ_0)	0.01	0.0025, 0.005, **0.01**, 0.02
Atlas, dry-air mass (derived operator)	MPGD (c)	16	4, 8, **16**, 32
	MPGD + norm match (λ_0)	0.0005†	**0.0005**, 0.001, 0.002, 0.004
	MPGD + noisy (c)	16†	2, 4, 8, **16**
	Ours (λ_0)	0.0006	0.0003, **0.0006**, 0.0012, 0.0024
Atlas, dry-air mass (native operator)	MPGD (c)	4	1, 2, **4**, 8, 16
	MPGD + norm match (λ_0)	0.0005	0.000125, 0.00025, **0.0005**, 0.001, 0.002, 0.004
	MPGD + noisy (c)	2	1, **2**, 4, 8
	Ours (λ_0)	0.0003	0.000075, 0.00015, 0.0002, 0.00025, **0.0003**, 0.0004, 0.0006 (0.00125 and ab

the initial residual exactly, loses far more skill than any gradient method on both backbones. Dry-air-mass guidance is neutral to slightly positive for all three variables on both backbones, except MMPS on GenCast, which loses t_{850} and u_{10} skill at both leads.

I Dry-air-mass guidance on Atlas with the native operator

Atlas predicts surface pressure and total column water vapour directly, so on Atlas the dry-air mass can be evaluated with a *native* operator that reads those two channels, as well as with the *derived* operator of Appendix B.2, which reconstructs surface pressure from mean sea-level pressure and integrates specific humidity and is the only option on GenCast. The main text uses the derived operator on both backbones so that the two rows of Table 2 solve the same inverse problem. This appendix reports the native operator: guidance objective, calibration scale σ_M (Table 6) and evaluation distance D_{ERA5} all use it, the scales were selected on the development set with the same rule as everywhere else (native rows of Table 10), and the test dates are the six March 2024 initializations of the other Atlas results.

Results. Table 15 shows that every gradient method removes 74–87 % of the native-mass distance to ERA5, against 57.8 % for the best of eight unguided members, at the same skill cost of about 0.4 % in z_{500} and MSLP CRPS at 12 h. The ordering among the methods, however, is the reverse of the derived-operator ordering in the main text: plain MPGD is the best arm (+86.5 %), each IGG component lowers the gain, and the full method is the weakest of the four (+74.1 %). Projection is also strong here (+76.8 % at no skill cost), whereas with the derived operator it is the weakest arm.

Table 16 scores every Atlas mass arm with both operators. Each arm improves the mass it was guided on, but guidance on the derived mass is neutral to harmful for the native mass (the derived-operator projection makes it 51.8 % worse than unguided), whereas guidance on the native mass still removes 28–55 % of the derived distance. Enforcing the mean-sea-level-pressure reconstruction of the mass

Table 10: Guidance scales used on the test set and the development-set ladders they were selected from (bold: selected). ζ': DPS step (Eq. 21); c: constant MPGD step in standardized units (Eq. 24); λ_0: norm-matched fraction (Algorithm 1); τ: MMPS observation noise with $n = 2$ solver iterations. Dagger: pick at a ladder edge.

Backbone / task	Method	Scale	Development ladder
GenCast, hydrostatic	DPS (ζ')	750	300, 450, 600, **750**, 1200
	MPGD (c)	150	40, 75, 100, **150**, 200, 300, 600
	MPGD + norm match (λ_0)	0.0125	0.006, 0.009, **0.0125**, 0.018, 0.025, 0.05, 0.1
	MPGD + noisy (c)	375	100, 150, 200, 250, 300, **375**, 450 (600 diverged)
	Ours (λ_0)	0.07	0.025, 0.035, 0.05, **0.07**, 0.1
GenCast, dry-air mass	DPS (ζ')	22	4, 8, 11, 15, **22**, 30, 60
	MMPS (τ, $n = 2$)	0.05†	0.2, 0.1, 0.07, 0.06, **0.05**
	MPGD (c)	8	2, 4, 6, **8**, 11, 15, 30
	MPGD + norm match (λ_0)	0.0025	0.0006, 0.00125, **0.0025**, 0.005, 0.01, 0.025
	MPGD + noisy (c)	8	2, 4, 6, **8**, 11, 15, 30
	Ours (λ_0)	0.025	0.0025, 0.005, 0.01, **0.025**, 0.05
Atlas, hydrostatic	MPGD (c)	150	50, 100, **150**, 200, 400, 800
	MPGD + norm match (λ_0)	0.001	0.00025, 0.0005, **0.001**, 0.002, 0.004
	MPGD + noisy (c)	300	50, 100, 200, **300**, 400
	Ours (λ_0)	0.01	0.0025, 0.005, **0.01**, 0.02
Atlas, dry-air mass	MPGD (c)	16	4, 8, **16**, 32
	MPGD + norm match (λ_0)	0.0005†	**0.0005**, 0.001, 0.002, 0.004
	MPGD + noisy (c)	16†	2, 4, 8, **16**
	Ours (λ_0)	0.0006	0.0003, **0.0006**, 0.0012, 0.0024

Table 11: Cost per sampling step of each method on GenCast, relative to unguided sampling. Network backward passes and derivative evaluations of $\mathcal{R}$ are counted separately, since the latter are orders of magnitude cheaper. For MMPS, n is the number of solver iterations; each iteration applies A and $A^\top$ once and costs one network vector–Jacobian product, and forming the score costs one more of each.

Method	Network backward	$\nabla\mathcal{R}$
Unguided	0	0
Best-of-k	0	0
Projection	0	**[n]**
DPS	1	1
MMPS (n iterations)	n+1	2n+1
MPGD	0	1
IGG (ours)	0	1

moves p_{msl} in a way that does not conserve the surface-pressure mass. The absolute gains are also not comparable across the two columns: the derived distance to ERA5 carries the semi-diurnal cycle of ERA5's own reconstructed mass, which no forecast can remove, so its gains saturate lower. These results are the basis of the sensitivity to the constraint formulation stated in Sec. 5: the choice of operator changes both the ranking of methods and the transfer between physically equivalent targets, and we do not have a rule for making that choice in advance.

I.1 Spectrum

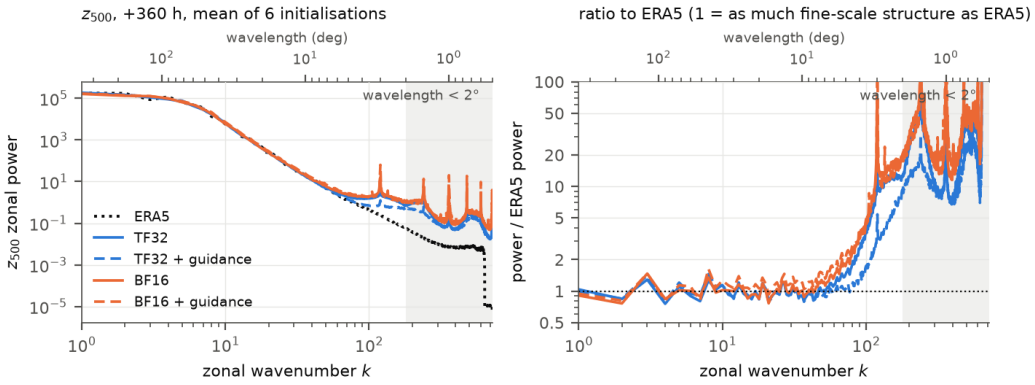

Figure 4: (Atlas with hydrostatic guidance) Zonal power spectrum of z_{500} at +120 h on the test dates (mean over dates of the eight-member mean spectrum), Left: power against zonal wavenumber k; right: ratio to the ERA5 spectrum at the same valid time. The shaded band marks wavelengths below 2°.

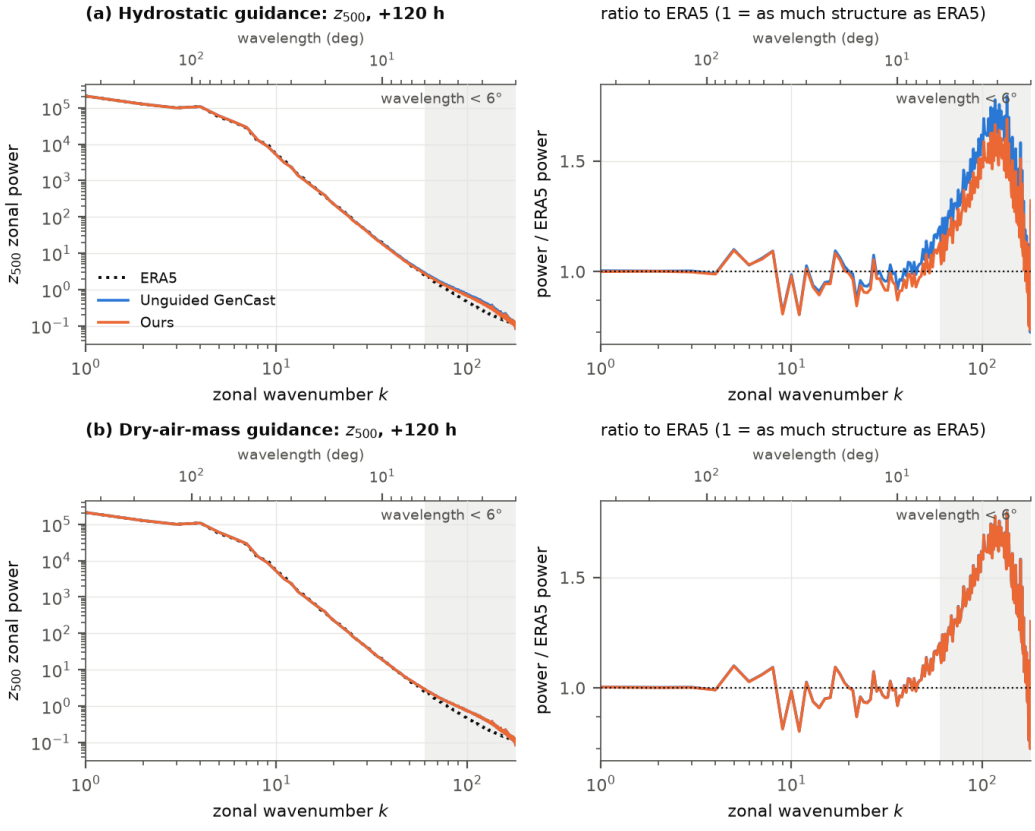

Figure 5: Zonal power spectrum of z_{500} at +120 h on the GenCast test set (24 initializations, mean over dates of the eight-member mean spectrum, 1° grid), unguided (blue) and with guidance (orange), for (a) hydrostatic and (b) dry-air-mass guidance. Left: power against zonal wavenumber k; right: ratio to ERA5. The shaded band marks wavelengths below 6°. Hydrostatic guidance removes part of the excess above $k \approx 60$ and leaves the synoptic band unchanged; dry-air-mass guidance, which does not act on geopotential or temperature, leaves the spectrum unchanged.

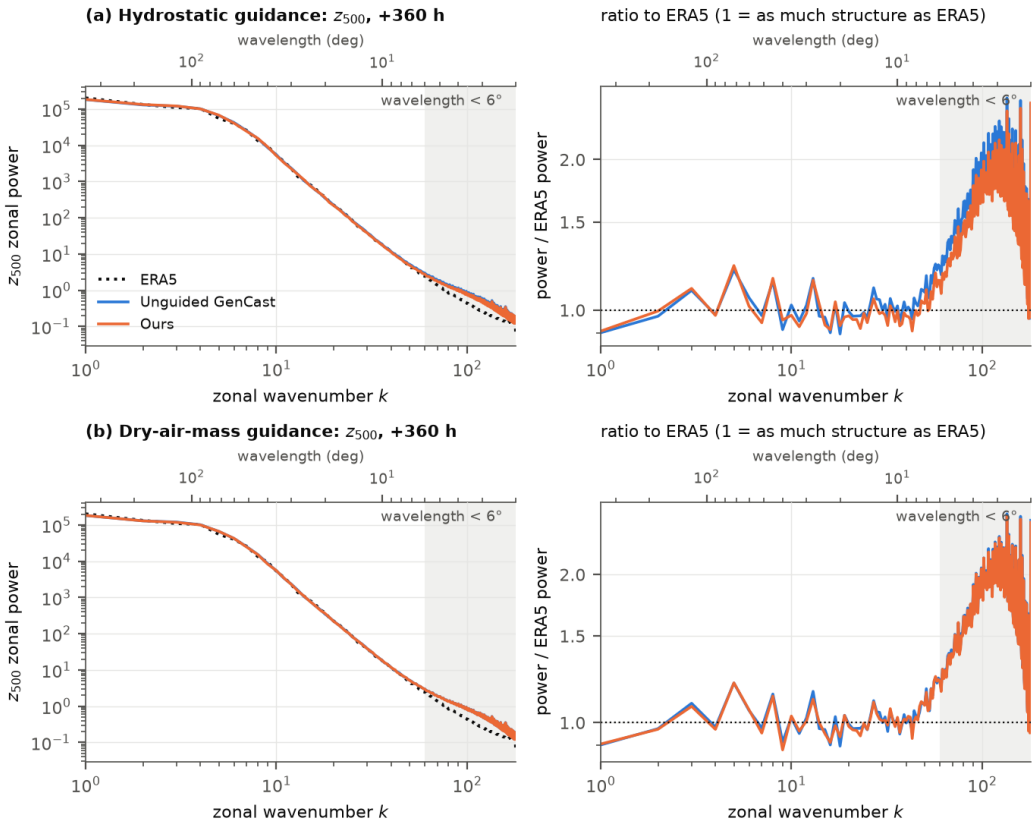

Figure 6: Zonal power spectrum of z_{500} at +360 h on the GenCast test set (24 initializations, mean over dates of the eight-member mean spectrum, 1° grid), unguided (blue) and with guidance (orange), for (a) hydrostatic and (b) dry-air-mass guidance. Left: power against zonal wavenumber k; right: ratio to ERA5. The shaded band marks wavelengths below 6°. Hydrostatic guidance removes part of the excess above $k \approx 60$ and leaves the synoptic band unchanged; dry-air-mass guidance, which does not act on geopotential or temperature, leaves the spectrum unchanged.

Table 12: Inference cost of Atlas per precision mode, with and without hydrostatic guidance, on one NVIDIA GB200 GPU (one ensemble member, 100 sampler steps per 6 h step). s/step: steady-state wall-clock seconds per 6 h autoregressive step, excluding the warm-up first step; speed-up is relative to unguided FP32. Torch peak: `max_memory_allocated` (PyTorch tensors only). nvidia-smi peak: total device memory in use, sampled every 10 s (includes the CUDA context and allocator cache). Modes: FP32, float32 weights without autocast and with TF32 disabled; TF32, float32 with TF32 tensor-core matrix products (the mode of all reported runs); AMP, float32 weights and state with bfloat16 autocast; BF16, bfloat16 weights, buffers and autocast. The sampler state and the physics objective are float32 in every mode. Guidance adds 0.8–1.2 s per step, at most 5% of the unguided time.

Mode	s/step	Speed-up	Torch peak (GiB)	nvidia-smi peak (GiB)
FP32	126.2	1.0×	26.8	33.4
FP32 guided	127.2	1.0×	26.8	33.4
TF32	24.0	5.3×	26.8	33.5
TF32 guided	25.2	5.0×	26.8	33.5
AMP	20.2	6.2×	24.0	30.5
AMP guided	21.0	6.0×	24.0	30.5
BF16	20.0	6.3×	15.1	20.6
BF16 guided	20.8	6.1×	15.1	20.6

Table 13: Hydrostatic-balance guidance: fair CRPS on variables the constraint does not act on, test set. Entries are % change relative to the unguided baseline, $100\,(1 - \text{method}/\text{baseline})$, so higher is better; *: paired date bootstrap 95 % interval excludes zero. GenCast: 24 dates; Atlas: six March 2024 dates.

		t_{850}		q_{700}		u_{10}	
Model	Method	12 h	120 h	12 h	120 h	12 h	120 h
GenCast	Projection	-0.43^*	-11.23^*	-0.02^*	-5.51^*	$+0.00$	-5.04^*
	DPS	-0.01^*	-0.32^*	-0.00^*	-0.49^*	-0.00^*	-0.33^*
	MPGD	-0.00	-0.25^*	-0.00	-0.51^*	-0.00	-0.25^*
	MPGD + norm match	-0.02^*	-0.20^*	$+0.00$	-0.27^*	-0.00^*	-0.19^*
	MPGD + noisy	-0.00	-0.48^*	-0.00^*	-0.54^*	$+0.00$	-0.41^*
	Ours	-0.00	-0.20^*	$+0.00^*$	-0.27^*	$+0.00$	-0.22^*
Atlas	Projection	-23.50^*	-93.48^*	-3.97^*	-28.19^*	-10.15^*	-46.35^*
	MPGD	-0.26^*	-0.62^*	-0.44^*	-2.18^*	-0.13^*	-0.91^*
	MPGD + norm match	-0.11^*	-0.19	-0.17^*	-1.11^*	-0.04^*	-0.43^*
	MPGD + noisy	-0.24^*	-0.61^*	-0.51^*	-2.25^*	-0.12^*	-0.99^*
	Ours	-0.13	-0.45^*	-0.26^*	-1.25^*	-0.03^*	-0.58^*

Table 14: Dry-air-mass guidance: fair CRPS on variables the constraint does not act on, test set. Same conventions as Table 13.

		t_{850}		q_{700}		u_{10}	
Model	Method	12 h	120 h	12 h	120 h	12 h	120 h
GenCast	Projection	-0.00	$+0.04$	-0.00	-0.05^*	$+0.00$	$+0.03$
	DPS	$+0.05^*$	$+0.16^*$	$+0.01^*$	-0.01	-0.01^*	$+0.01$
	MMPS	-0.19^*	-0.69^*	-0.01	-0.04	-0.21^*	-0.28
	MPGD	$+0.04^*$	$+0.12^*$	$+0.00$	-0.03^*	-0.00^*	$+0.02$
	MPGD + norm match	$+0.08^*$	$+0.15^*$	$+0.00$	-0.01	-0.01^*	-0.00
	MPGD + noisy	$+0.03^*$	$+0.12^*$	$+0.00$	-0.03^*	-0.00^*	$+0.02$
	Ours	$+0.03^*$	$+0.12^*$	$+0.00$	-0.02^*	-0.00^*	$+0.02$
Atlas	Projection	$+0.03$	$+0.05$	$+0.01^*$	$+0.04$	-0.05^*	$+0.00$
	MPGD	$+0.06^*$	$+0.10$	$+0.02^*$	$+0.01$	-0.02^*	-0.02
	MPGD + norm match	$+0.06^*$	$+0.13$	$+0.02^*$	$+0.04$	-0.02^*	-0.00
	MPGD + noisy	$+0.05^*$	$+0.09$	$+0.01^*$	$+0.04$	-0.05^*	-0.01
	Ours	$+0.05^*$	$+0.11$	$+0.01^*$	$+0.04$	-0.01^*	$+0.02$

Table 15: Dry-air-mass guidance on Atlas with the native operator, six March 2024 test dates. Entries are % change relative to unguided sampling, $100\,(1 - \text{method}/\text{unguided})$, so higher is better in every column. Int.: D_{ERA5} integrated over 12–360 h. Best-of-k ($k = 8$) returns a single member, so no CRPS is defined. Bold: best per column.

	D_{ERA5}				CRPS			
					12 h		120 h	
Method	Int.	12 h	120 h	360 h	z500	mslp	z500	mslp
Best-of-k	$+57.8$	$+2.0$	$+57.1$	$+59.7$	–	–	–	–
Projection	$+76.8$	$+71.2$	$+76.2$	$+77.9$	$\mathbf{-0.00}$	$\mathbf{-0.02}$	$+0.02$	$\mathbf{+0.04}$
MPGD	$\mathbf{+86.5}$	$\mathbf{+74.0}$	$\mathbf{+85.2}$	$+87.4$	-0.35	-0.38	$+0.10$	-0.07
+ norm matching	$+84.6$	$+61.6$	$+82.6$	$\mathbf{+87.6}$	-0.37	-0.36	$\mathbf{+0.13}$	-0.05
+ intermediate state	$+80.9$	$+72.0$	$+81.6$	$+82.1$	-0.23	-0.28	$+0.04$	-0.28
Ours (both)	$+74.1$	$+49.1$	$+73.6$	$+81.3$	-0.36	-0.37	$+0.06$	-0.23

Table 16: Every Atlas mass arm scored with both operators (six March 2024 test dates). Entries are the % change of the D_{ERA5} integral relative to unguided sampling when the mass is evaluated with the native operator (left) or the derived operator (right); each arm is guided with the operator named in its block. Higher is better.

Guided on	Method	Native D_{ERA5}	Derived D_{ERA5}
native mass	Projection	$+76.8$	$+1.2$
	MPGD	$+86.5$	$+47.3$
	+ norm matching	$+84.6$	$+54.6$
	+ intermediate state	$+80.9$	$+28.0$
	Ours (both)	$+74.1$	$+41.4$
derived mass	Projection	-51.8	$+52.8$
	MPGD	$+27.3$	$+54.1$
	+ norm matching	$+41.6$	$+53.3$
	+ intermediate state	-13.2	$+61.6$
	Ours (both)	$+11.0$	$+67.3$