How much ensemble does a frozen weather model contain?

Anonymous authors
Paper under double-blind review

Abstract

Data-driven weather models now match numerical weather prediction; to represent uncertainty, the field trains probabilistic models end to end, at the full training cost of the forecaster. We ask instead how much ensemble skill a pretrained deterministic model provides with its weights frozen, measured against its own forecast. HyperENS trains low-rank adapters on the backbone's attention projections and a hypernetwork that generates their cores from a Gaussian latent: each new latent is a new member, whose adapter stays fixed along its forecast. The member distribution is trained end to end with an almost-fair CRPS on the forecasts and their power spectra, and no spread amplitude is tuned afterwards. Applied to the pretrained 0.25° Aurora, HyperENS+, which also adapts the encoder's query latents, trains 3,235,544 parameters, 0.26% of the model, within two days on eight A100 GPUs for both training runs together, and leaves the original deterministic forecast unchanged. Scored against ERA5 on 178 initialisations of 2024, averaging per-variable relative scores over eight variables, its 32-member ensemble mean has 18.8% and 24.7% lower RMSE than the frozen forecast at days five and ten, and its fair CRPS is 0.569 and 0.525 times that forecast's mean absolute error. Aurora 1.5's ensemble, a whole-model fine-tune, has the lower absolute fair CRPS: ours is 9.6% and 6.8% higher at days five and ten. The mean spread–skill ratio is within six percent of one from day three, but short leads remain underdispersed (0.861 at day one) and the rank histograms show residual errors. On an earlier checkpoint evaluated on 418 common initialisations, the adapter's static part alone is only slightly better than the frozen forecast; most of the gain appears only in the ensemble.

1 Introduction

The last few decades have seen a steady rise in the quality of weather forecasts (Bauer et al., 2015; Haiden et al., 2025). One of the main reasons is the advance of data assimilation, which merges observations of every kind into a prior numerical weather prediction (NWP) forecast and so produces initial conditions of a quality no observing system alone could give. Applied retrospectively, the same techniques yielded the reference representation of the atmospheric state, the ERA5 reanalysis (Hersbach et al., 2020). The past five years have accordingly seen deep networks trained on this archive, a gap-free record that is itself a model product, as emulators of the atmosphere. Despite coarse early resolutions and solutions that were not always physical (Bonavita, 2024), these models have progressed to the ensemble (Price et al., 2025; Bonev et al., 2025; Alet et al., 2025; Xiong et al., 2025; Lang et al., 2026; Weyn et al., 2026; Rasp et al., 2026), the form that forecasting a chaotic system such as the atmosphere requires. In contrast to a deterministic model, an ensemble produces not one evolution of the atmosphere but a distribution of them, sampling the growth of errors in the initial state, the sensitive dependence on initial conditions that Lorenz (1963) established, and, in operational systems, the errors of the model itself, which stochastic physics was introduced to represent (Buizza et al., 1999).

Two arguments carry data-driven models into operations: i) they can learn from data relationships that the governing equations and their parametrisations do not capture, and ii) they match NWP skill at orders of magnitude lower inference cost. The cost is concentrated in training, hundreds to thousands of accelerator-days for the current global models at quarter-degree grid spacing (about 25 km), and is paid once, and paid again, multiplied, for the ensemble. Deterministic training is ac-

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.

complished with an L1 or L2 loss at a batch size of one sample per device, so that the effective batch equals the number of devices; the standard probabilistic objectives are evaluated over an ensemble of members, so every sample now costs M forward passes and the number of independent samples per step falls by that factor (Lang et al., 2026). Diffusion models are the exception: trained sample by sample on a denoising objective, they pay at inference instead, with tens of network evaluations per step (Price et al., 2025).

This naturally prompted the community to convert deterministic models into ensembles by comparatively short fine-tuning (Xiong et al., 2025; Weyn et al., 2026). These methods, however, tune the entire model, which at foundation scale is still expensive, and they produce a second model to maintain. When downstream post-processing and products are built on a deterministic forecast, keeping that forecast unchanged has value of its own: a retrofit leaves the deterministic model bit-identical and serves the ensemble from the same weights plus a small adapter.

In this paper we ask a simple question: can an already-trained deterministic model be turned into an ensemble without changing its weights, with the member distribution trained rather than injected? Or, rephrased: how much ensemble does a frozen weather model contain? The contributions are the following: i) HYPERENS, a probabilistic retrofit for frozen models: a hypernetwork turns the $r \times r$ core of a low-rank adapter into a random variable, and the member distribution is trained end to end with a CRPS-based objective on the forecasts (Sec. 3); ii) an evaluation of the retrofit of a frozen Aurora against its own deterministic forecast and against the whole-model ensemble fine-tune of Aurora 1.5 on shared 2024 initialisations, in skill and spread–skill consistency; iii) a decomposition into the frozen backbone, the static part of the adapter, a single member and the ensemble, in which the static part alone is only slightly better than the frozen forecast and most of the gain appears only in the ensemble; iv) evaluation-time ablations of the spread amplitude and of the direction of the perturbation, and an analysis of the short-lead under-dispersion, which is smaller in the arm that also perturbs the encoder's query latents. All four are in Sec. 4, in that order.

2 RELATED WORK

Ensembles by perturbing deterministic models from outside. The first data-driven ensembles attached untrained stochasticity to trained deterministic forecasters: Gaussian initial-condition noise in FourCastNet (Pathak et al., 2022), Perlin noise in Pangu-Weather (Bi et al., 2023), perturbed analyses fed to GraphCast (Lam et al., 2023; Price et al., 2025), random and singular-vector initial perturbations of a low-resolution network (Scher & Messori, 2021), and post-hoc distributional regression (Bülte et al., 2026). They replay the lesson of the earliest Monte Carlo forecasting experiments (Hollingsworth, 1979): perturbations imposed from outside, with no dynamical structure, are damped faster than they grow, and the ensembles are under-dispersed; the uncertainty is a property of the input, and nothing in the network is asked to represent it. SEEDS (Li et al., 2024) is the learned counterpart: a diffusion model that extends an operational ensemble from a few of its members, learned from that ensemble's own samples rather than from a deterministic model's weights.

End-to-end probabilistic training. The current state of the art moves the uncertainty inside the optimisation: a generative sampler learned by diffusion (GenCast, Price et al. 2025), the whole trunk trained under CRPS-based losses (AIFS-CRPS, Lang et al. 2026; FuXi-ENS, Zhong et al. 2025), noise from diffusion processes on the sphere injected at every step (FourCastNet 3, Bonev et al. 2025), or a shared low-dimensional noise draw conditioning the whole network plus a deep ensemble of seeds for epistemic coverage (FGN, Alet et al. 2025; Rasp et al. 2026). U-Cast (Cachay et al., 2026) shows that even a dropout inlet suffices once the trunk is fine-tuned under CRPS, and that for a single 1.5° model the probabilistic fine-tuning is about fifteen percent of the combined pre-training and fine-tuning cost. Diaconu et al. (2026) retrofit deterministic PDE emulators by training the backbone together with a noise-conditioning branch under a CRPS objective; HyperENS keeps the backbone fixed. Whatever the injection mechanism, *the trunk itself is the object of probabilistic training*, at the full training cost of the forecaster, modest at 1.5° and foundation-scale at 0.25°. Even so the initial-condition class is not dispensed with: GenCast's ablations find EDA-perturbed initialisation (Bonavita et al., 2016) helps beyond the shortest leads, and the operational AIFS ENS starts from the IFS ensemble's perturbed analyses on top of its own stochasticity (Lang et al., 2026); the two classes remain complementary.

Distributions over weights. A third family represents uncertainty as a distribution over parameters: Bayesian hypernetworks (Krueger et al., 2018), HyperGAN (Ratzlaff & Fuxin, 2019) and multiplicative normalising flows (Louizos & Welling, 2017) map a latent to generated weights; BatchEnsemble (Wen et al., 2020) and rank-1 Bayesian networks (Dusenberry et al., 2020) perturb shared weights with per-member rank-one factors; MC-dropout (Gal & Ghahramani, 2016), Laplace approximations and deep ensembles (Lakshminarayanan et al., 2017) are the generic tools, applied to the adapter itself as a post-hoc Laplace approximation (Yang et al., 2024), variational factors (Wang et al., 2024) or adapter ensembles (Mühlematter et al., 2026). For weather, Xiong et al. (2025) fine-tune the mean and variance of a Gaussian posterior over the weights of a Pangu-class model under an L1 loss with a KL term, and add a learned perturbation of the input state scaled by the six-hour increment; their weight-noise-only ablation, perturbing *all* parameters with no low-rank structure, stays under-dispersed at every lead. What the family shares is how the shape of the distribution is chosen: by SGD's accidents for seed ensembles, in parameter space against a prior or a likelihood for posteriors, not by a score on the forecasts.

Low-rank adapters and hypernetworks. Low-rank adaptation is standard (Hu et al., 2022), and a small square matrix between the factors is a recurring idea: AdaLoRA (Zhang et al., 2023) prunes a diagonal core by importance, LoRA-XS (Bałazy et al., 2025) trains only an $r \times r$ core between factors frozen from the SVD of the pretrained weight, structurally the $B\,C\,A$ we use with learned templates in place of the spectrum, and VeRA (Kopiczko et al., 2024) shares one random factor pair across layers with per-layer scalings. Generating adapters with a hypernetwork is likewise established (Ha et al., 2017): HyperFormer (Mahabadi et al., 2021) conditions on task and layer embeddings, the ancestor of our (stage, type, block) embedding, and HyperDreamBooth (Ruiz et al., 2024) generates LoRA weights outright. Closer still, ScalaBL (Samplawski et al., 2025) learns a variational distribution over a diagonal core between low-rank factors, Singular Value Ensembles (Turkoglu et al., 2026) train per-member singular values on a frozen singular basis, and Aurora itself uses LoRA on its attention projections for roll-out fine-tuning (Bodnar et al., 2025). HyperENS generates dense $r \times r$ cores for all adapted layers from one latent per member, over a frozen weather backbone, and trains the member distribution with a CRPS-based objective on the forecasts.

This work. Our method combines three elements, none of which is new on its own. First, the backbone stays frozen: we do not ask whether end-to-end probabilistic training yields calibrated ensembles, but how much ensemble skill can be obtained from a deterministic model whose weights stay unchanged. Second, the adapter core is a random variable, and its distribution is trained end to end with a CRPS-based objective on the forecasts, with no amplitude tuned after training. Third, the perturbation belongs to the model-error class of NWP, like SPPT (Buizza et al., 1999) and SKEB (Berner et al., 2009): all members are identical at $t = 0$, and spread grows only as the perturbed model is integrated. Alet et al. (2025) treat a learned, CRPS-fitted distribution over parameters as aleatoric and cover epistemic uncertainty with an ensemble of independently trained models; our cores are of the former kind, held fixed per member rather than redrawn at every step, and silent about the between-model axis that independent seeds address.

3 METHODS

3.1 HYPERENS: A HYPERNETWORK OVER LOW-RANK CORES

Let f_W be the deterministic forecast model with weights W, here Aurora (Bodnar et al., 2025): it maps the atmospheric state at $t - 6\,\mathrm{h}$ and t to the state at $t + 6\,\mathrm{h}$ and is applied autoregressively. We take Aurora's public pretrained checkpoint, the deterministic model before any fine-tuning, and every one of its weights stays frozen throughout this work. An ensemble member is the frozen model with a member-specific low-rank update added to a fixed subset of its weight matrices. The update is a deterministic function of a member latent $z \sim \mathcal{N}(0, I_{32})$, so a fresh draw of z is a fresh member and the ensemble size is unlimited at inference. Within a member the update is fixed in time and applied identically at every autoregressive step, which places the scheme in the model-error perturbation class of numerical weather prediction, the class SPPT (Buizza et al., 1999) belongs to, rather than in the initial-condition class.

The adapted subset is the fused query–key–value (QKV) projection and the output projection of every Swin attention block (Liu et al., 2021) of the backbone, encoder and decoder stages included:

$J = 96$ modules, two per block. For module j, write $W_j \in \mathbb{R}^{d_j^{\text{out}} \times d_j^{\text{in}}}$ for the frozen weight. Each module carries a pair of trained templates $A_j \in \mathbb{R}^{r \times d_j^{\text{in}}}$ and $B_j \in \mathbb{R}^{d_j^{\text{out}} \times r}$ of rank $r = 8$, shared by every member; member dependence enters only through a small $r \times r$ core generated by a hypernetwork:

$$W_j(z) \;=\; W_j + \Delta W_j(z), \qquad \Delta W_j(z) \;=\; B_j\, C_j(z)\, A_j, \qquad C_j(z) \;=\; I_r + s\, \tanh\big(\delta_j(z)\big), \tag{1}$$

where $\delta_j(z) \in \mathbb{R}^{r \times r}$ is the generated core, $\tanh$ acts elementwise and s is a fixed core scale ($s = 1$ throughout). Three properties of this form are used deliberately. With $C_j(z) = I_r$ the update is exactly a conventional member-independent low-rank fine-tune $B_j A_j$, so a hypernetwork that learns nothing degrades to a static fine-tune and not to a broken model; with B_j zero-initialised, $\Delta W_j(z) \equiv 0$ at initialisation and the ensemble reproduces the pretrained model exactly. The $\tanh$ bounds every entry of the core perturbation by s, so that for fixed templates $\|\Delta W_j\|$ is bounded uniformly in z, which also guards low-precision training against tail draws of z. **Member-to-member variation is confined to the fixed r-dimensional column space of B_j and row space of A_j: diversity mixes a learned set of directions per module and cannot invent new ones.**

The hypernetwork. One network serves all J modules. Its input is the latent concatenated with a learned module embedding $e_j \in \mathbb{R}^{128}$, itself the sum of three embedding-table rows selected by the module's encoder or decoder stage, its projection type (QKV or output) and its block position within the stage:

$$\delta_j(z) \;=\; g_\phi\big([\, z;\, e_j \,]\big), \qquad e_j \;=\; e^{\text{stage}}_{\sigma(j)} + e^{\text{type}}_{\tau(j)} + e^{\text{block}}_{\beta(j)}. \tag{2}$$

The trunk g_ϕ is a two-layer perceptron, $\mathbb{R}^{32+128} \to \mathbb{R}^{512} \to \mathbb{R}^{512}$ with SiLU activations, followed by a linear head to the $r^2 = 64$ entries of the core, reshaped to $r \times r$. The head is zero-initialised in weight and bias, so $\delta_j \equiv 0$ and $C_j = I_r$ at initialisation, the identity invariant above. All J cores are produced in one batched call with the module index as the batch axis, once per forward pass of the backbone and at negligible cost beside it; the factored embedding is what lets one trunk share structure across the modules instead of learning each core on its own.

Two arms. *HyperENS* adapts the J backbone modules and nothing else. *HyperENS+* additionally adapts the encoder's learned query latents: the 3 vectors of dimension 512 with which the Perceiver encoder's (Jaegle et al., 2021) level-aggregation cross-attention interrogates the input. They form one 3×512 tensor, so Eq. (1) applies unchanged, with its own rank-r templates and a core from the same hypernetwork under its own module embedding. That tensor is broadcast over every spatial column, so its perturbation is global and enters through the encoding of the state rather than through its integration: members of HyperENS+ differ from the first step, members of HyperENS only once the perturbed processor has acted, and Sec. 4 (*The short-lead deficit*) prices the difference. The trainable set is 3,231,040 parameters for HyperENS and 3,235,544 for HyperENS+, 0.26% of the frozen model (Table 2, App. B).

Static–stochastic decomposition. Let $E_j = \mathbb{E}_z[\tanh \delta_j(z)]$ denote the mean core of module j. Eq. (1) splits identically into a member-independent and a member-dependent part:

$$\Delta W_j(z) \;=\; B_j\big(I_r + s\, E_j\big)A_j \;+\; s\, B_j\big(\tanh \delta_j(z) - E_j\big)A_j, \tag{3}$$

where the first term is the static part and the second the stochastic part: the former is the effective fine-tune the ensemble learns as a whole, the latter the only source of ensemble spread. The static part is a forecast model in its own right, the *mean core*: every member identical, spread exactly zero. The decomposition organises the analysis of Sec. 4 and the evaluation-time ablations of App. C; its norms are computed exactly and offline from $r \times r$ Gram matrices (App. F.1).

3.2 Training and inference

The objective is a fixed-weight sum, over the nine training variables, of two latitude-weighted proper scores evaluated on the M_{train}-member ensemble: the almost-fair CRPS of Lang et al. (2026) at three spatial scales, and a CRPS on radially averaged power spectra. Per-term weights are pinned so that every term contributes unit magnitude to the initial objective, with surface variables down-weighted by a quarter. The terms and their estimator are written out in App. A. Training rolls the

model out over four autoregressive 6-h steps and applies the loss at leads 6 and 24 hours. HyperENS is trained in a single pass over ERA5 2008–2019 ($0.25°$, 13 pressure levels) and HyperENS+'s fine-tune over 2020–2022, with AdamW and a cosine schedule restarted on every three-month chunk of data. Both training runs together took under two days on eight A100 80 GB GPUs, at most sixteen A100-days for the HyperENS to HyperENS+ chain, against 456 A100-days for the three ensemble stages of Aurora 1.5 (Weyn et al., 2026), figures that also depend on each site's data pipeline. The ensemble of $M_{\mathrm{train}} = 8$ members is realised as one member per device, each drawing its own latent, and the member forecasts are gathered into one differentiable ensemble before the loss; the layout of the frozen backbone, the adapters and the loss across devices is App. B. M_{train} is a hyperparameter of the gradient estimator, since the fair-CRPS dispersion term averages $\binom{M}{2}$ member pairs. What the retrofit costs, then, is the trainable set above, that training run, and at inference one backbone forward per member and step plus a hypernetwork call that is negligible beside it.

The recipe was trained more than once during development, from cold and warm starts, at $M_{\mathrm{train}} = 4$ and 8, with i.i.d. and with antithetic latents (App. G), with the same qualitative outcome; the runs reported are the longest of each arm, not a selection.

At test time, M_{test} latents are drawn independently, each defining one member whose weight perturbation is held fixed along its full autoregressive trajectory, and ensemble products (mean, spread, probabilistic scores) are computed from the member set. Members differ only through the frozen model's response to bounded perturbations of its weights, and no post-hoc filtering or amplitude tuning is applied at any stage.

3.3 Verification

Forecasts are verified against ERA5 at $0.25°$ on eight variables: 2-m temperature ($T_{2\mathrm{m}}$), the 10-m wind components ($U_{10\mathrm{m}}$, $V_{10\mathrm{m}}$), mean sea-level pressure (MSLP), geopotential at 500 hPa (Z_{500}), temperature and zonal wind at 850 hPa (T_{850}, U_{850}) and specific humidity at 700 hPa (Q_{700}), at forty lead times from six hours to ten days, latitude-weighted throughout. Deterministic skill is the RMSE of the ensemble mean; probabilistic skill is the fair CRPS (Gneiting & Raftery, 2007; Ferro, 2014). Spread–skill consistency is measured by the spread–skill ratio

$$\mathrm{SSR} \;=\; \sqrt{\frac{M+1}{M}}\,\frac{\text{RMS ensemble spread}}{\text{RMSE of the ensemble mean}}, \tag{4}$$

whose prefactor makes a calibrated M-member ensemble read exactly one, values below one meaning an under-dispersed ensemble. We report it as the mean over initialisations of the per-initialisation ratio, not as the ratio of pooled variance to pooled squared error (Fortin et al., 2014). Distributional errors the ratio cannot see are read from the rank histograms (Hamill, 2001), summarised where a scalar is needed by their *tilt*: the pooled mass in the upper half of the bins minus that in the lower half, median bin excluded, zero for a symmetric histogram and positive when the analysis ranks above the members more often than below, that is, when the forecast is biased low. Per-member spectral realism is the mean log-ratio of radially averaged power spectra between member and verification over the upper half of the resolved wavenumbers, ideal zero, negative for a small-scale deficit.

Three conventions hold throughout. Systems are compared only on initialisations they share, and every ratio is computed on that pairing; systems initialised at different synoptic hours are never compared at short lead, because the frozen backbone's own six-hour error differs between them. The campaigns are listed in Table 3 (App. D). Cross-variable numbers are means over the eight variables of per-variable values, for a ratio the mean of per-variable ratios and never a ratio of native-unit sums, which Z_{500} and MSLP would own. Intervals are 95% paired moving-block bootstraps over initialisations, block length about $n^{1/3}$, the dates resampled once and both systems recomputed on the resample before the ratio is taken; they cover sampling within 2024 and say nothing about other years. Calibration metrics are within-model absolutes and are never ratioed between systems; skill metrics are ratios, candidate over reference, so below one is better. Every caption states the sample size and the window behind it.

4 RESULTS

Every number that follows comes from one campaign on 2024, a year no system was trained on: 32-member ensembles, 40 leads to 10 d, the 8 variables and the conventions of Sec. 3.3, the same ERA5 for every system. Two arms are scored throughout, HyperENS and HyperENS+ (Sec. 3.1), whose parent is the frozen Aurora backbone. The external reference is Aurora 1.5's ensemble (Weyn et al., 2026), a fine-tune of the whole model, scored beside its deterministic counterpart, the deterministic Aurora 1.5 released with it. Both are run here from ERA5, on which their final fine-tuning stages were not trained, and without perturbed initial conditions, whereas the report's own scorecard starts the members from perturbed IFS analyses (Weyn et al., 2026); Aurora 1.5 may be handicapped by both. We did not retrofit that deterministic model: it consumes an extended set of input fields that the ERA5 archive the retrofit is trained from, WeatherBench2 (Rasp et al., 2024), does not carry. The comparison is therefore across two deterministic models, and that is why every gain below is normalised on the deterministic model it was measured against. Table 3 (App. D) lists the campaigns with their windows and sample sizes. HyperENS+ appears through two checkpoints (Table 3): the later one in the comparison with Aurora 1.5, the earlier one in the decomposition and the ladders, which were run before it existed.

How much ensemble does a frozen model contain? The question is answered against each system's own deterministic forecast: the retrofit against the frozen Aurora it perturbs, Aurora 1.5's ensemble against the deterministic Aurora 1.5, all four scored on the same 178 initialisations (2 Jan–21 Dec 2024). Fig. 1(a, b) is the paper's answer. HyperENS+ takes the RMSE of the ensemble mean to 0.937 of the frozen model's at day one, 0.812 at day five and 0.753 at day ten, and the fair CRPS to 0.677, 0.569 and 0.525 of the deterministic MAE. Aurora 1.5's ensemble, against its own deterministic model, reaches 0.872 and 0.761 in RMSE and 0.630 and 0.543 in fair CRPS at days five and ten; through the first day its ensemble mean is no better than that model (1.003). **Each measured against its own deterministic model, HyperENS+ has the lower RMSE ratio at every lead and the lower fair-CRPS ratio from eighteen hours on; at six and twelve hours Aurora 1.5's ensemble has the lower CRPS ratio.** HyperENS, the backbone-only arm, is a few percent behind HyperENS+ and level with Aurora 1.5 by day ten. Variable by variable (Fig. 1) the ordering holds, with one exception: on Z_{500} HyperENS+'s day-one gain is far larger than on any other variable, and *Where the skill comes from* says why.

On a second backbone, a small PyTorch re-implementation of Pangu-Weather (Bi et al., 2023) trained to the 6-h step from a single input time and scored on 154 initialisations of 2019, HyperENS reaches the same day-ten ratios as on Aurora but a larger short-lead deficit, and its absolute errors stay those of its weaker backbone (App. D.7).

Is the spread consistent with the error? The mean spread–skill ratio over the eight variables is within six percent of one from day three (Fig. 1c); the rank histograms nevertheless show residual distributional errors. Aurora 1.5 is within a few percent of one from day one. The retrofit starts under-dispersed, since every member is identical at $t = 0$ and spread has to be manufactured by integrating a weight perturbation forward: 0.861 at day one, then it crosses one before day three and settles slightly above it (1.037 at day five, 1.057 at day ten), within six percent of one from day three, slightly over-dispersed, against Aurora 1.5's within four. The mean hides a heterogeneity the figure shows: the mass field is dispersed enough from the first day, the boundary layer, the winds and the moisture are not, and by day ten $T_{2\mathrm{m}}$ has overshot while the winds sit at one. HyperENS never gets there, under-dispersed at every lead (*The short-lead deficit*). The spread is what the frozen model has: the members are its own responses to bounded perturbations of its weights, no amplitude was tuned after training, and the trained amplitude is the day-five optimum among the scales tested (*Sensitivity to perturbation scale*); whether perturbed initial conditions can buy the rest is open (a pilot: App. E).

Against Aurora 1.5 on the same initialisations the retrofit's fair CRPS is 1.096 times Aurora 1.5's at day five (1.037 at day one, 1.068 at day ten; Table 4, App. D): the whole-model fine-tune is the system with the lower fair CRPS, by a few percent, and its deterministic counterpart is the better model. Member spectra (Fig. 6) stay on the frozen backbone's own curve on $T_{2\mathrm{m}}$, a small-scale deficit at every lead, where Aurora 1.5's members carry a surplus; and the rank histograms (Fig. 5) show what a scalar cannot: HyperENS+ is a U at six hours and a shallow dome from day three

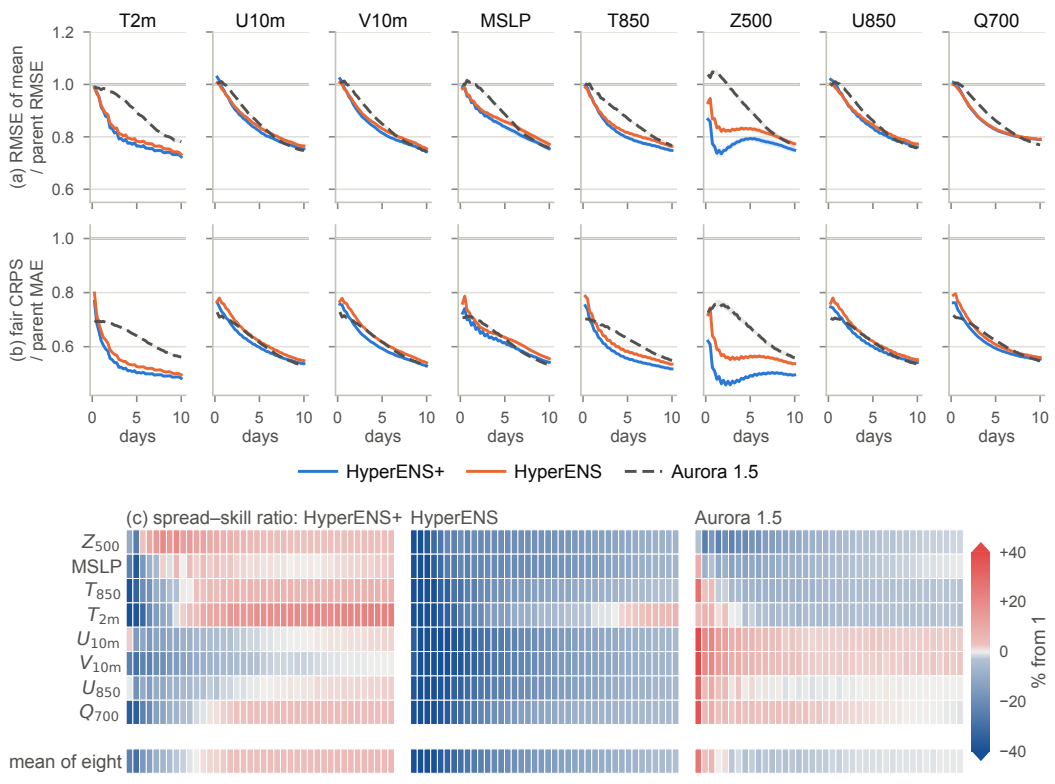

Figure 1: The gain over the deterministic forecast, per variable, and the spread–skill ratio behind it. (a) RMSE of the ensemble mean over the deterministic RMSE and (b) fair CRPS over the deterministic MAE, against lead, one column per variable, for HyperENS+ and HyperENS over the frozen Aurora backbone and for Aurora 1.5's ensemble over Aurora 1.5's deterministic model; shading is the 95% paired moving-block bootstrap, the solid line parity. (c) Spread–skill ratio as a scorecard: each system's spread–skill ratio minus one, in percent, on its own error, per variable (rows, with their mean) and lead (columns); blue below one, red above one, cells within three percent of parity drawn near-neutral. All four systems paired on 178 initialisations (2 Jan–21 Dec 2024; $M{=}32$). The cross-variable means are Fig. 2 (App. D).

with extreme bins that stay over-populated, a shape error no rescaling of the spread removes, while HyperENS is a U at every lead with a slope on it, under-dispersion and a low bias together.

Where the skill comes from The static–stochastic split of Eq. (3) lets the retrofit be taken apart into models that can each be scored. Four forecasts of one frozen backbone are compared on the same initialisations: the frozen model itself; the mean core, the static part alone, which is what the fine-tune did to the deterministic forecast; one member at a fixed latent, a single draw of the trained distribution; and the 32-member ensemble mean. The chain frozen $\rightarrow$ mean core $\rightarrow$ one member $\rightarrow$ ensemble mean is multiplicative, and Table 1 gives each link as a ratio to the one before it. *The fine-tune alone changes little:* the mean core's RMSE is 0.968 of the frozen model's at day five and its MAE 0.929, three to seven percent. *One draw is deliberately worse than the mean core*, by 1.188 in RMSE at day five: a member of a calibrated ensemble carries the error of the forecast plus its own perturbation, and it is meant to. *The averaging buys that back and more:* the ensemble mean has 0.711 of one member's error, which lands it at 0.815 of the frozen model's RMSE, with fair CRPS at 0.572 of the deterministic MAE. For this earlier checkpoint on 418 common initialisations, then, the mean-core model has 0.968 of the frozen model's RMSE and 0.929 of its MAE at day five, and the full ensemble 0.815 and 0.572. **The static part is a small share of the gain; the rest appears only with the ensemble.** Because the forecast model is nonlinear, the mean-core forecast is not the ensemble mean, and the difference between the two cannot be attributed to the averaging alone. HyperENS tells the same story with a smaller stochastic part (Table 5 and Fig. 3, App. D).

Table 1: Where the skill comes from: the frozen backbone taken to the mean core (the static path alone, $C = I + s\,E_j$), the mean core to one stochastic member at a fixed latent, and that member to the 32-member ensemble mean, each stage a ratio of two models scored on the same initialisations; cross-variable means of per-variable ratios; the product of the stages and the measured M=32/frozen ratio differ only because a mean of ratios does not commute with multiplication (Table 5). Fair CRPS is M-corrected and comparable across ensemble sizes, the RMSE of the ensemble mean is not. HyperENS+, paired on the initialisations common to all four rungs (n=418, 27 May–21 Dec 2024); both arms and the per-variable control-run table are Table 5.

	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
the fine-tune alone: mean core / frozen backbone	0.994	0.966	0.968	0.980	0.991
one draw: one member / mean core	1.209	1.226	1.188	1.143	1.102
the averaging: M=32 ensemble / one member	0.782	0.724	0.711	0.705	0.696
product of the three stages	0.941	0.858	0.817	0.790	0.760
the retrofit: M=32 ensemble / frozen backbone	0.938	0.854	0.815	0.787	0.758
fair CRPS (= the MAE at M=1)					
the fine-tune alone: mean core / frozen backbone	0.972	0.927	0.929	0.948	0.962
one draw: one member / mean core	1.254	1.300	1.261	1.206	1.151
the averaging: M=32 ensemble / one member	0.558	0.504	0.491	0.486	0.480
product of the three stages	0.680	0.608	0.576	0.555	0.531
the retrofit: M=32 ensemble / frozen backbone	0.677	0.602	0.572	0.552	0.528

The static part alone. The mean core is a deterministic forecast in its own right, and the largest paired sample in the paper (1421 initialisations, Table 5b) scores it against the frozen model it was fitted around. It is better, at every lead from day one: 0.990 of the frozen model's RMSE at day one, 0.964 at day five and 0.987 at day ten on the cross-variable mean, with intervals that exclude one. The cross-variable mean hides which fields the fine-tune helps and which it hurts. It helps the mass field and the surface temperature and hurts moisture: Z_{500} is at 0.790 on day one and $T_{2\mathrm{m}}$ at 0.880 on day five, against Q_{700} at 1.035 on day five and $U_{10\mathrm{m}}$ at 1.034 on day one. A fifth of the day-one Z_{500} error removed by a static fine-tune of three million parameters is consistent with a low bias of the frozen backbone in the mass field: on the campaign itself the analysis ranks above the frozen Z_{500} forecast more often than below (tilt +0.532 at day one, Sec. 3.3), and the pilot of App. E shows the same. A systematic error is the cheapest thing such a fit removes, and the effect is largest at day one and fades to 0.969 by day ten; it is also why Z_{500} alone shows a day-one gain of that size in Fig. 1.

Is the trained distribution needed? The trained amplitude is compared with rescaled ones without retraining (App. C): a fluctuation scale k multiplies the centred stochastic part of Eq. (3) while the static part is held at its trained value, so the shared fine-tune is bit-identical on every rung and only the spread moves (Table 6 and Fig. 8, App. D). At day five the trained rung is the optimum among the scales tested: shrinking the fluctuation to k=0.8 costs 1.003, to k=0.6 1.034, and inflating it to k=1.3 costs 1.025. The optimum tracks the spread–skill ratio: the rung below is marginally better at day ten (0.994). Removing the identity path of Eq. (1) at evaluation time, with the mean core kept, multiplies day-five fair CRPS by 1.239 and lowers the ratio to 0.811 (Table 6).

The sharper test replaces the generated cores by noise of the same size in the same place: on every adapted module the fluctuation $\tanh \delta_j(z) - E_j$ becomes Gaussian noise of the same per-module RMS, the static part and the templates unchanged (App. C). Scored on 1460 initialisations spread over the year (3 Jan–4 Dec 2024; Table 7, App. D), the random cores give a fraction of the trained spread, a spread–skill ratio of 0.187 at day one and 0.406 at day five against 0.849 and 1.019 for the trained model on the initialisations the two share, and a fair CRPS 1.250 times the trained model's at day one, 1.236 at day five and 1.131 at day ten. The frozen model damps random directions and amplifies the generated ones; that is what the training buys.

Evaluation-time checks. Rescoring the same 32 forecasts on random member subsets (Table 6b, App. D): fair CRPS is flat within a percent from two members up (0.992 at M=2, day five); the RMSE of the ensemble mean falls from 1.398 of the full ensemble's at one member through 1.049

at eight to 1.017 at sixteen, as $\mathrm{RMSE}^2(M) = \mathrm{RMSE}^2(\infty) + \mathrm{spread}^2/M$ says it should, with no parameter fitted to the points it is judged against (Fig. 10, App. D); M_{test} is free to choose.

The short-lead deficit, and the channel that closes it What remains is at the shortest leads, and we can say why. Every member of a model-error ensemble is identical at $t = 0$: spread has to be manufactured by integrating the weight perturbation forward, and how fast it reaches a given field is set by how fast the frozen model's response to a perturbation of its processor reaches that field. We measured that response. Drawing members in antithetic pairs $\pm z$ and correlating the two deviations from a common static reference (App. G, Table 19) measures how far the trunk's response is still linear in its own perturbation: strongest at six hours, a quarter weaker by day two, and by day five a property of the field, the mass and thermodynamic fields retaining roughly half of their antisymmetry and the wind components essentially none. The same ordering is in the calibration row of Fig. 1c: at day one HyperENS+ is over-dispersed on Z_{500} and under-dispersed on everything else, the winds and the boundary layer last, in the order in which the trunk's response reaches them. This ordering is consistent with the reading that the deficit is the signature of the perturbation class rather than of the parameterisation; it does not establish it. The full-dimensional weight noise of Xiong et al. (2025) shows the same deficit. The remedy, then, is to let members differ before the processor has acted. Adapting the encoder's 3 learned query latents, 4,120 parameters, lets them differ from the first step. On the same backbone and the same 418 initialisations of a separate campaign (27 May–21 Dec 2024), the arms differ in the adapter and in the data and gradient window of their fine-tune, so the contrast does not isolate the encoder channel: day-one SSR rises from 0.562 to 0.830, the deep U of Fig. 5 becomes a shallow one and the low bias goes with it, and fair CRPS improves at every lead, most at six hours (0.931) and still at day ten (0.976; Table 8, App. D). In the development-configuration probe (App. F.2), random perturbations of the query latents generate spread with little change in the spectral diagnostic; at similar day-five spread–skill ratios the trained processor perturbation scores about ten to thirteen percent lower on the four normalised scores of Table 18, but the two configurations differ in both channel and training, so neither contribution is isolated. What is left at six hours (0.559 on that campaign) is the deficit the class implies; perturbed inputs at training time or perturbed initial conditions at inference (App. E) are the open remedies.

5 Discussion

We asked how much ensemble skill can be obtained from a frozen deterministic weather model. On 178 shared 2024 initialisations, HyperENS+ reduces the ensemble-mean RMSE by 18.8% and 24.7% relative to the frozen model at days five and ten, and its fair CRPS is 0.569 and 0.525 of the frozen model's MAE. That fair CRPS is 3.7%, 9.6% and 6.8% higher than that of Aurora 1.5's whole-model ensemble fine-tune at days one, five and ten. The mean spread–skill ratio is within six percent of one from day three, with residual errors in the rank histograms. A separate decomposition on an earlier checkpoint and 418 common initialisations shows larger improvements for the full ensemble than for the mean-core model, without isolating the nonlinear effects on the mean forecast.

Limitations. One backbone scored to the end, one year. The retrofit has the higher fair CRPS of the two systems, and spread–skill consistency is a cross-variable statement ($T_{2\mathrm{m}}$ over-dispersed by day ten, boundary layer and winds under-dispersed through day one, extreme rank bins over-populated at every lead). A gain over a deterministic forecast is easier to realise from a weaker one, and the deterministic Aurora 1.5, the better model, was not retrofitted (Sec. 4). The decomposition's campaign starts on 27 May and is not an annual sample. The random-core baseline covers 1460 initialisations; an ensemble of the frozen Aurora with perturbed initial conditions exists only as a two-initialisation pilot (App. E), and the random-direction probe only in the development configuration (App. F.2). The second backbone is scored on half of 2019 against our re-implementation of its deterministic model, not the published one.

Implications. For an operational service, the retrofit adds an ensemble without changing the deterministic model on which post-processing and products are built, and serves both from one set of weights and a small adapter. For the science of ensembles the result is a measurement: a three-million-parameter neighbourhood of a trained deterministic solution holds an ensemble whose spread matches its error on average from day three.

AI USE STATEMENT

In this work, we used generative AI tools for implementing analysis and plotting scripts on top of our in-house training library, for preparing the standalone code release, and for editing and restructuring the manuscript, including the abstract and the appendices, from the authors' text, outlines and results; the authors revised every section. We also used them to propose hypotheses and experiments; those we tested performed worse than our own designs, and none is part of the final method. Additionally, we used generative AI tools for literature search, discussion of the methodology and arguments, and a mock review of the draft. The code release was tested end to end on a fresh machine following its README; the numbers of the main text and of the campaign appendix are produced by generator scripts from the evaluation artifacts; the references were checked against their DOIs or arXiv records by the authors. We take responsibility for the final content of this work, including text, claims or artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work trains and evaluates weather forecasting models on ERA5 and on perturbations taken from the publicly archived ECMWF ensemble; it involves no human subjects, no personal data and no new data release. Forecasts are verified against reanalysis and are not used operationally, and the method adds no capability beyond what the underlying deterministic model already provides. The authors declare no conflicts of interest beyond their affiliation.

REPRODUCIBILITY STATEMENT

The architecture and objective are specified in Sec. 3 and Apps. A and B, the evaluation conventions in Sec. 3.3, and every table and figure states the sample it was computed on. Aurora is used frozen from its public release and ERA5 is public. The evaluation code and the trained adapters of both arms are provided as supplementary material; its README documents the data access and the Pangu re-implementation.

REFERENCES

Ferran Alet, Ilan Price, Andrew El-Kadi, Dominic Masters, Stratis Markou, Tom R. Andersson, Jacklynn Stott, Remi Lam, Matthew Willson, Alvaro Sanchez-Gonzalez, and Peter Battaglia. Skillful joint probabilistic weather forecasting from marginals, 2025. URL https://arxiv.org/abs/2506.10772.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015. doi: 10.1038/nature14956. URL https://www.nature.com/articles/nature14956.

Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. LoRA-XS: Low-rank adaptation with extremely small number of parameters, 2025. URL https://arxiv.org/abs/2405.17604.

J. Berner, G. J. Shutts, M. Leutbecher, and T. N. Palmer. A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *Journal of the Atmospheric Sciences*, 66(3):603–626, 2009. doi: 10.1175/2008JAS2677.1. URL https://journals.ametsoc.org/view/journals/atsc/66/3/2008jas2677.1.xml.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3. URL https://doi.org/10.1038/s41586-023-06185-3.

Cristian Bodnar, Wessel P. Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A. Weyn, Haiyu Dong, Jayesh K. Gupta, Kit Thambiratnam, Alexander T. Archibald, Chun-Chieh Wu, Elizabeth Heider, Max Welling, Richard E. Turner, and Paris Perdikaris. A foundation model for the Earth system. *Nature*,

641(8065):1180–1187, May 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09005-y. URL https://doi.org/10.1038/s41586-025-09005-y.

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

Massimo Bonavita, Elias Hólm, Lars Isaksen, and Mike Fisher. The evolution of the ECMWF hybrid data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 142(694):287–303, 2016. doi: 10.1002/qj.2652. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2652.

Boris Bonev, Thorsten Kurth, Ankur Mahesh, Mauro Bisson, Jean Kossaifi, Karthik Kashinath, Anima Anandkumar, William D. Collins, Michael S. Pritchard, and Alexander Keller. FourCastNet 3: A geometric approach to probabilistic machine-learning weather forecasting at scale, 2025. URL https://arxiv.org/abs/2507.12144.

Philippe Bougeault, Zoltan Toth, Craig Bishop, Barbara Brown, David Burridge, De Hui Chen, Beth Ebert, Manuel Fuentes, Thomas M. Hamill, Ken Mylne, Jean Nicolau, Tiziana Paccagnella, Young-Youn Park, David Parsons, Baudouin Raoult, Doug Schuster, Pedro Silva Dias, Richard Swinbank, Yoshiaki Takeuchi, Warren Tennant, Laurence Wilson, and Steve Worley. The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, 91(8):1059–1072, 2010. doi: 10.1175/2010BAMS2853.1.

R. Buizza, M. Miller, and T. N. Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908, 1999. doi: 10.1002/qj.49712556006. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712556006.

Christopher Bülte, Nina Horat, Julian Quinting, and Sebastian Lerch. Uncertainty quantification for data-driven weather models. *Artificial Intelligence for the Earth Systems*, 5(1):240049, 2026. doi: 10.1175/AIES-D-24-0049.1. URL https://journals.ametsoc.org/view/journals/aies/5/1/AIES-D-24-0049.1.xml.

Salva Rühling Cachay, Duncan Watson-Parris, and Rose Yu. U-Cast: A surprisingly simple and efficient frontier probabilistic AI weather forecaster. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2604.09041.

Cristiana Diaconu, Miles Cranmer, Richard E. Turner, Tanya Marwah, and Payel Mukhopadhyay. Probabilistic retrofitting of learned simulators, 2026. URL https://arxiv.org/abs/2603.01949.

Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2782–2792. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/dusenberry20a.html.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Vincent Fortin, Mabrouk Abaza, François Anctil, and Richard Turcotte. Why should ensemble spread match the RMSE of the ensemble mean? *Journal of Hydrometeorology*, 15(4):1708–1713, 2014. doi: 10.1175/JHM-D-14-0008.1.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

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.

David Ha, Andrew Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.09106.

Thomas Haiden, Martin Janousek, Frédéric Vitart, Fernando Prates, Michael Maier-Gerber, Cathy Wing Yi Li, and Matthieu Chevallier. Evaluation of ECMWF forecasts. Technical Report 931, ECMWF, Reading, 09/2025 2025. URL https://www.ecmwf.int/en/elibrary/81680-evaluation-ecmwf-forecasts.

Thomas M. Hamill. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, 129(3):550–560, 2001. doi: 10.1175/1520-0493(2001)129⟨0550:IORHFV⟩2.0.CO;2.

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. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Anthony Hollingsworth. An experiment in Monte Carlo forecasting. In *ECMWF Workshop on Stochastic Dynamic Forecasting, 17-19 October 1979*, pp. 65–84, Shinfield Park, Reading, 1979. European Centre for Medium-Range Weather Forecasts.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and João Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664, 2021.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11454.

David Krueger, Chin-Wei Huang, Riashat Islam, Ryan Turner, Alexandre Lacoste, and Aaron Courville. Bayesian hypernetworks, 2018. URL https://arxiv.org/abs/1710.04759.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. AIFS-CRPS: ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1):18, Feb 2026. ISSN 3005-1460. doi: 10.1038/s44387-026-00073-7. URL https://doi.org/10.1038/s44387-026-00073-7.

Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024. doi: 10.1126/sciadv.adk4489. URL https://www.science.org/doi/abs/10.1126/sciadv.adk4489.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Edward N. Lorenz. Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi: 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/20/2/1520-0469_1963_020_0130_dnf_2_0_co_2.xml.

Christos Louizos and Max Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2218–2227. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/louizos17a.html.

Rabeeh Karimi Mahabadi, Sebastian Ruder, Mostafa Dehghani, and James Henderson. Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks, 2021. URL https://arxiv.org/abs/2106.04489.

Dominik J. Mühlematter, Michelle Halbheer, Alexander Becker, Dominik Narnhofer, Helge Aasen, Konrad Schindler, and Mehmet Ozgur Turkoglu. LoRA-Ensemble: Efficient uncertainty modelling for self-attention networks. *Transactions on Machine Learning Research*, 2026. URL https://arxiv.org/abs/2405.14438.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. FourCastNet: A global data-driven high-resolution weather model using adaptive fourier neural operators, 2022. URL https://arxiv.org/abs/2202.11214.

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, Jan 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019.

Stephan Rasp, Boris Babenko, Dominic Masters, Andrew El-Kadi, Samier Merchant, Guy Shalev, Ilan Price, Fred Zyda, Remi Lam, Sasha Shysheya, Matthew Willson, Stratis Markou, Shreya Agrawal, Suhani Vora, Mohammed Alewi Hassen, Sunny Mak, Tom R. Andersson, Megan Bela, Akib Uddin, Nofar Peled Levi, Ben Gaiarin, Ferran Alet, Aaron Bell, Peter Battaglia, and Alvaro Sanchez-Gonzalez. WeatherNext 3: Increasing resolution and performance of global weather models with raw observations, 2026. URL https://arxiv.org/abs/2609.03582.

Neale Ratzlaff and Li Fuxin. HyperGAN: A generative model for diverse, performant neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5361–5369. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ratzlaff19a.html.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. HyperDreamBooth: HyperNetworks for fast personalization of text-to-image models, 2024. URL https://arxiv.org/abs/2307.06949.

Colin Samplawski, Adam D. Cobb, Manoj Acharya, Ramneet Kaur, and Susmit Jha. Scalable Bayesian low-rank adaptation of large language models via stochastic variational subspace inference. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, 2025. URL https://arxiv.org/abs/2506.21408.

Sebastian Scher and Gabriele Messori. Ensemble methods for neural network-based weather forecasts. *Journal of Advances in Modeling Earth Systems*, 13(2), 2021. doi: 10.1029/2020MS002331. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020MS002331.

Mehmet Ozgur Turkoglu, Dominik J. Mühlematter, Alexander Becker, Konrad Schindler, and Helge Aasen. Quantifying the uncertainty of foundation models with singular value ensembles, 2026. URL https://arxiv.org/abs/2601.22068. version 2, 29 May 2026; ICML 2026.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. BLoB: Bayesian low-rank adaptation by backpropagation for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.11675.

Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: An alternative approach to efficient ensemble and lifelong learning. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.06715.

Jonathan A. Weyn, Zekun Ni, Amit Misra, Will Fein, Haiyu Dong, Wessel P. Bruinsma, Richard E. Turner, Matt Corey, Kit Thambiratnam, Kevin White, Kenji Takeda, and Hongyu Sun. Aurora 1.5: Fine-tuning a foundation model for medium-range ensemble weather prediction. Microsoft Research, July 2026. URL https://www.microsoft.com/en-us/research/publication/aurora-1-5-fine-tuning-a-foundation-model-for-medium-range-ensemble-weather-prediction/.

Xinlei Xiong, Wenbo Hu, Shuxun Zhou, Kaifeng Bi, Lingxi Xie, Ying Liu, Richang Hong, and Qi Tian. Bridging the gap between Bayesian deep learning and ensemble weather forecasts, 2025. URL https://arxiv.org/abs/2511.14218v1. version 1, 18 November 2025.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian low-rank adaptation for large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.13111.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2303.10512.

Xiaohui Zhong, Lei Chen, Hao Li, Roberto Buizza, Jun Liu, Jie Feng, Zijian Zhu, Xu Fan, Kan Dai, Jing jia Luo, Jie Wu, and Bo Lu. FuXi-ENS: A machine learning model for efficient and accurate ensemble weather prediction. *Science Advances*, 11(44):eadu2854, 2025. doi: 10.1126/sciadv.adu2854. URL https://www.science.org/doi/abs/10.1126/sciadv.adu2854.

A TRAINING OBJECTIVE

Let $X_1, \dots, X_M$ be the members of the training ensemble at one grid point of one variable, y the ERA5 value there, and $\langle\cdot\rangle$ the latitude-weighted mean over the grid. The score the trainer implements is the almost-fair CRPS of Lang et al. (2026),

$$\mathcal{L}_\alpha(X, y) \;=\; \Big\langle \frac{1}{M}\sum_i |X_i - y| \;-\; \frac{\alpha}{2}\,\frac{1}{M(M-1)}\sum_{i\neq j}|X_i - X_j| \;-\; (1-\alpha)\,\frac{1}{2M^2}\sum_{i,j}|X_i - X_j|\Big\rangle, \tag{5}$$

with $\alpha = 0.95$. At $\alpha = 1$ the dispersion term is the unbiased estimator of $\frac{1}{2}\,\mathbb{E}|X - X'|$ over independent members and the score is the fair CRPS, the score every table of the paper reports; at $\alpha = 0$ it is the plain ensemble CRPS, whose dispersion term is the same sum over all ordered pairs including $i = j$ and is therefore low by the factor $(M-1)/M$. The value $\alpha = 0.95$ is the compromise of Lang et al. (2026) between the two. The member-pair sums are evaluated without materialising the $M \times M$ difference tensor, which is what lets a multi-step roll-out at 0.25° fit in memory.

Spatial scales. The score is evaluated at three resolutions: on the fields as they are, and after average-pooling members, target and latitude weights over 4×4 and 16×16 grid boxes, about 1° and 4°. The three values are averaged with equal weight,

$$\mathcal{L}^{\mathrm{ms}} \;=\; \frac{1}{3}\sum_{s\in\{1,4,16\}} \mathcal{L}_\alpha\big(\mathrm{pool}_s X,\; \mathrm{pool}_s y\big), \tag{6}$$

so that the members' agreement with the analysis and their spread are scored at the large scales as well as on the full field.

Spectral term. For each member and for the target, the field is multiplied by $\sqrt{\cos\varphi}$ and transformed with a two-dimensional FFT; the power is averaged over 32 linearly spaced annuli in total wavenumber, giving a 32-vector $P(X_i)$ of radially averaged power. The spectral term is Eq. (5) applied to these vectors, $\mathcal{L}^{\mathrm{sp}} = \mathcal{L}_\alpha(P(X), P(y))$, with the mean over annuli in place of the grid mean and no further latitude weighting. It rewards members whose power is distributed over scales as the analysis's is and whose spectra disperse as those of independent draws would; it is the term that keeps the member spectra of Fig. 6 on the backbone's own curve.

Variables and weights. The nine training variables are the four surface fields ($T_{2\mathrm{m}}$, $U_{10\mathrm{m}}$, $V_{10\mathrm{m}}$, MSLP) and the five atmospheric fields (Z, T, U, V, Q) on all thirteen pressure levels, an atmospheric term being the mean over its levels. The objective is

$$\mathcal{L} \;=\; \sum_v w_v\Big(\lambda_v^{\mathrm{ms}}\,\mathcal{L}_v^{\mathrm{ms}} + \lambda_v^{\mathrm{sp}}\,\mathcal{L}_v^{\mathrm{sp}}\Big), \tag{7}$$

where w_v is one for an atmospheric and a quarter for a surface variable, and the per-term weights λ_v are pinned so that each of the eighteen terms contributes unit magnitude to the objective at the start of training, then held fixed. Pinning is what lets Z in $\mathrm{m^2\,s^{-2}}$ and Q in $\mathrm{kg\,kg^{-1}}$ share one sum; the ratio of the two terms within a variable is set by the same rule, not by hand.

Roll-out. Training rolls the ensemble out over four autoregressive six-hour steps from one initial condition, every member keeping its own $\Delta W(z_m)$ along the trajectory, and evaluates Eq. (7) at the first and the fourth step, leads six and 24 hours; both gradients flow through the whole roll-out. With one member per device (App. B) the pair sums of Eq. (5) run over the $M_{\mathrm{train}} = 8$ members of one ensemble group, 28 unordered pairs per grid point, which is the sense in which M_{train} is a parameter of the gradient estimator (Sec. 3.2).

B TRAINING SYSTEM

Two constraints shape the training system: a fair-CRPS objective needs all M members of an ensemble inside one loss, and a 0.25° backbone of 1.3 B parameters does not leave room on one 80 GB device for eight of its roll-outs. The layout below resolves both, and any retrofit of this kind on a foundation-scale backbone will need something like it. Table 2 gives the trainable set of each arm.

Table 2: The trainable set of each arm, counted from the trained overlays. The frozen Aurora backbone has 1.3 B parameters and none of them is touched. Both arms share the $J = 96$ backbone modules, their rank-8 templates and one hypernetwork; HyperENS+ adds one module, the encoder's query latents, and with it one more row in each of the three embedding tables.

	HyperENS	HyperENS+
adapted modules	96	97
templates $\{A_j, B_j\}$	2,850,816	2,850,816
query-latent adapter	–	4,120
hypernetwork ϕ (embeddings, trunk, head)	380,224	380,608
trainable parameters	3,231,040	3,235,544
share of the frozen model	0.26%	0.26%

One member per device. The world of N devices is split into ensemble groups of M_{train} consecutive ranks, and each rank is one member. At every step all ranks of a group load the same initial condition; rank m draws its own latent z_m from a generator seeded by the sample index and its position in the group, so a run is reproducible and no two members share a draw, and the hypernetwork on that rank generates the cores $\delta_j(z_m)$ once, before the backbone's forward pass. The backbone then runs with member m's $\Delta W(z_m)$ in every adapted module, for all steps of the roll-out. The batch of one ensemble is thus one sample per rank; with $N > M_{\text{train}}$ several ensembles train in parallel as independent groups, and their gradients are averaged by the sharded reduction below.

The differentiable gather. After the roll-out each rank holds its member's forecast. The member forecasts are gathered within the group by an all-gather with a defined backward pass (`torch.distributed.nn.all_gather`), so that the ensemble loss, evaluated identically on every rank of the group, backpropagates through the gathered tensor into each rank's own forecast and from there into that rank's hypernetwork and templates. No member's gradient is truncated: the pair terms of Eq. (5) reach both members of every pair. The peak memory of a step is set at this point, by the gathered M-member forecast and its gradient.

Sharding the frozen backbone. The whole model, frozen backbone and trainable adapters alike, is wrapped in fully sharded data parallelism (FSDP, full sharding) across all N ranks, with the backbone's blocks as shard units, the sharded parameters and optimiser state offloaded to host memory between uses, and backward prefetching. Each rank therefore holds $1/N$ of the frozen weights at rest and receives a block's full weights only for the duration of its forward and backward, which is what makes M simultaneous roll-outs of a 1.3 B model fit beside their activations. Only the adapter parameters carry gradients; the trainable set is asserted against the state dictionary at start-up and checkpointed separately from the backbone, so a checkpoint of the retrofit is tens of megabytes and the backbone file is never rewritten.

Activation checkpointing. The backbone's blocks are wrapped in non-reentrant activation checkpointing, and for multi-step roll-outs the whole layer stack in one outer checkpoint region, so that a four-step roll-out with the loss at steps one and four stores one step's activations at a time and recomputes the rest in the backward pass. The hypernetwork runs above the checkpointed region, once per forward pass; the cores it emits are held for the whole roll-out and are never recomputed.

Where the adapters sit. The backbone adapters are not hooks: each adapted projection of every Swin attention block owns a small adapter module whose output is added to the frozen projection's, $W_j x + B_j\, C_j(z)\, A_j\, x$, the core selected by the module's index into the (J, r, r) tensor the hypernetwork emitted. Given no core the same module is a static low-rank adapter, which is how the identity check of Sec. 3.1 and the mean-core runs of Sec. 4 execute on one code path. The query-latent adapter of HyperENS+ is the one exception: the encoder's query latents are a parameter of the frozen model, not an activation, so their perturbation is applied by a forward pre-hook on the input of the level-aggregation cross-attention, split into two hooks — the hypernetwork call on the encoder, outside the checkpointed region, and the addition inside it — so that it survives the recomputation activation checkpointing performs. The frozen model's own code is not modified in either case.

Precision and optimiser. Training runs under mixed-precision autocast; AdamW, a cosine schedule restarted on every three-month data chunk, and a global gradient-norm clip that binds only as a divergence guard. The bound of Eq. (1) is what makes low-precision training of the generator safe: a tail draw of z cannot produce an unbounded weight update.

C THE EVALUATION-TIME ABLATION FAMILY

The diagnostic experiments of Sec. 4 use a family of inference-time modifications of the trained core, obtained without any retraining:

$$C_j(z;\, \iota, k) \;=\; \iota\, I_r \;+\; s\, E_j \;+\; s\, k\big(\tanh \delta_j(z) \;-\; E_j\big), \qquad \iota \in \{0,1\},\; k \geq 0, \tag{8}$$

where ι toggles the identity path and k rescales the fluctuation about the mean core, which is kept. The pair $(\iota, k) = (1, 1)$ is the trained model, bit for bit. The mean cores E_j are estimated once from 2048 fixed-seed draws and are bit-identical on every device; a device-dependent estimate would shift each member by a different constant and manufacture spurious spread. Three members of the family do the diagnostic work:

- *Spread ladder* $(\iota, k) = (1, k)$, k varied: the static part of Eq. (3), $B_j(I_r + sE_j)A_j$, is the same on every rung *independently of* k. The ladder therefore varies ensemble spread while holding the shared fine-tune exactly fixed, which is what permits a clean test of whether spread is coupled to the fine-tune (Sec. 4). Rescaling s instead would scale the mean core sE_j together with the fluctuation and confound the two.
- *Static model* $(\iota, k) = (1, 0)$, $C_j = I_r + sE_j$: every member identical, spread exactly zero; isolates the shared fine-tune.
- *Identity off* $(\iota, k) = (0, 1)$: the identity path removed, the mean core and the fluctuation kept.
- *Random core* $(\iota, k) = (1, 1)$ with $\tanh \delta_j(z) - E_j$ replaced by $\tanh(\tau_j \varepsilon_j)$, $\varepsilon_j \sim \mathcal{N}(0, I)$ i.i.d. per member and τ_j the per-module RMS of the trained fluctuation over the same 2048 draws: the same static part, templates and magnitude, a random direction.
- *Single member*, $M_{\text{test}} = 1$: isolates the cost of one stochastic draw.

Together with the full ensemble these variants close a multiplicative decomposition of the skill (Sec. 4).

Table 3: The evaluation campaigns of Sec. 4, all on ERA5 2024; n: initialisations scored. Systems are compared only on the initialisations they share, so the paired n of a comparison is smaller than either row.

system	M	window (2024)	n	used for
frozen backbone, $M=1$	1	1 Jan–21 Dec	1422	parent of both arms; every decomposition and ladder
Aurora 1.5 deterministic, $M=1$	1	2 Jan–21 Dec	178	parent of Aurora 1.5
Aurora 1.5 ensemble	32	1 Jan–21 Dec	711	Fig. 1, Table 4
HyperENS+ (later checkpoint)	32	2 Jan–21 Dec	178	Fig. 1, Table 4
HyperENS	32	2 Jan–21 Dec	178	Fig. 1, Table 4
HyperENS+ (earlier checkpoint)	32	27 May–21 Dec	418	decomposition, within-model contrast
HyperENS	32	1 Jan–21 Dec	712	decomposition, within-model contrast
HyperENS+ mean core, $M=1$	1	1 Jan–21 Dec	1422	decomposition
HyperENS mean core, $M=1$	1	1 Jan–21 Dec	1422	decomposition
HyperENS+ one member, $M=1$	1	1 Jan–21 Dec	1422	decomposition
HyperENS one member, $M=1$	1	1 Jan–21 Dec	1422	decomposition
HyperENS+ (earlier checkpoint), ladder reference	32	1 Jan–18 Dec	68	spread and M_{test} ladders
random core, $M=32$ (HyperENS+ static part)	32	1 Jan–31 Dec	1460	*Is the trained distribution needed?*

D CAMPAIGN FIGURES AND TABLES

Table 3 lists the campaigns of Sec. 4 with their windows and sample sizes; everything below is referenced from the main text where it is used.

D.1 GAIN OVER THE PARENT, SUMMARISED, AND THE TABLES BEHIND SEC. 4

Table 4 gives, at days one, five and ten, the spread–skill ratios behind Fig. 1c and the retrofit's fair CRPS and RMSE as ratios to Aurora 1.5, with the paired intervals on every ratio. Table 5 is the full decomposition (*Where the skill comes from*) for both arms, with the per-variable mean-core panel (*The static part alone*); Fig. 3 draws the same chain.

D.2 CALIBRATION, HISTOGRAM SHAPE AND SPECTRA AGAINST LEAD

D.3 THE SPREAD LADDER AND THE MEMBERS AT INFERENCE

Table 6(a) and Fig. 8 are the spread ladder of Sec. 4 (*Engineered, not lucky*); Table 6(b), Fig. 9 and Fig. 10 the M_{test} ladder (*Evaluation-time checks*).

D.4 THE RANDOM-CORE RUNG

Table 7 is the random-core rung of Sec. 4 (*Is the trained distribution needed?*): the generated cores replaced by Gaussian noise of the same per-module RMS, scored on 1460 initialisations spread over 2024 and paired with the mean core, the frozen backbone and, on the 8 initialisations they share, the trained HyperENS+. The random members are damped rather than wrong: their spectral ratios sit on the mean core's curve and their rank histograms show little tilt, but the spread is a fraction of the trained model's at every lead and the fair CRPS is a quarter higher through day five.

D.5 THE WITHIN-MODEL CONTRAST

Table 8 tabulates the HyperENS+ against HyperENS contrast that Sec. 4 (*The short-lead deficit*) quotes: spread–skill ratios of both arms and the paired fair-CRPS and RMSE ratios on the initialisations the two campaigns share.

D.6 PER-VARIABLE SCORECARDS

Tables 9, 10 and 11 give the per-variable values behind the campaign, each system on all of its own initialisations; the HyperENS+ scorecard is the campaign of the earlier checkpoint (Sec. 4), not the campaign of Table 4. The tables also report a rank-histogram reliability index, $r_i = \sum_{b=1}^{M+1} \left| p_b -$

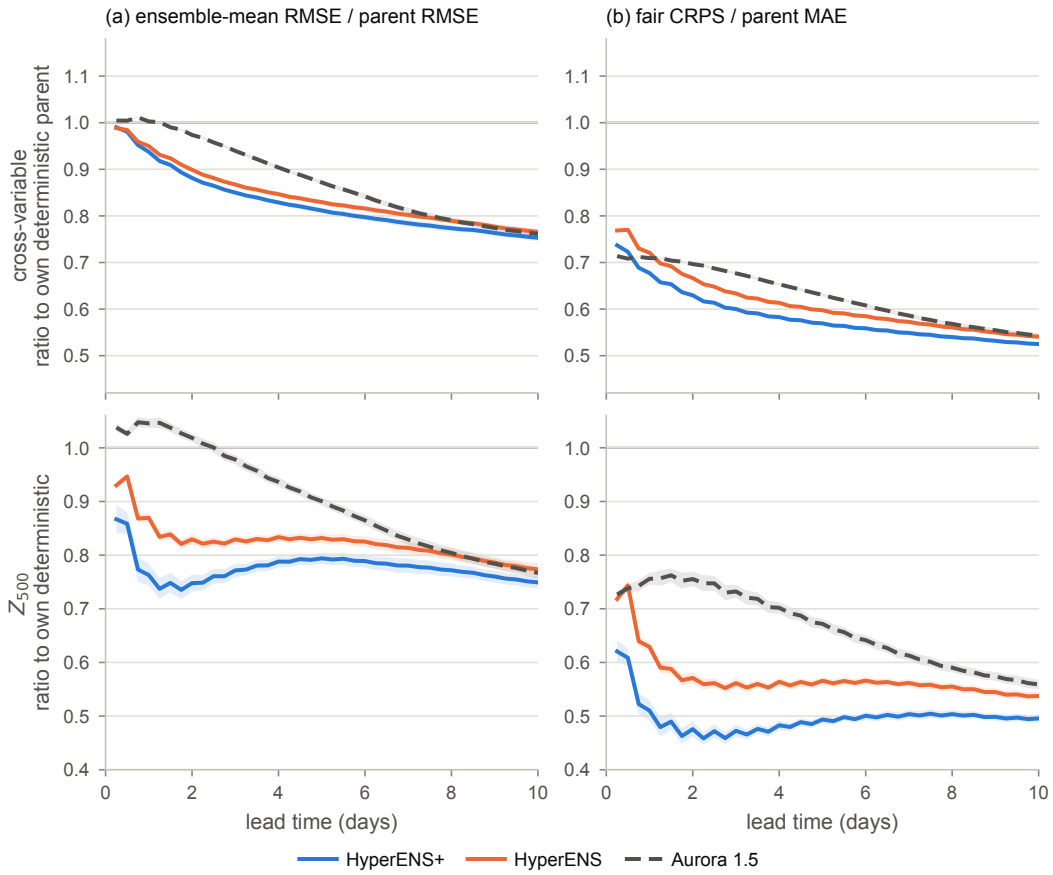

Figure 2: Fig. 1(a, b) summarised: (a) RMSE of the ensemble mean over the parent's RMSE and (b) fair CRPS over the parent's MAE, against lead, as the cross-variable mean of per-variable ratios over the 8 scored variables (top) and on Z_{500} (bottom), for the systems and pairing of Fig. 1.

$\left|\frac{1}{M+1}\right|$ with p_b the observed frequency of rank b, zero for a flat histogram. It is a per-variable absolute and it is unsigned, so it cannot tell under-dispersion from over-dispersion or from bias; the sign of the departure from flatness is the tilt of Sec. 3.3 and the histograms themselves, Figs. 5 and 7. For the frozen backbone, whose two-bin histogram makes the tilt the excess of analyses above the forecast over those below, the Z_{500} tilt on all 1422 initialisations is +0.284 at six hours, +0.532 at day one, +0.560 at day five and +0.471 at day ten, the low bias that *The static part alone* quotes.

D.7 The second backbone

The second backbone is our own PyTorch re-implementation of Pangu-Weather (Bi et al., 2023), a 3D Earth-specific transformer of 64 million parameters, a twentieth of Aurora's, which takes a single atmospheric state as input and is trained by us to the 6-h step only, so that a ten-day forecast is forty autoregressive steps of one model rather than the published hierarchy of 24-, 6-, 3- and 1-h models. The retrofit is HyperENS as in Sec. 3: rank-8 templates on the query–key–value and output projections of every attention block, encoder and decoder included, one hypernetwork with the factored module embedding, the same objective, and no perception channel, since the model has no query latents. Both the deterministic model and the $M{=}32$ retrofit are scored on 154 daily initialisations of 2019 (2 January–4 June) to day ten with the metric handlers of Sec. 3.3; the deterministic model is the retrofit's own frozen backbone, so the ratios of Fig. 11 are of the same kind as those of Fig. 1.

Two things carry over from Aurora and one does not. By day ten the retrofit reaches the same ratios as on Aurora, about 0.8 of the deterministic RMSE and 0.55 of its MAE, on every variable shown. The short-lead deficit is larger: the ensemble mean is worse than the deterministic forecast through

day two on Z_{500} and through day one on the winds, and the fair-CRPS ratio at day one is about 0.85 on Z_{500} against 0.68 on Aurora, the signature of a mass field whose members are over-dispersed before the mean has settled. And the absolute skill is that of the backbone (Table 12): the retrofit's day-five Z_{500} fair CRPS is a quarter higher than that of HyperENS on Aurora (Table 10), because a re-implementation trained to a single step from a single input time is a weaker forecaster than Aurora, and a weaker backbone yields a weaker ensemble however well its spread is placed. The markers in Fig. 11 are read off the figures of Xiong et al. (2025), whose ensemble is a whole-model variational fine-tune with a learned input perturbation, as ratios to their own deterministic model; they are ahead of the retrofit through day five and level with it by day ten, with a read-off error that dominates the day-one bars.

Table 4: Headline comparison against Aurora 1.5 on the initialisation grid the two systems share (HyperENS+: n=178, 2 Jan–21 Dec 2024; HyperENS: n=178, 2 Jan–21 Dec 2024). Upper panel: calibration, the spread-skill ratio of each system against its own error. Lower panel: skill, each arm over Aurora 1.5. SSR is the $\sqrt{(M+1)/M}$-corrected RMS spread over the RMSE of the ensemble mean, averaged over initialisations as a per-date ratio (not a ratio of aggregates); it is a within-model absolute and is never ratioed between systems. Skill entries are means over the eight scored variables of per-variable ratios, never ratios of native-unit sums. Rows are five of the eight scored variables; the last row of each block is the mean over all eight. Aurora 1.5 appears in both blocks of the upper panel because it is recomputed on each block's own pairing. Brackets: 95% paired moving-block bootstrap over initialisations (block ℓ=6, 2000 draws), dates resampled once and both arms recomputed on the resample.

	retrofit			Aurora 1.5		
	1 d	5 d	10 d	1 d	5 d	10 d
HyperENS+ against Aurora 1.5, n=178						
Z_{500}	1.108	1.050	1.038	0.763	0.929	0.982
T_{850}	0.820	1.065	1.077	0.976	0.935	0.970
T_{2m}	0.788	1.186	1.236	1.009	0.940	0.969
U_{10m}	0.841	0.983	1.017	1.136	1.021	1.024
Q_{700}	0.792	1.038	1.039	1.034	1.021	0.992
mean, eight variables	0.861	1.037	1.057	0.997	0.975	0.995
HyperENS against Aurora 1.5, n=178						
Z_{500}	0.696	0.818	0.927	0.763	0.929	0.982
T_{850}	0.585	0.862	0.967	0.976	0.935	0.970
T_{2m}	0.587	0.938	1.075	1.009	0.940	0.969
U_{10m}	0.549	0.812	0.940	1.136	1.021	1.024
Q_{700}	0.565	0.876	0.941	1.034	1.021	0.992
mean, eight variables	0.594	0.843	0.957	0.997	0.975	0.995

	fair CRPS			RMSE of the ensemble mean		
	1 d	5 d	10 d	1 d	5 d	10 d
HyperENS+ / Aurora 1.5, n=178						
Z_{500}	1.006 [0.980, 1.031]	1.064 [1.045, 1.081]	1.042 [1.031, 1.054]	0.995 [0.979, 1.011]	1.060 [1.050, 1.070]	1.043 [1.034, 1.053]
T_{850}	1.019 [1.015, 1.022]	1.080 [1.075, 1.086]	1.063 [1.054, 1.071]	0.965 [0.961, 0.970]	1.046 [1.041, 1.052]	1.038 [1.031, 1.046]
T_{2m}	1.069 [1.060, 1.078]	1.155 [1.148, 1.163]	1.131 [1.124, 1.138]	0.941 [0.924, 0.957]	1.057 [1.049, 1.066]	1.053 [1.045, 1.061]
U_{10m}	1.023 [1.021, 1.026]	1.084 [1.080, 1.088]	1.055 [1.048, 1.062]	0.995 [0.993, 0.997]	1.059 [1.054, 1.064]	1.037 [1.031, 1.043]
Q_{700}	1.072 [1.070, 1.074]	1.083 [1.079, 1.086]	1.067 [1.061, 1.073]	1.026 [1.024, 1.028]	1.068 [1.065, 1.071]	1.061 [1.056, 1.066]
mean, eight variables	1.037 [1.032, 1.043]	1.096 [1.091, 1.102]	1.068 [1.061, 1.074]	0.987 [0.982, 0.992]	1.059 [1.054, 1.064]	1.043 [1.038, 1.050]
HyperENS / Aurora 1.5, n=178						
Z_{500}	1.240 [1.202, 1.277]	1.219 [1.208, 1.230]	1.130 [1.119, 1.141]	1.135 [1.111, 1.158]	1.111 [1.099, 1.122]	1.077 [1.067, 1.087]
T_{850}	1.071 [1.066, 1.076]	1.132 [1.126, 1.138]	1.099 [1.091, 1.107]	0.964 [0.960, 0.967]	1.072 [1.066, 1.079]	1.059 [1.052, 1.067]
T_{2m}	1.128 [1.118, 1.139]	1.210 [1.202, 1.218]	1.161 [1.153, 1.168]	0.948 [0.931, 0.965]	1.083 [1.075, 1.091]	1.066 [1.059, 1.074]
U_{10m}	1.076 [1.073, 1.079]	1.127 [1.122, 1.132]	1.077 [1.070, 1.083]	0.992 [0.990, 0.993]	1.079 [1.073, 1.084]	1.052 [1.047, 1.059]
Q_{700}	1.119 [1.117, 1.121]	1.110 [1.106, 1.114]	1.079 [1.073, 1.085]	1.021 [1.019, 1.023]	1.072 [1.069, 1.075]	1.062 [1.057, 1.067]
mean, eight variables	1.109 [1.101, 1.118]	1.152 [1.147, 1.158]	1.100 [1.094, 1.107]	1.004 [0.998, 1.010]	1.083 [1.077, 1.088]	1.061 [1.055, 1.067]

Table 5: **(a)** Multiplicative skill decomposition, cross-variable: the frozen backbone taken to the mean core (the static path alone, $C = I + s\,E_j$), the mean core to one stochastic member at a fixed latent, and that member to the 32-member ensemble mean. The product of the three stages and the directly measured M=32/frozen ratio are both shown; they differ only because a mean of per-variable ratios does not commute with multiplication. Fair CRPS is M-corrected and comparable across ensemble sizes, the RMSE of the ensemble mean is not. Paired on the initialisations common to all four rungs (HyperENS+: n=418, 27 May–21 Dec 2024; HyperENS: n=711, 1 Jan–21 Dec 2024). **(b)** Control-run preservation per variable: the RMSE of each arm's mean core over the frozen backbone, 1 being unchanged (HyperENS+: n=1421, 1 Jan–21 Dec 2024; HyperENS: n=1421, 1 Jan–21 Dec 2024). The spread across variables is the finding and the cross-variable mean conceals it: for HyperENS+ the eight per-variable ratios span 0.790 (Z_{500}) to 1.065 (Q_{700}) at day one; for HyperENS the eight per-variable ratios span 0.879 (Z_{500}) to 1.028 (Q_{700}) at day one. Brackets: 95% paired moving-block bootstrap over initialisations (block ℓ = 7/9 for the arms of (a), 11/11 for (b); 2000 draws), dates resampled once and both arms recomputed on the resample.

	1 d	3 d	5 d	7 d	10 d
HyperENS+, RMSE of the ensemble mean, n=418					
the fine-tune alone: mean core / frozen backbone	0.994	0.966	0.968	0.980	0.991
one draw: one member / mean core	1.209	1.226	1.188	1.143	1.102
the averaging: M=32 ensemble / one member	0.782	0.724	0.711	0.705	0.696
product of the three stages	0.941	0.858	0.817	0.790	0.760
the retrofit: M=32 ensemble / frozen backbone	0.938	0.854	0.815	0.787	0.758
HyperENS+, fair CRPS (= the MAE at M=1*),* n=418					
the fine-tune alone: mean core / frozen backbone	0.972	0.927	0.929	0.948	0.962
one draw: one member / mean core	1.254	1.300	1.261	1.206	1.151
the averaging: M=32 ensemble / one member	0.558	0.504	0.491	0.486	0.480
product of the three stages	0.680	0.608	0.576	0.555	0.531
the retrofit: M=32 ensemble / frozen backbone	0.677	0.602	0.572	0.552	0.528
HyperENS, RMSE of the ensemble mean, n=711					
the fine-tune alone: mean core / frozen backbone	0.988	0.961	0.966	0.978	0.986
one draw: one member / mean core	1.097	1.119	1.102	1.079	1.056
the averaging: M=32 ensemble / one member	0.878	0.816	0.788	0.766	0.739
product of the three stages	0.952	0.877	0.839	0.808	0.770
the retrofit: M=32 ensemble / frozen backbone	0.951	0.875	0.838	0.808	0.769
HyperENS, fair CRPS (= the MAE at M=1*),* n=711					
the fine-tune alone: mean core / frozen backbone	0.970	0.930	0.936	0.951	0.964
one draw: one member / mean core	1.121	1.155	1.140	1.113	1.082
the averaging: M=32 ensemble / one member	0.667	0.597	0.568	0.546	0.521
product of the three stages	0.725	0.642	0.606	0.578	0.543
the retrofit: M=32 ensemble / frozen backbone	0.725	0.639	0.604	0.577	0.542

	6 h	1 d	5 d	10 d
HyperENS+, mean core / frozen backbone, n=1421				
Z_{500}	0.904	0.790	0.905	0.969
T_{850}	1.017	1.014	0.948	0.974
T_{2m}	0.999	0.949	0.880	0.922
U_{10m}	1.032	1.034	0.990	1.001
Q_{700}	1.030	1.065	1.035	1.037
mean of the eight variables	1.006	0.990	0.964	0.987
	[1.005, 1.008]	[0.988, 0.992]	[0.960, 0.969]	[0.982, 0.991]
HyperENS, mean core / frozen backbone, n=1421				
Z_{500}	0.973	0.879	0.922	0.973
T_{850}	1.003	0.989	0.946	0.975
T_{2m}	0.992	0.944	0.883	0.924
U_{10m}	1.017	1.018	0.987	1.001
Q_{700}	1.011	1.028	1.009	1.020
mean of the eight variables	1.004	0.986	0.962	0.985
	[1.003, 1.004]	[0.985, 0.987]	[0.958, 0.966]	[0.980, 0.989]

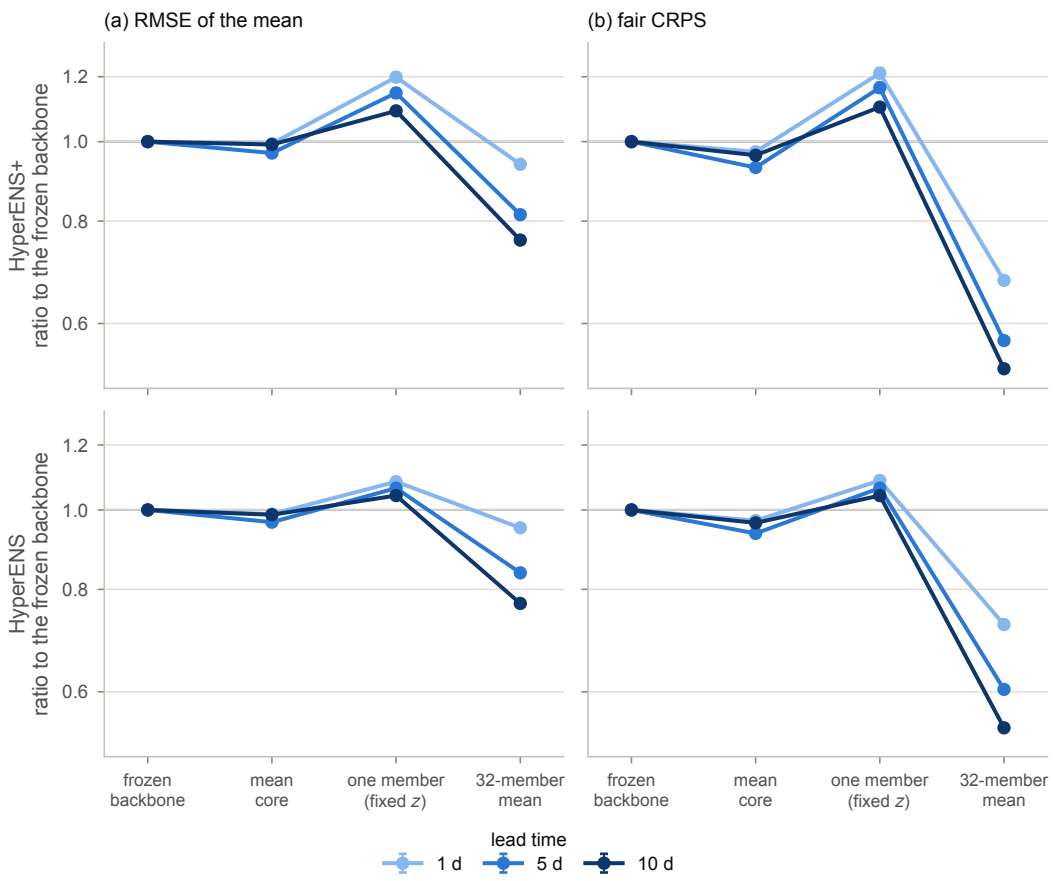

Figure 3: The decomposition as a ladder: (a) RMSE of the ensemble mean and (b) fair CRPS as ratios to the frozen backbone, at days one, five and ten, along the four rungs frozen backbone → mean core → one member at a fixed latent → 32-member mean, for HyperENS+ (top; n=418, 27 May–21 Dec 2024) and HyperENS (bottom; n=711, 1 Jan–21 Dec 2024). Logarithmic ratio axis; bars are the paired moving-block bootstrap. The single-member rung sits above one while the ensemble mean sits below it: members are degraded on purpose and the averaging buys it back.

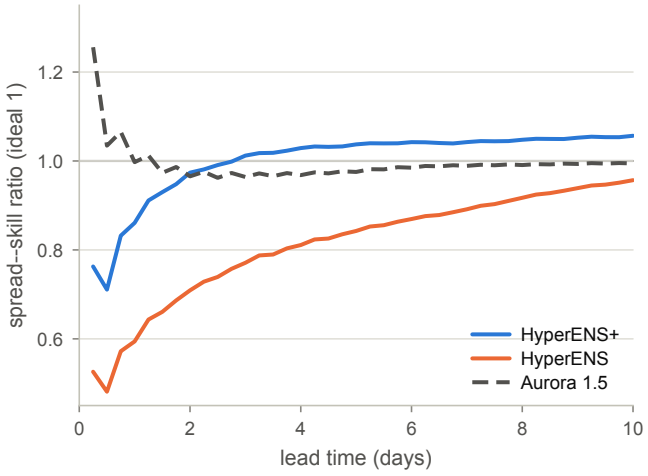

Figure 4: Spread–skill ratio against lead, 6 h to 10 d, cross-variable mean over the 8 scored variables, for HyperENS+, HyperENS and Aurora 1.5 on the 178 initialisations they share (2 Jan–21 Dec 2024), M=32 throughout, the per-date estimand of Sec. 3.3. The horizontal line is the ideal value.

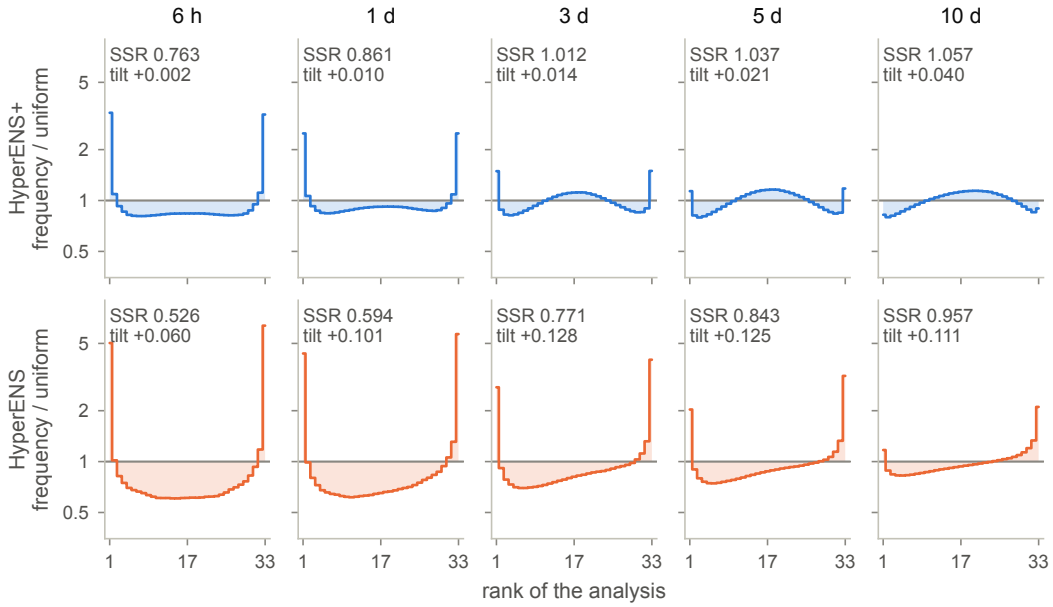

Figure 5: Rank-histogram shape against lead for HyperENS+ (top) and HyperENS (bottom) on the pairing of Fig. 4: bin counts pooled over initialisations, grid points and the 8 scored variables, relative to a flat histogram (the horizontal line), on a logarithmic axis, $M+1$ bins. Each panel is annotated with the cross-variable spread–skill ratio and the tilt of Sec. 3.3. A U is under-dispersion, a dome over-dispersion, a slope a bias; no scalar index separates the three. Aurora 1.5's campaign artifact stores no bin counts, so it cannot be drawn here; Fig. 7 draws it from the one artifact that does.

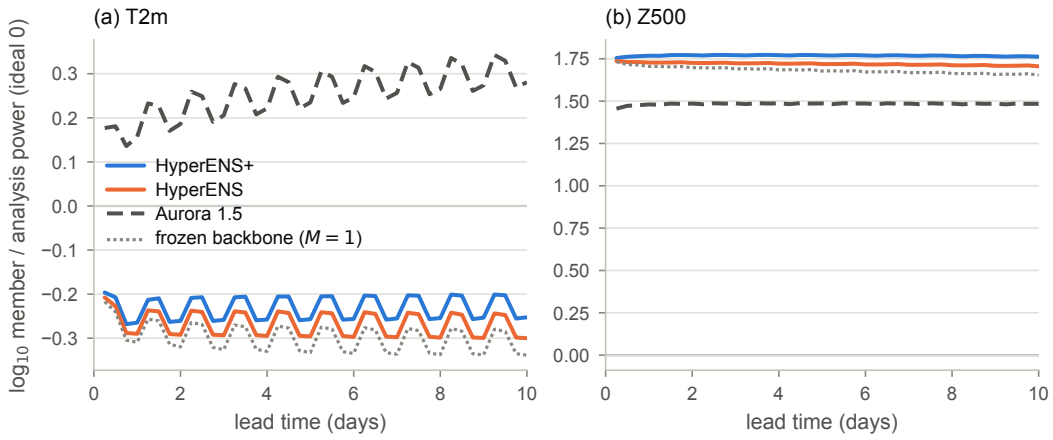

Figure 6: Spectral realism of the members: mean $\log_{10}$ ratio of member to analysis radially averaged power over the upper half of the resolved wavenumbers, for (a) $T_{2\mathrm{m}}$ and (b) Z_{500}, the two variables on which the metric is computed, for the systems and pairing of Fig. 4, with the frozen backbone's own forecast ($M=1$) as the dotted curve. The horizontal line is the ideal value; negative values are a small-scale power deficit, positive values a surplus. On Z_{500} every system, the frozen backbone and Aurora 1.5 included, carries a large lead-independent surplus: it is the backbone family's, against a geopotential spectrum that is nearly empty at these wavenumbers, and the members neither add to it nor remove it.

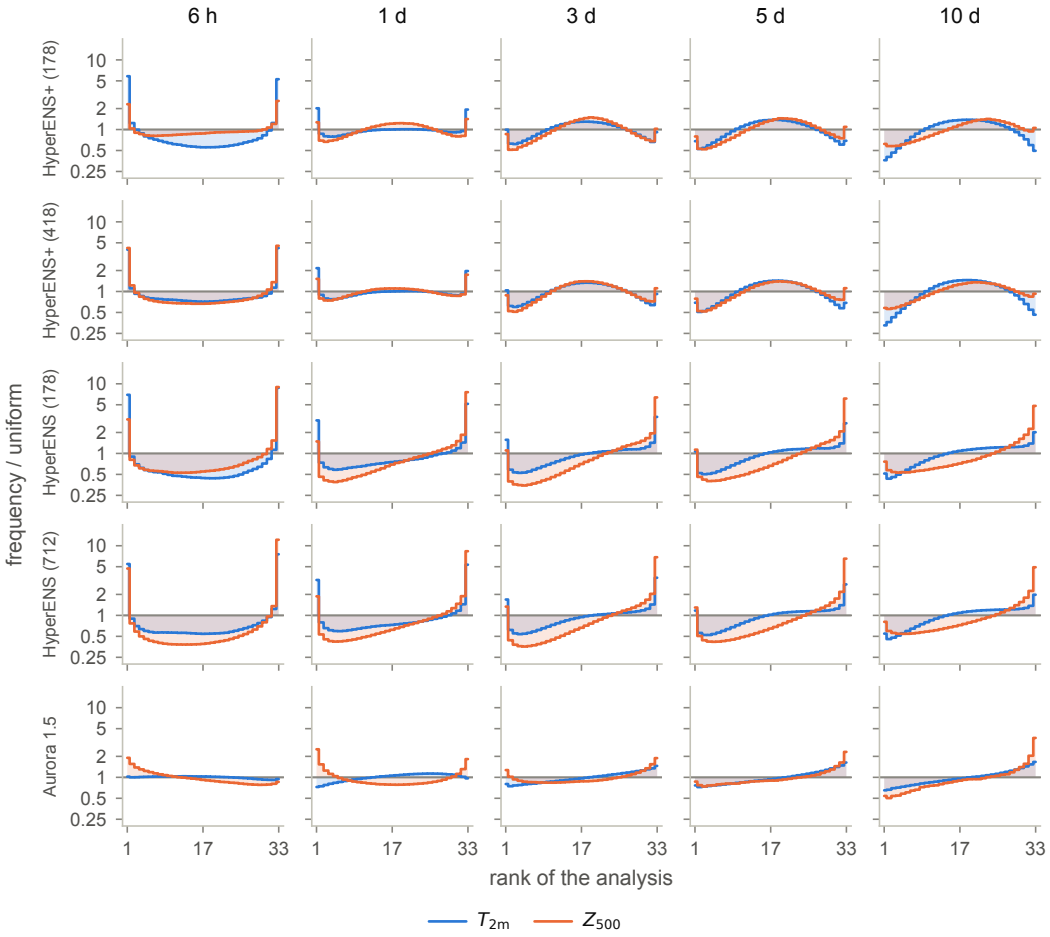

Figure 7: Average rank histograms on $T_{2\mathrm{m}}$ and Z_{500} for the plain arms and Aurora 1.5, each drawn on all of its own initialisations: HyperENS+ on the campaign of Table 4 (178 initialisations) and on the campaign of the earlier checkpoint (418, 27 May–21 Dec 2024); HyperENS on its two campaigns (178 and 712); Aurora 1.5 from the only artifact of it that stores bin counts, twelve initialisations over 1–7 January 2024. Counts pooled over initialisations and grid points, normalised against a uniform $1/(M{+}1)$ (the horizontal line), logarithmic axis. Rows are on different initialisation sets, so the six-hour column compares shapes, not values.

Table 6: Evaluation-time ladders for the HyperENS+ arm, both cross-variable. **(a)** The fluctuation scale k around the trained operating point ($C_k = I + s\,E + k\,s\,(\tanh\delta - E)$, centred with the mean kept, so the static part is held at its trained value and only the fluctuation scales), plus the identity-off rung, which removes the identity path alone ($C = s\,\tanh\delta$). Ratios are to the ladder's own k=1.0 rung. **(b)** The same 32 forecasts of the k=1.0 rung rescored on random member subsets (8 subsets per initialisation and size, seed 0; M=32 is the full ensemble and reproduces the scorecard exactly), ratios to M=32. At M=1 the fair correction degenerates and the spread is identically zero, so only the RMSE of the single member is defined; the other cells are marked $--$. Both blocks use the ladder's n=68 designed initialisations (1 Jan–18 Dec 2024): 68 distinct days spanning 12 months, separate days rather than adjacent synoptic hours. SSR is the $\sqrt{(M+1)/M}$-corrected RMS spread over the RMSE of the ensemble mean, averaged over initialisations as a per-date ratio (not a ratio of aggregates); it is a within-model absolute and is never ratioed between systems. Skill entries are means over the eight scored variables of per-variable ratios, never ratios of native-unit sums. Brackets: 95% paired moving-block bootstrap over days (block ℓ=4, 2000 draws), dates resampled once and both arms recomputed on the resample.

	SSR (ideal 1)			fair CRPS		RMSE of the mean	
	6 h	5 d	10 d	5 d	10 d	5 d	10 d
(a) spread ladder: the fluctuation scale k; ratios to the k=1.0 rung							
k=0.6	0.420	0.796	0.959	1.034 [1.028,1.039]	1.001 [0.993,1.008]	1.024 [1.019,1.028]	1.005 [0.999,1.011]
k=0.8	0.545	0.936	1.022	1.003 [1.000,1.005]	0.994 [0.990,0.997]	1.004 [1.001,1.006]	0.998 [0.995,1.001]
k=1.0 (reference rung)	0.663	1.045	1.070	1	1	1	1
k=1.3	0.837	1.165	1.118	1.025 [1.021,1.030]	1.025 [1.021,1.028]	1.016 [1.011,1.021]	1.014 [1.010,1.018]
identity off (ι=0, k=1)	0.575	0.811	0.897	1.239 [1.225,1.252]	1.184 [1.171,1.198]	1.159 [1.144,1.174]	1.117 [1.103,1.132]
(b) M_{test} ladder on the k=1.0 rung; ratios to M=32							
M=1	–	–	–	–	–	1.398 [1.376,1.420]	1.428 [1.415,1.441]
M=2	0.678	1.014	1.047	0.992 [0.977,1.009]	0.991 [0.980,1.004]	1.214 [1.202,1.227]	1.224 [1.215,1.234]
M=4	0.687	1.037	1.060	0.996 [0.989,1.005]	0.998 [0.992,1.003]	1.105 [1.099,1.112]	1.112 [1.106,1.117]
M=8	0.674	1.041	1.064	1.001 [0.999,1.004]	1.001 [0.999,1.003]	1.049 [1.046,1.052]	1.050 [1.048,1.053]
M=16	0.665	1.041	1.065	1.002 [1.000,1.003]	1.001 [1.000,1.003]	1.017 [1.016,1.019]	1.018 [1.017,1.020]
M=32	0.663	1.045	1.070	1	1	1	1

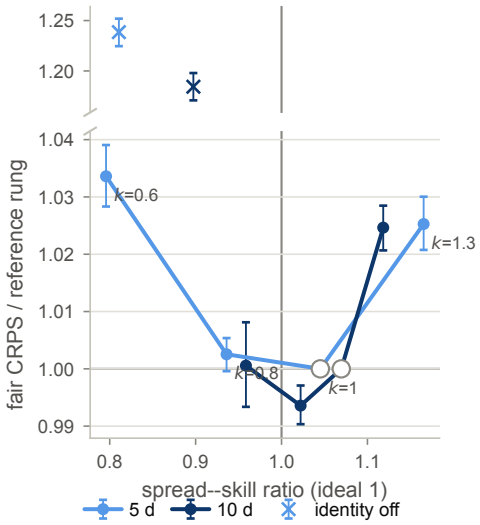

Figure 8: The CRPS optimum in spread: fair CRPS relative to the ladder's reference rung against the spread–skill ratio, cross-variable, at day five and day ten, along four fluctuation scales (k labelled), with the identity-off rung as crosses. Paired on 68 distinct days (1 Jan–18 Dec 2024); bars as in Table 6. The vertical line is the ideal ratio, the horizontal one the reference rung, which is algebraically the trained model but draws its own latents.

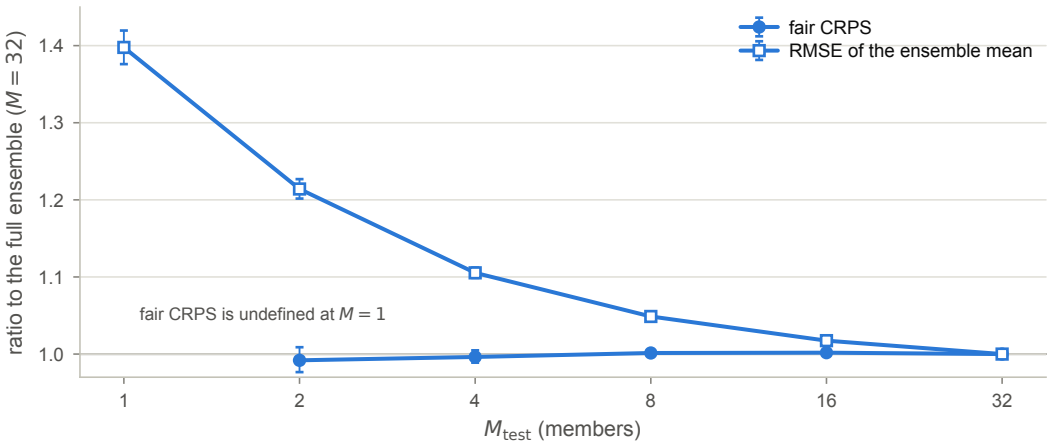

Figure 9: The M_{test} ladder of Table 6(b) at day five: fair CRPS and the RMSE of the ensemble mean as ratios to the full ensemble, against the number of members, from random member subsets of the same forecasts (8 subsets per initialisation and size; $M{=}32$ reproduces the scorecard exactly), on the 68 designed initialisations of the ladder (1 Jan–18 Dec 2024). The single-member point is drawn for RMSE only: at one member the fair correction degenerates to the MAE and the spread is identically zero. Bars: paired moving-block bootstrap, as in Table 6.

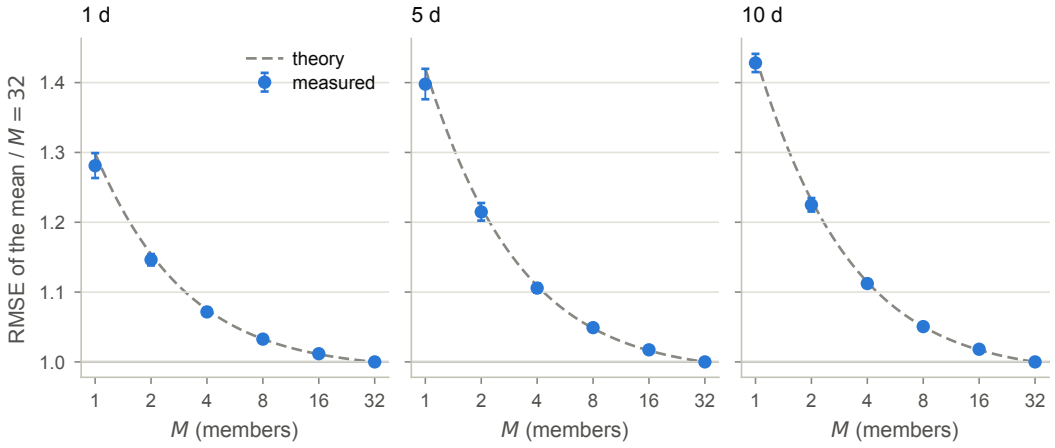

Figure 10: Ensemble size against the exchangeable-member prediction, at days one, five and ten: the RMSE of the ensemble mean at $M \in \{1, 2, 4, 8, 16, 32\}$ members as a ratio to M=32, measured on the subsets of Fig. 9 (points, paired bootstrap bars) and predicted by $\mathrm{RMSE}^2(M) = \mathrm{RMSE}^2(\infty) + \mathrm{spread}^2/M$ (dashed), where $\mathrm{RMSE}(\infty)$ is fixed by the M=32 point and the measured spread, so the curve has no parameter fitted to the points it is judged against. Same initialisations as Fig. 9.

Table 7: The random-core rung: the trained fluctuation of every adapted module replaced by Gaussian noise of the same per-module RMS, the static part and the templates unchanged (App. C), M=32, on n=1460 initialisations of 2024 shared with the mean core and the frozen backbone, and on the n=8 of them shared with the trained HyperENS+ campaign. Spread–skill ratios are within-model absolutes; skill entries are cross-variable means of per-variable ratios, below one is better. Brackets: 95% paired moving-block bootstrap over initialisations (2000 draws), dates resampled once and both arms recomputed on the resample.

	6 h	1 d	3 d	5 d	10 d
spread–skill ratio, pooled (ideal 1)					
random core	0.164	0.187	0.303	0.406	0.711
HyperENS+ (trained), n=8 shared	0.755	0.849	1.001	1.019	1.045
fair CRPS					
random core / mean core (n=1460)	0.896 [0.895, 0.898]	0.883 [0.880, 0.886]	0.821 [0.820, 0.823]	0.762 [0.759, 0.766]	0.612 [0.605, 0.618]
HyperENS+ / mean core (n=8)	0.745 [0.741, 0.749]	0.708 [0.705, 0.710]	0.658 [0.652, 0.664]	0.619 [0.609, 0.631]	0.549 [0.540, 0.558]
random core / HyperENS+ (n=8)	1.206 [1.197, 1.212]	1.250 [1.246, 1.257]	1.254 [1.242, 1.266]	1.236 [1.217, 1.256]	1.131 [1.098, 1.162]
RMSE of the ensemble mean					
random core / mean core (n=1460)	0.999 [0.998, 1.000]	0.994 [0.993, 0.994]	0.973 [0.972, 0.974]	0.940 [0.937, 0.943]	0.826 [0.819, 0.833]
HyperENS+ / mean core (n=8)	0.996 [0.993, 1.000]	0.959 [0.957, 0.962]	0.894 [0.888, 0.900]	0.844 [0.827, 0.864]	0.766 [0.751, 0.778]
random core / HyperENS+ (n=8)	1.004 [1.000, 1.007]	1.037 [1.034, 1.040]	1.091 [1.082, 1.098]	1.115 [1.091, 1.136]	1.093 [1.066, 1.119]

Table 8: The within-model contrast of Sec. 4 (*The short-lead deficit*): HyperENS+ (earlier checkpoint) against HyperENS on the n=418 initialisations of 2024 they share (27 May–21 Dec 2024), M=32. Spread–skill ratios are per-initialisation means, within-model absolutes; ratios are cross-variable means of per-variable ratios, below one favouring HyperENS+. This campaign is initialised at different times of day from the ladder's initialisations (Table 6), so the six-hour values are not comparable between the two tables. Brackets: 95% paired moving-block bootstrap over initialisations (block ℓ=7, 2000 draws), dates resampled once and both arms recomputed on the resample.

	6 h	1 d	5 d	10 d
spread–skill ratio (ideal 1)				
HyperENS	0.370	0.562	0.812	0.957
HyperENS+	0.559	0.830	1.035	1.077
fair CRPS, HyperENS+ / HyperENS	0.931 [0.929, 0.934]	0.929 [0.924, 0.934]	0.948 [0.945, 0.952]	0.976 [0.972, 0.979]
RMSE of the ensemble mean, HyperENS+ / HyperENS	0.994 [0.992, 0.996]	0.983 [0.980, 0.986]	0.974 [0.971, 0.977]	0.987 [0.983, 0.990]

Table 9: Per-variable scorecard for Aurora 1.5 (M=32) on all of its n=711 initialisations (1 Jan–21 Dec 2024), five of the forty scored leads. Means over initialisations; MSLP in hPa and Q_{700} in g kg^{-1}, other variables in native units. Arms are scored on different initialisation sets, so rows of different tables are not paired – use the paired tables for any comparison.

variable [unit]	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
T_{2m} [K]	0.68	0.95	1.28	1.66	2.11
U_{10m} [m s^{-1}]	0.81	1.42	2.12	2.77	3.41
V_{10m} [m s^{-1}]	0.84	1.46	2.19	2.90	3.57
MSLP [hPa]	0.53	1.28	2.55	4.00	5.60
T_{850} [K]	0.66	1.05	1.55	2.13	2.81
Z_{500} [m^2 s^{-2}]	41	122	257	417	614
U_{850} [m s^{-1}]	1.24	2.10	3.08	4.01	4.96
Q_{700} [g kg^{-1}]	0.626	0.917	1.126	1.306	1.494
fair CRPS					
T_{2m} [K]	0.29	0.41	0.55	0.70	0.89
U_{10m} [m s^{-1}]	0.40	0.65	0.96	1.27	1.60
V_{10m} [m s^{-1}]	0.41	0.68	0.99	1.32	1.67
MSLP [hPa]	0.26	0.57	1.05	1.66	2.38
T_{850} [K]	0.33	0.51	0.73	0.99	1.33
Z_{500} [m^2 s^{-2}]	22	58	112	179	267
U_{850} [m s^{-1}]	0.63	1.03	1.48	1.94	2.44
Q_{700} [g kg^{-1}]	0.274	0.414	0.525	0.625	0.734
spread-skill ratio, per-date mean (ideal 1)					
T_{2m}	1.019	0.946	0.945	0.959	0.972
U_{10m}	1.131	1.034	1.022	1.024	1.025
V_{10m}	1.131	1.040	1.025	1.023	1.026
MSLP	0.891	0.891	0.948	0.976	0.996
T_{850}	0.975	0.913	0.936	0.958	0.971
Z_{500}	0.762	0.850	0.934	0.972	0.986
U_{850}	1.030	0.993	0.992	0.999	1.000
Q_{700}	1.033	1.032	1.021	1.007	0.993
reliability index (ideal 0)					
T_{2m}	0.094	0.125	0.149	0.163	0.174
U_{10m}	0.135	0.089	0.079	0.075	0.072
V_{10m}	0.134	0.093	0.082	0.078	0.072
MSLP	0.165	0.154	0.135	0.139	0.148
T_{850}	0.145	0.155	0.160	0.163	0.167
Z_{500}	0.370	0.444	0.451	0.431	0.402
U_{850}	0.040	0.033	0.041	0.050	0.057
Q_{700}	0.052	0.044	0.049	0.055	0.064

Table 10: Per-variable scorecard for HyperENS (M=32) on all of its n=712 initialisations (1 Jan–21 Dec 2024), five of the forty scored leads. Means over initialisations; MSLP in hPa and Q_{700} in g kg^{-1}, other variables in native units. Arms are scored on different initialisation sets, so rows of different tables are not paired – use the paired tables for any comparison.

variable [unit]	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
$T_{2\mathrm{m}}$ [K]	0.66	1.03	1.45	1.87	2.30
$U_{10\mathrm{m}}$ [m s^{-1}]	0.86	1.60	2.39	3.06	3.63
$V_{10\mathrm{m}}$ [m s^{-1}]	0.89	1.64	2.47	3.20	3.79
MSLP [hPa]	0.55	1.49	2.99	4.56	6.08
T_{850} [K]	0.65	1.14	1.73	2.36	3.01
Z_{500} [m^2 s^{-2}]	49	148	307	482	673
U_{850} [m s^{-1}]	1.29	2.33	3.44	4.40	5.27
Q_{700} [g kg^{-1}]	0.667	0.989	1.223	1.412	1.592
fair CRPS					
$T_{2\mathrm{m}}$ [K]	0.33	0.51	0.68	0.86	1.05
$U_{10\mathrm{m}}$ [m s^{-1}]	0.45	0.77	1.13	1.45	1.74
$V_{10\mathrm{m}}$ [m s^{-1}]	0.47	0.80	1.17	1.52	1.82
MSLP [hPa]	0.29	0.70	1.33	2.00	2.68
T_{850} [K]	0.35	0.58	0.85	1.15	1.48
Z_{500} [m^2 s^{-2}]	29	76	145	221	306
U_{850} [m s^{-1}]	0.71	1.21	1.72	2.20	2.65
Q_{700} [g kg^{-1}]	0.318	0.469	0.591	0.695	0.796
spread-skill ratio, per-date mean (ideal 1)					
$T_{2\mathrm{m}}$	0.566	0.802	0.895	0.959	1.056
$U_{10\mathrm{m}}$	0.523	0.685	0.786	0.858	0.935
$V_{10\mathrm{m}}$	0.520	0.675	0.780	0.850	0.927
MSLP	0.589	0.726	0.756	0.816	0.919
T_{850}	0.579	0.773	0.835	0.883	0.961
Z_{500}	0.613	0.731	0.766	0.822	0.915
U_{850}	0.552	0.713	0.796	0.861	0.934
Q_{700}	0.547	0.775	0.869	0.912	0.939
reliability index (ideal 0)					
$T_{2\mathrm{m}}$	0.453	0.312	0.283	0.270	0.270
$U_{10\mathrm{m}}$	0.615	0.412	0.274	0.187	0.112
$V_{10\mathrm{m}}$	0.638	0.438	0.290	0.197	0.116
MSLP	0.464	0.297	0.254	0.215	0.179
T_{850}	0.545	0.311	0.262	0.244	0.235
Z_{500}	0.628	0.611	0.588	0.536	0.480
U_{850}	0.592	0.374	0.243	0.164	0.099
Q_{700}	0.552	0.305	0.196	0.136	0.095

Table 11: Per-variable scorecard for HyperENS+ (M=32) on all of its n=418 initialisations (27 May–21 Dec 2024), five of the forty scored leads. Means over initialisations; MSLP in hPa and Q_{700} in g kg^{-1}, other variables in native units. Arms are scored on different initialisation sets, so rows of different tables are not paired – use the paired tables for any comparison. Not an annual sample: the perc arm starts on 27 May.

variable [unit]	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
T_{2m} [K]	0.65	0.99	1.37	1.76	2.17
U_{10m} [m s^{-1}]	0.86	1.58	2.35	2.99	3.58
V_{10m} [m s^{-1}]	0.89	1.62	2.41	3.12	3.73
MSLP [hPa]	0.54	1.46	2.91	4.40	5.94
T_{850} [K]	0.65	1.11	1.65	2.24	2.86
Z_{500} [m^2 s^{-2}]	44	137	288	453	642
U_{850} [m s^{-1}]	1.30	2.31	3.39	4.31	5.21
Q_{700} [g kg^{-1}]	0.660	0.976	1.209	1.395	1.571
fair CRPS					
T_{2m} [K]	0.31	0.48	0.64	0.81	1.00
U_{10m} [m s^{-1}]	0.42	0.74	1.08	1.40	1.71
V_{10m} [m s^{-1}]	0.44	0.76	1.12	1.46	1.77
MSLP [hPa]	0.28	0.67	1.26	1.90	2.59
T_{850} [K]	0.33	0.55	0.80	1.08	1.39
Z_{500} [m^2 s^{-2}]	23	63	124	193	276
U_{850} [m s^{-1}]	0.67	1.16	1.67	2.13	2.61
Q_{700} [g kg^{-1}]	0.302	0.453	0.576	0.679	0.780
spread-skill ratio, per-date mean (ideal 1)					
T_{2m}	0.793	1.112	1.208	1.240	1.303
U_{10m}	0.803	0.915	0.973	1.003	1.017
V_{10m}	0.740	0.888	0.958	0.985	1.005
MSLP	0.837	0.985	0.983	1.000	1.029
T_{850}	0.834	1.039	1.076	1.086	1.111
Z_{500}	0.995	1.069	1.035	1.040	1.058
U_{850}	0.837	0.952	0.990	1.014	1.028
Q_{700}	0.799	1.010	1.056	1.064	1.062
reliability index (ideal 0)					
T_{2m}	0.180	0.248	0.301	0.320	0.343
U_{10m}	0.246	0.132	0.090	0.079	0.075
V_{10m}	0.363	0.174	0.093	0.070	0.060
MSLP	0.209	0.239	0.235	0.220	0.210
T_{850}	0.221	0.155	0.183	0.191	0.194
Z_{500}	0.231	0.327	0.317	0.301	0.286
U_{850}	0.227	0.115	0.086	0.081	0.080
Q_{700}	0.248	0.100	0.078	0.074	0.075

Table 12: Absolute scores of the HyperENS retrofit on the re-implemented Pangu-Weather, M=32, on the 154 initialisations of Fig. 11: RMSE of the ensemble mean and fair CRPS, with the 95% intervals of that evaluation. Z_{500} in $\mathrm{m^2\,s^{-2}}$, temperatures in K, wind in $\mathrm{m\,s^{-1}}$.

	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
Z_{500}	60 ± 6	197 ± 10	396 ± 20	562 ± 28	735 ± 74
T_{850}	0.83 ± 0.08	1.48 ± 0.07	2.19 ± 0.11	2.70 ± 0.13	3.26 ± 0.33
T_{2m}	0.82 ± 0.08	1.32 ± 0.13	1.81 ± 0.18	2.25 ± 0.22	2.69 ± 0.27
U_{10m}	1.06 ± 0.11	2.01 ± 0.10	2.89 ± 0.14	3.35 ± 0.17	3.76 ± 0.38
fair CRPS					
Z_{500}	33.8 ± 3.4	97 ± 5	186 ± 10	260 ± 14	345 ± 35
T_{850}	0.460 ± 0.046	0.776 ± 0.039	1.107 ± 0.057	1.347 ± 0.074	1.63 ± 0.17
T_{2m}	0.418 ± 0.044	0.682 ± 0.070	0.908 ± 0.094	1.11 ± 0.12	1.31 ± 0.15
U_{10m}	0.582 ± 0.058	1.016 ± 0.051	1.431 ± 0.073	1.667 ± 0.085	1.88 ± 0.19

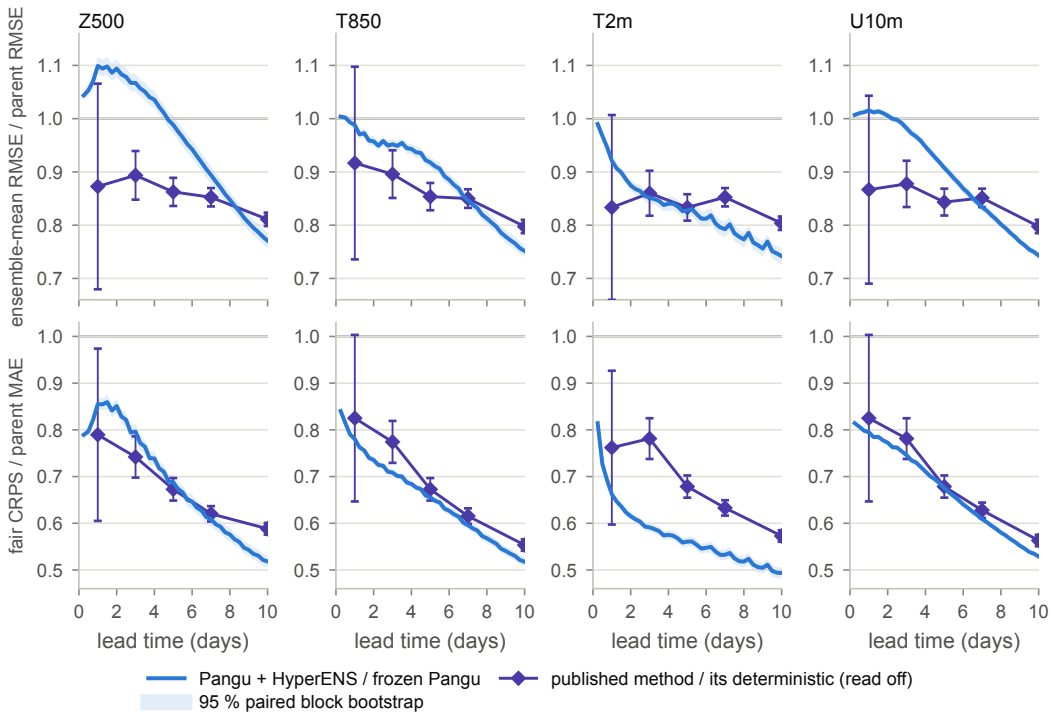

Figure 11: Gain over the deterministic forecast on the second backbone: RMSE of the ensemble mean over the deterministic RMSE (top) and fair CRPS over the deterministic MAE (bottom), against lead, for HyperENS on the re-implemented Pangu-Weather (line, 95% paired moving-block bootstrap band) on 154 daily initialisations of 2019 (2 January–4 June), $M=32$. Diamonds: the ratios Xiong et al. (2025) report for their ensemble over their own deterministic model, read off their figures, with the read-off error as bars; the axis labels' "parent" is the deterministic model of each system.

E PERTURBED INITIAL CONDITIONS: CONSTRUCTION, AND A PILOT

E.1 CONSTRUCTION

Each member of a perturbed-initial-condition configuration starts from ERA5 plus a perturbation taken from the operational ECMWF ensemble (ENS) through the TIGGE archive (Bougeault et al., 2010): $\mathrm{IC}(m,t) = \mathrm{ERA5}(t) + \delta(m,t)$, with $\delta(m,t)$ the difference between perturbed member m and the control forecast of the same ENS run at the same lead. Two choices in that line are deliberate. The perturbation is centred on the control rather than on the ensemble mean, because ENS members are generated as perturbations about the control analysis, so member minus control *is* the perturbation by construction rather than an estimate of one. And it is added to ERA5 rather than substituted for the member state, which keeps the unperturbed baseline exactly ERA5 and keeps the frozen backbone started from ERA5 a valid control; substituting the raw member state would fold the IFS-minus-ERA5 analysis difference into what is called the perturbation. Verification remains against ERA5 although the perturbations come from the IFS; this puts a small analysis-difference bias into every perturbed configuration equally, so it does not distort the comparisons between them, but it is not zero.

Aurora is initialised from two states, at $T - 6\,\mathrm{h}$ and T, and ENS runs twice daily, so initialisations are placed six hours after an ENS run and both states are taken from that single ENS run, initialised at $T - 6\,\mathrm{h}$: the state at $T - 6\,\mathrm{h}$ from its initial condition, where member minus control is the initial perturbation the IFS itself applied, and the state at T from its 6-h lead, each differenced against the control of its own lead. The two states therefore lie on one physically consistent member trajectory six hours apart, and the perturbation is flow-dependent at both; freezing one perturbation across both timestamps, or leaving $T - 6\,\mathrm{h}$ unperturbed, would inject a spurious six-hour tendency equal to the perturbation and dominate short-lead spread.

The archive serves eight pressure levels for all five atmospheric variables (200, 250, 300, 500, 700, 850, 925 and 1000 hPa) against Aurora's thirteen. At 50, 100 and 150 hPa the perturbation is exactly zero: extrapolating three levels off the top of the data is indefensible, and perturbing only geopotential there would produce a hydrostatically inconsistent state Aurora was never trained on. At 400 and 600 hPa the perturbation is filled by linear interpolation in log-pressure, exact at its knots and never extrapolating. The fill costs amplitude, because ENS perturbations have fine vertical structure: adjacent served levels are essentially uncorrelated (the correlation of the temperature perturbation is -0.10 between 300 and 500 hPa and -0.05 between 500 and 700 hPa, measured on one member), so a linear blend of two of them has less variance than either, and the interpolated levels carry a perturbation about 30% smaller in RMS than their neighbours. All eight scored variables sit on measured levels or at the surface, so the direct effect on the scorecard is bounded, but the state Aurora is initialised from is under-perturbed in the upper troposphere and unperturbed in the stratosphere; the initial-condition-driven spread below is therefore a lower bound on what a true IFS ENS member would produce. This is a property of the archive, not of the method.

TIGGE serves geopotential height in geopotential metres; it is converted to geopotential with $g = 9.80665\,\mathrm{m\,s^{-2}}$ once, when the archive is built, and the magnitude of the control field is asserted before the difference is taken, since after differencing a factor of g is no longer detectable. The perturbation, retrieved on a regular 0.5° grid, is interpolated bilinearly (periodic in longitude, clamped at the poles) onto Aurora's 0.25° grid and then added; ERA5 itself never passes through an interpolator. Specific humidity is clamped at zero after the perturbation is added, and the clamped fraction is recorded.

The design is a crossed factorial: K initial conditions, the first K ENS members by number, times M latent draws, the same M latents reused for every initial condition, with $K \cdot M = 32$ for every configuration so that the fair-CRPS estimator bias and the rank-histogram bin count are identical across the rows of Table 13. Reusing the latents is what makes the design crossed and permits a decomposition of spread into initial-condition, weight and interaction components; redrawing them per initial condition would cost the same and destroy it. The pilot, referred to in the Limitations paragraph of Sec. 5, carries two initialisations, 2024-01-05 and 2024-07-03, with 32 perturbed members each in the store; every arm — the frozen backbone at $K{=}32$, $M{=}1$, and each retrofit at $K \in \{2, 4, 8\}$ — is scored on both by the same roll-out, metric handlers and artifact writer as the campaign of Sec. 4, so every comparison in the pilot is paired and none is an annual estimate. Two initialisations bound what the pilot can say to signs and orderings, and every caption says so; a store on more dates is the

obvious next step, and is a matter of archive volume, of the order of a gigabyte of retrieved fields per initialisation.

The perturbations were built offline from TIGGE (the ECMWF Data Store, dataset `tigge-forecasts`, origin `ecmwf`, on the $0.5°$ grid), verified against the live archive — perturbation RMS of the right magnitude for an ENS perturbation, growing from the earlier to the later slice, and near-zero correlation between members — and uploaded as a pre-cooked archive that the evaluation job reads at run time exactly like the model weights. The job itself has no network dependency: "retrieved from TIGGE" describes how the archive was built, not what the evaluation did.

E.2 What the pilot shows

The design is a mixture at the fixed budget of 32 members, K perturbed initial conditions times M latent draws with $K\,M = 32$; K=4, M=8 matches the training-time ensemble size, and K=32, M=1 on the frozen backbone, no adapter attached, is the literal reading of the title question and the reference every ratio is taken against. The runs are zero-shot for both arms. Three things hold on both dates. The frozen backbone under perturbed initial conditions alone is not an ensemble: its spread–skill ratio is below one at every lead in every variable and U-shaped in lead (Table 16), Hollingsworth's lesson recovered in the ML setting, and its rank histograms are deep Us with a large positive tilt on Z_{500}, $T_{2\mathrm{m}}$ and T_{850} that survives to day ten (Table 17), a low bias of the frozen checkpoint. The retrofit's mixtures remove the U and reach one by day five, HyperENS+ closer than HyperENS, with fair CRPS about a third lower than the initial-condition ensemble's on $T_{2\mathrm{m}}$ and Z_{500} at day five and a few percent worse on the winds and the pressure at day one (Table 13). And the best split moves with lead (Table 15): K=8 at day one for every variable in both arms, K=4 or K=2 by day ten — the two classes are complementary. Fair CRPS and the $\sqrt{(M+1)/M}$ prefactor assume independent members, which a $K\times M$ design does not satisfy; neither is adjusted.

Table 13 is the pilot's headline — per-variable dispersion and skill of the two retrofits against the initial-condition-only ensemble on the frozen backbone at the matched rung K=4 × M=8; Table 14 makes its ratios dimensional again; Table 15 gives the per-rung sweep at three leads, Table 16 the shape of the spread–skill profile over the forty leads, and Table 17 the rank-histogram statistics, with the tilt in the convention of Sec. 3.3.

Figs. 12 and 13 are the pilot's rank histograms on $T_{2\mathrm{m}}$, Z_{500} and $U_{10\mathrm{m}}$.

F Two controls from the development configuration

Two controls from development remain here because nothing in the evaluation campaign of Sec. 4 replaces them: the weight-space capacity audit, which answers whether more rank or a larger core scale was ever the constraint, and the random-direction perception probe, which puts an untrained perturbation into the encoder channel, a control that the HyperENS+-versus-HyperENS contrast of Sec. 4 does not provide. They were run during development, before the campaign was affordable, on the checkpoints of the development run from which the HyperENS arm was taken, and are reported in that configuration rather than blurred into the main text: scored as described in the harness paragraph that follows, against the frozen backbone itself rather than against Aurora 1.5. No value here is comparable with a value in the main text, even for the same checkpoint. The spread–skill ratios here use the development scorecard's earlier convention — a mean-of-standard-deviation numerator against an RMS denominator, and a population variance paired with the unbiased finite-ensemble factor — which reads low relative to the RMS convention of the main text by a factor that depends on the model and cannot be converted; they are consistent among themselves and nowhere else. The variable set is the development scorecard's, one variable larger than that of Sec. 3.3. The processor-side ladder that the probe's second result refers to was the development-configuration ladder, not reproduced here; its successor at the campaign configuration is Fig. 8.

Every ablation below is an *evaluation-time* intervention on already-trained weights unless stated otherwise, scored on the harness of that configuration: $M = 8$ members, daily initialisations over 2019-01-01 to 2019-06-03, twenty roll-out steps of six hours, to day 5, nine variables scored (v_{850} included, unlike the main campaign), latitude-weighted throughout. Ratios are candidate over baseline, so < 1 is better; the spread–skill ratio (SSR) and the rank-histogram reliability index are

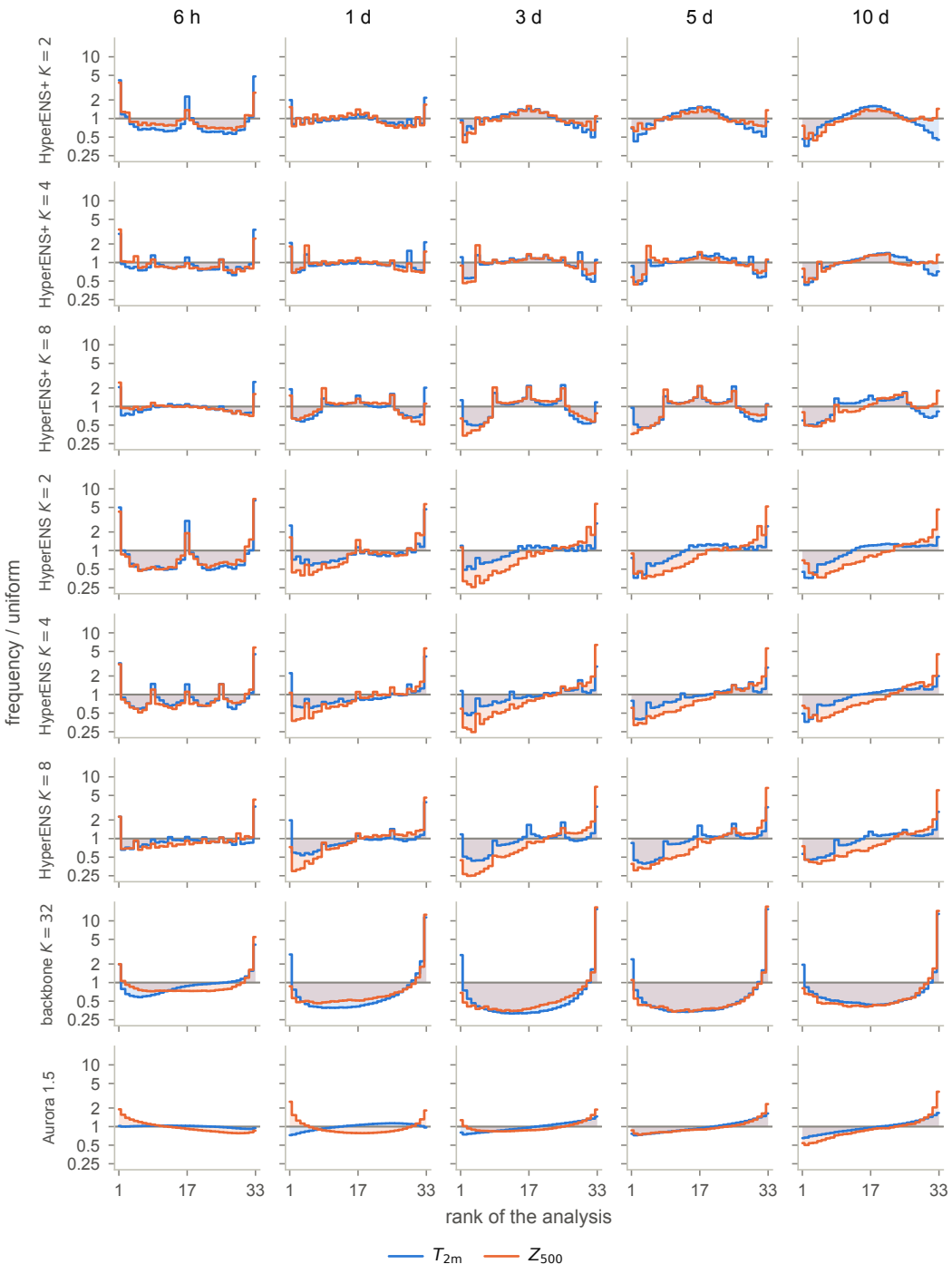

Figure 12: Average rank histograms of the pilot on $T_{2\mathrm{m}}$ and Z_{500}: one row per arm — HyperENS+ and HyperENS at $K \in \{2, 4, 8\}$ with $K\,M = 32$, the frozen backbone at $K{=}32$, and Aurora 1.5 — one column per lead; counts pooled over the two initialisations and all grid points and normalised against a uniform $1/(M{+}1)$ (the horizontal line), logarithmic axis. Aurora 1.5 is drawn from the only artifact of it that stores bin counts, twelve initialisations over 1–7 January 2024, and is a reference for shape, not a paired row.

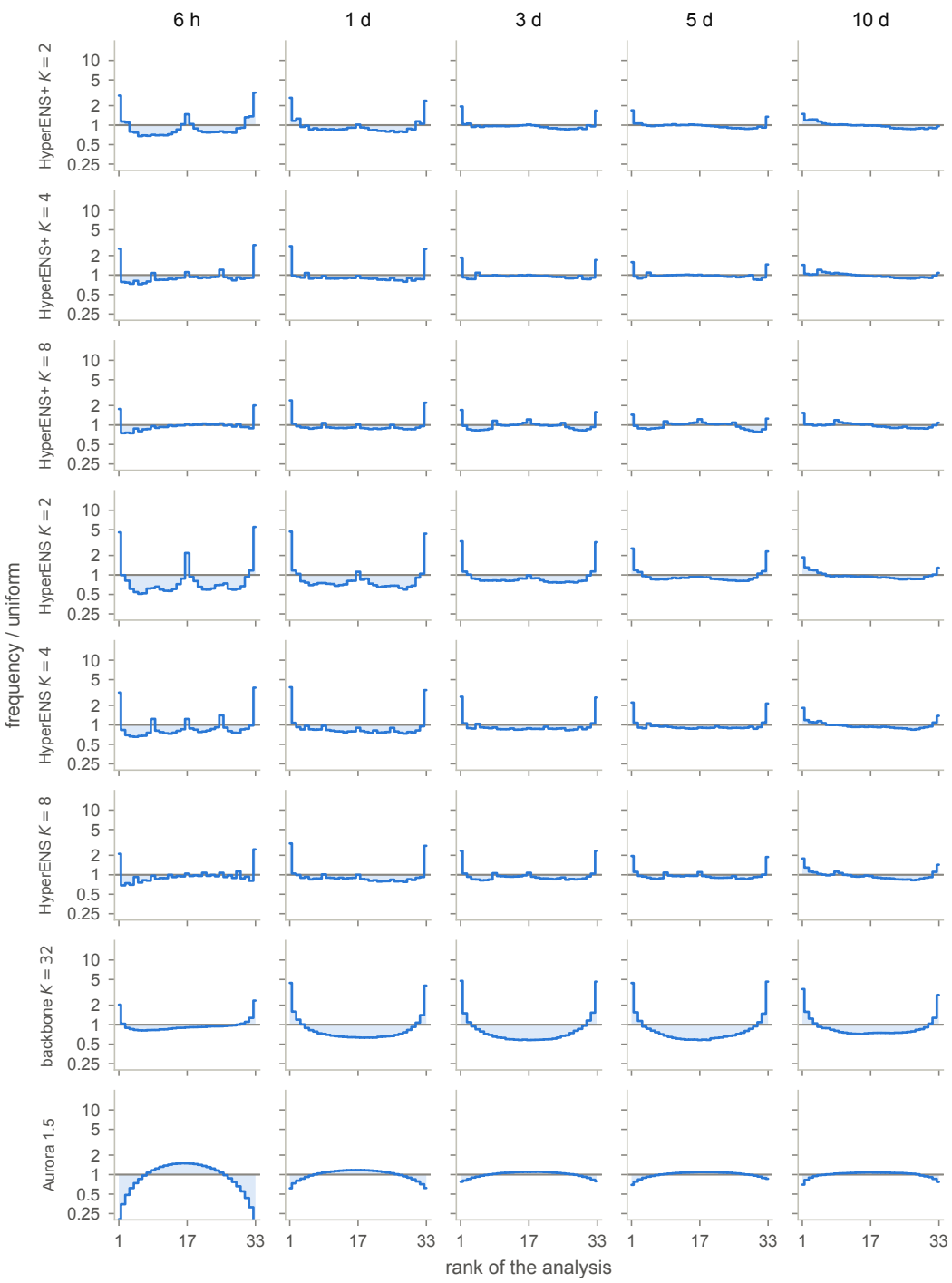

Figure 13: As Fig. 12, for $U_{10\mathrm{m}}$.

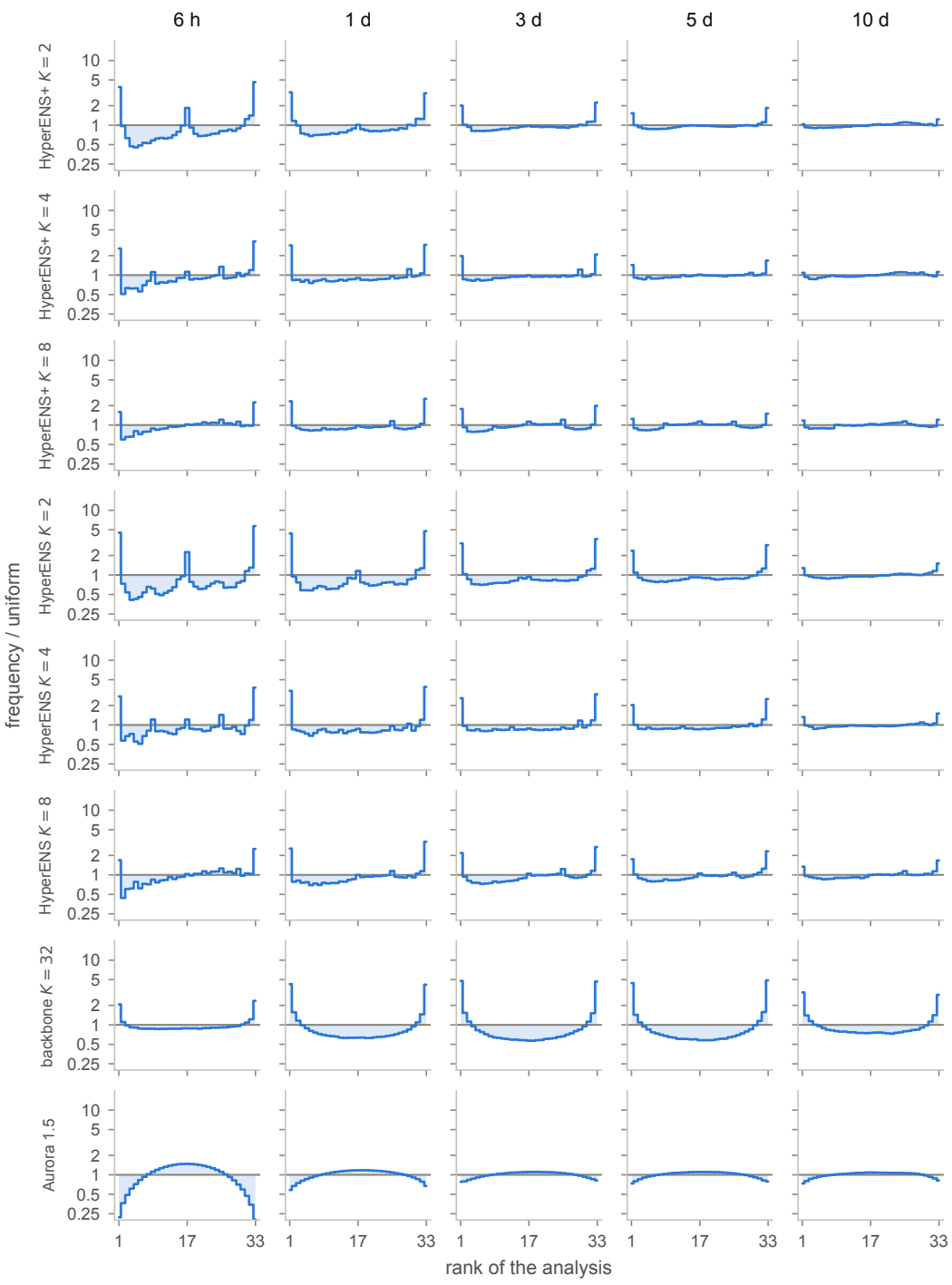

Figure 14: As Fig. 12, for $V_{10\mathrm{m}}$.

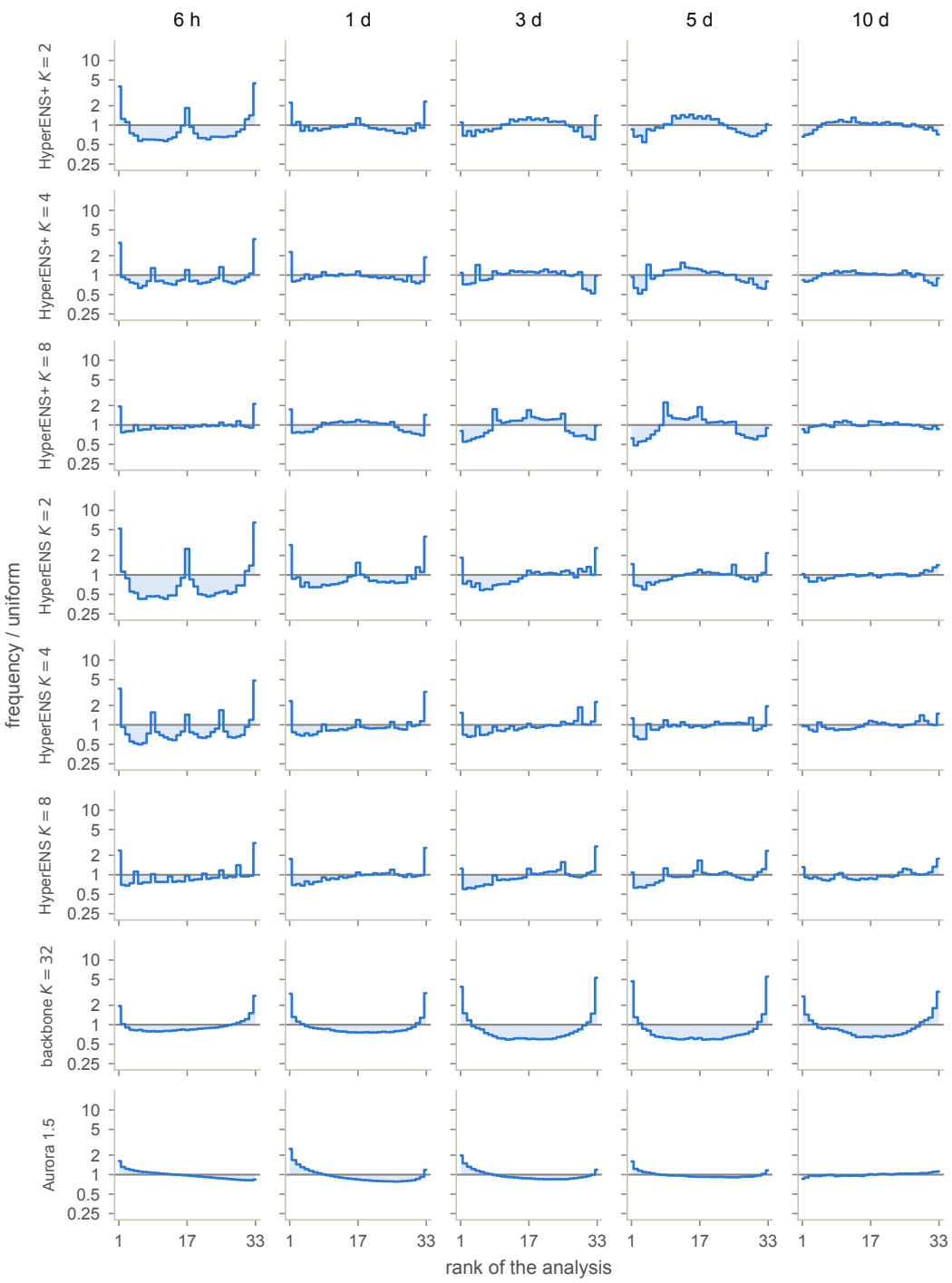

Figure 15: As Fig. 12, for MSLP.

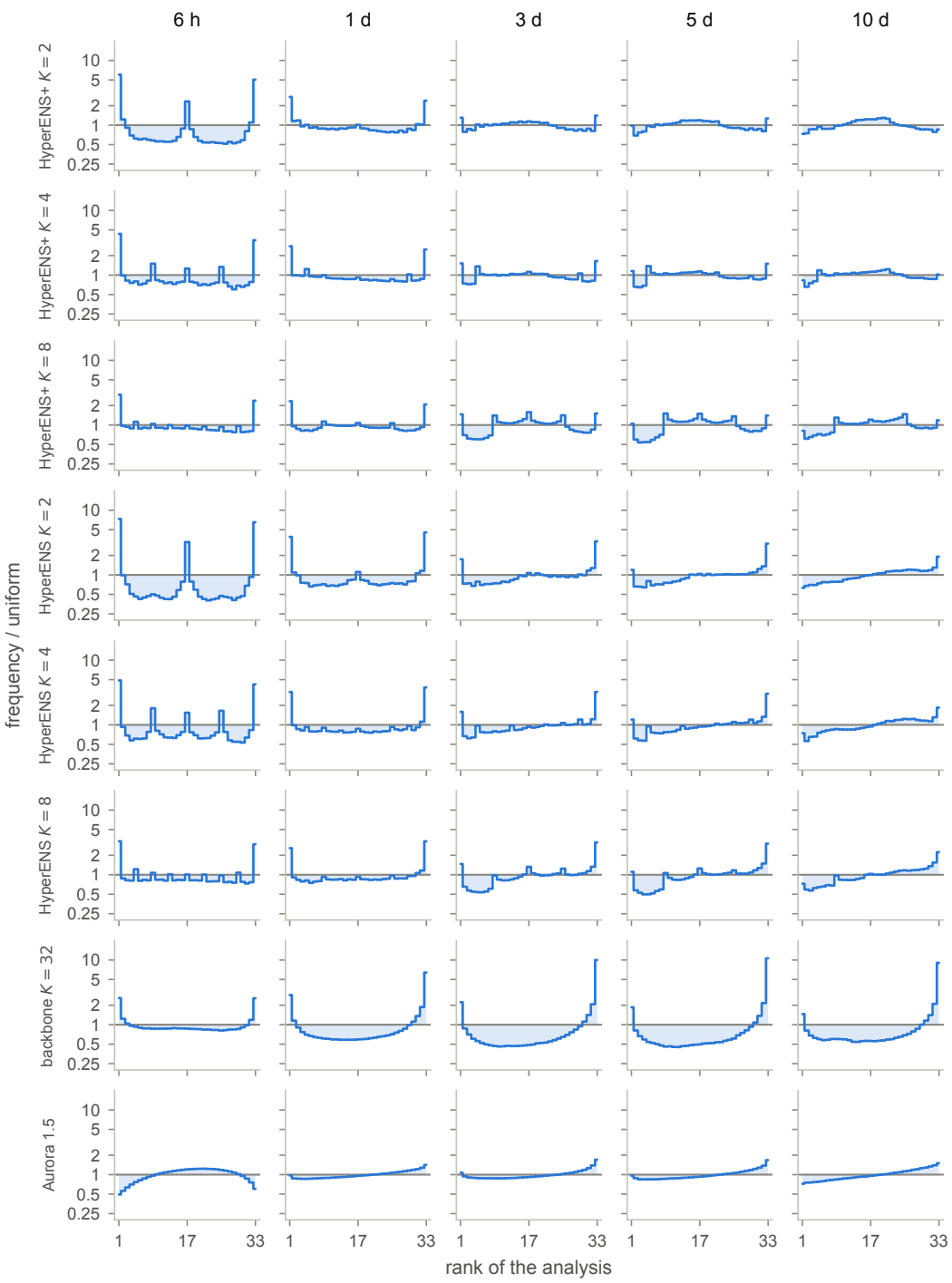

Figure 16: As Fig. 12, for T_{850}.

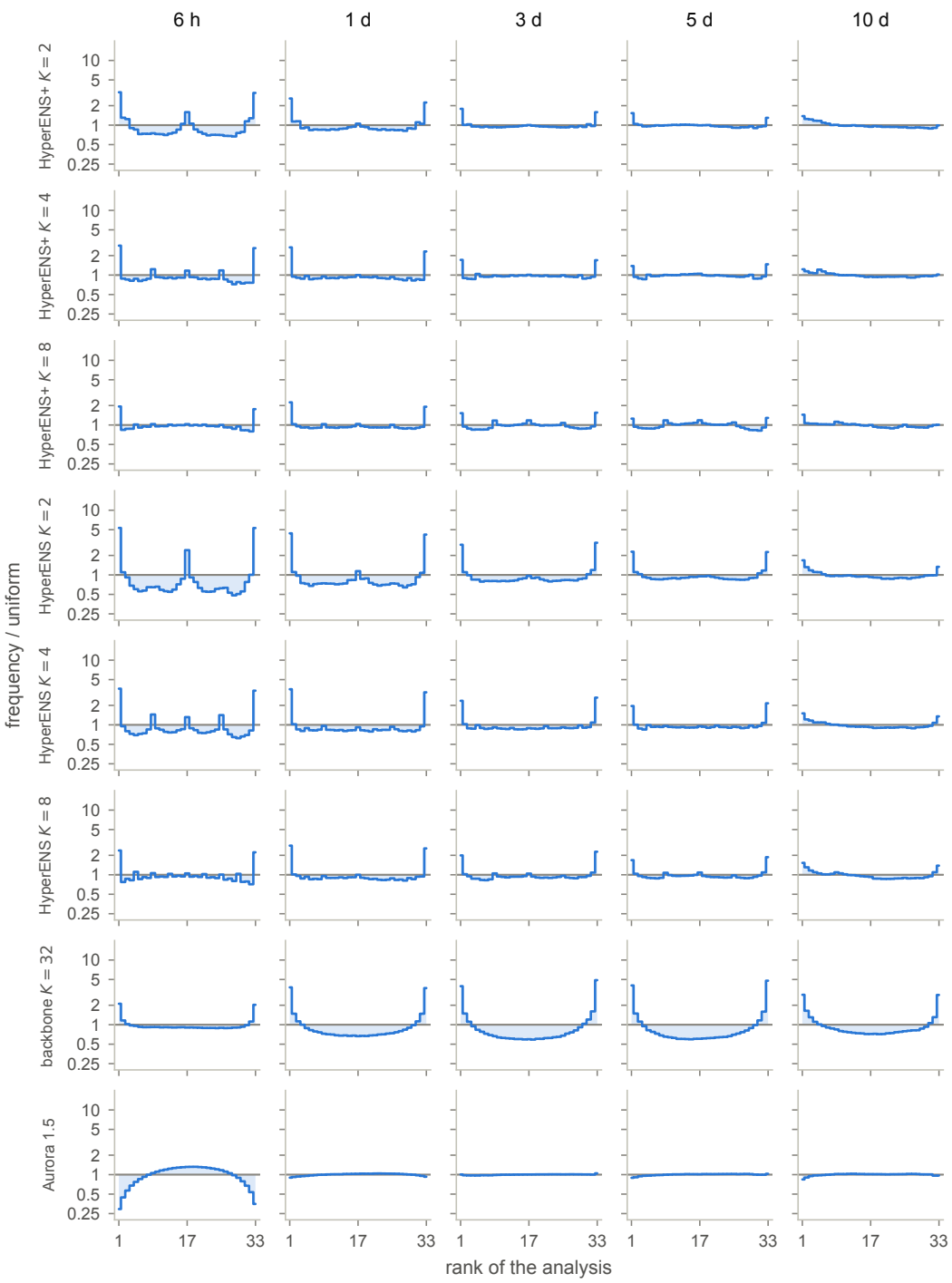

Figure 17: As Fig. 12, for U_{850}.

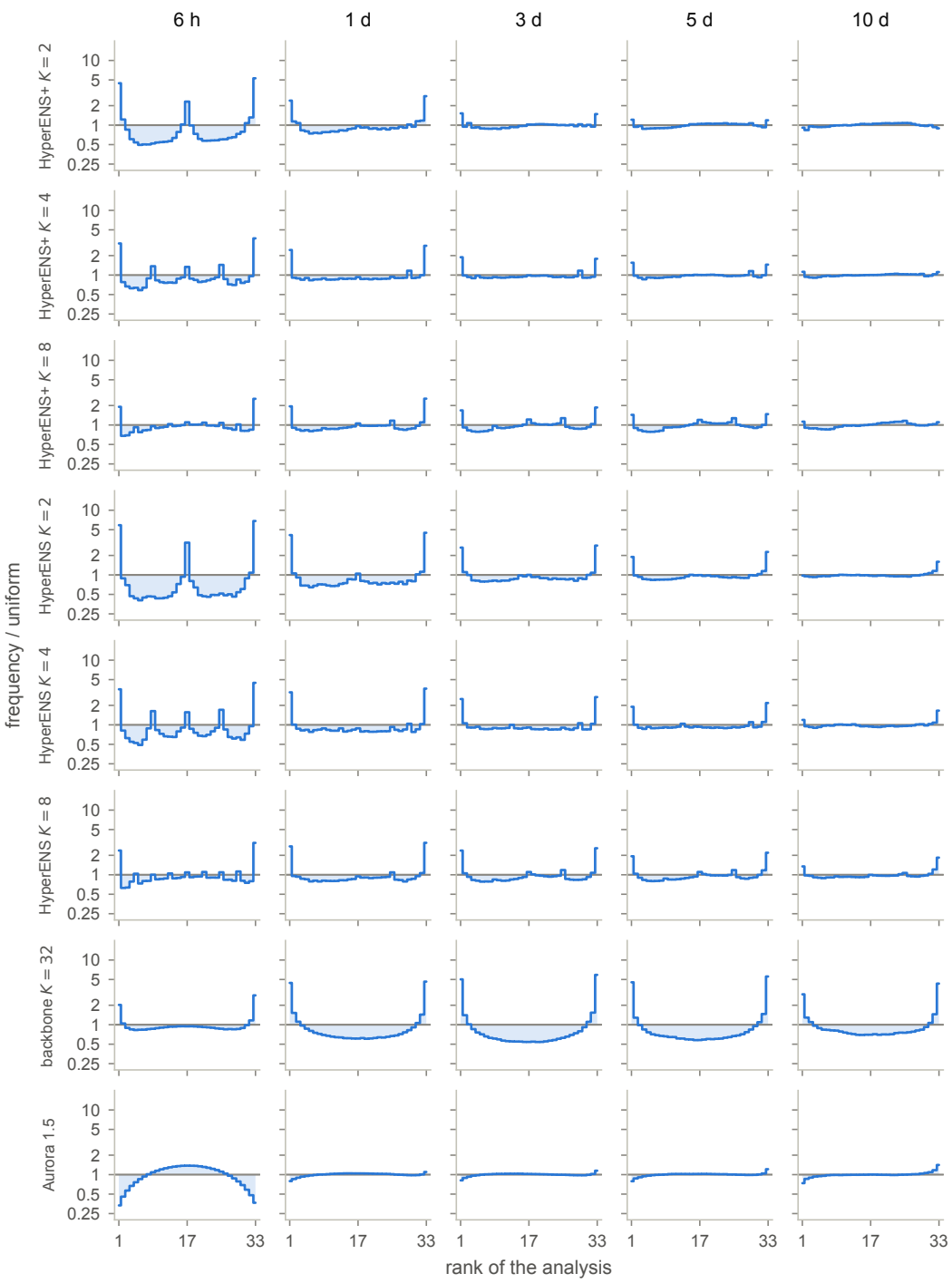

Figure 18: As Fig. 12, for Q_{700}.

Table 13: Per-variable dispersion and skill of the two hypernet-LoRA retrofits against an initial-condition-only ensemble on the frozen backbone, at a matched budget of 32 members and the matched middle rung $K{=}4 \times M{=}8$ (K perturbed ICs $\times$ M latent draws). All three arms are scored by the same rollout, metric handlers and output writer on the same $n{=}2$ initialisations, 2024-01-05 and 2024-07-03, so every column is paired. Lower is better in the lower panel; 1 is ideal in the upper one. SSR is $\sqrt{(M{+}1)/M}\,\langle\sigma\rangle_{\mathrm{rms}}/\mathrm{RMSE}$, the convention a calibrated ensemble scores 1 on, averaged over initialisations; ratios are ratios of the two arms' means over initialisations. At $n{=}2$ no interval is quoted; these are the two pilot initialisations of the store and the table is a pilot-scale readout, not an annual estimate.

	IC-only (frozen)			HyperENS			HyperENS+		
	1 d	5 d	10 d	1 d	5 d	10 d	1 d	5 d	10 d
spread-skill ratio, per-date mean (ideal 1)									
$T_{2\mathrm{m}}$	0.470	0.610	0.749	0.682	0.985	1.079	0.792	1.100	1.110
$U_{10\mathrm{m}}$	0.745	0.740	0.856	0.793	0.889	0.952	0.870	0.962	0.979
$V_{10\mathrm{m}}$	0.745	0.762	0.860	0.794	0.897	0.966	0.854	0.982	1.009
MSLP	0.861	0.811	0.884	0.838	0.939	0.978	0.895	1.036	1.017
Z_{500}	0.778	0.722	0.807	0.897	0.947	0.938	1.045	1.076	0.995
T_{850}	0.681	0.684	0.779	0.779	0.907	1.025	0.841	0.985	1.046
U_{850}	0.723	0.734	0.845	0.802	0.907	0.945	0.875	0.976	0.973
Q_{700}	0.665	0.623	0.724	0.766	0.941	0.986	0.852	1.006	1.048
fair CRPS, relative to the IC-only ensemble									
$T_{2\mathrm{m}}$	1	1	1	0.835	0.685	0.704	0.828	0.689	0.728
$U_{10\mathrm{m}}$	1	1	1	1.047	0.935	0.976	1.044	0.934	0.988
$V_{10\mathrm{m}}$	1	1	1	1.044	0.928	0.947	1.033	0.909	0.942
MSLP	1	1	1	1.096	0.945	0.951	1.096	0.950	0.963
Z_{500}	1	1	1	0.860	0.761	0.835	0.792	0.702	0.812
T_{850}	1	1	1	1.002	0.857	0.823	1.009	0.849	0.833
U_{850}	1	1	1	1.060	0.930	0.983	1.059	0.930	1.004
Q_{700}	1	1	1	1.019	0.899	0.908	1.012	0.897	0.910

absolute. This design is deliberate: it lets a hypothesis whose fix would cost a full training run be tested for the price of a few inference runs, and the ordering of the subsections is the order in which the hypotheses were killed.

F.1 Is the parameterisation the constraint?

The norms of Eq. (3) are computed exactly and offline via the identity

$$\left\|B_j M A_j\right\|_F^2 \;=\; \mathrm{tr}\big(M^\top (B_j^\top B_j)\, M\, (A_j A_j^\top)\big), \tag{9}$$

which reduces every weight-space computation to $r \times r$ algebra regardless of layer width, so this audit runs from a 13 MB checkpoint extract in seconds on CPU.

Capacity is not binding. Rebuilding the hypernet offline from the trained extract and drawing 2048 latents: the $\tanh$ bound is used to 27% of its range (a $3.77\times$ amplitude headroom), the rank-8 core is half used (4.13 effective directions), and the worst stage loses 13% of its gradient to saturation. Raising r or the core scale therefore cannot be the fix, and we do not report it as one.

F.2 Perturbing perception rather than processing

Every member encodes the same fields into the same latent representation: the adapters act only on the processor, so members are bit-identical at $t = 0$ and all spread must be manufactured by integrating a model-error perturbation forward. That is the natural explanation for the short-lead deficit. We therefore priced a second channel before training it, by perturbing the encoder's learned query latents — the vectors its cross-attention uses to interrogate the input — with a random direction of relative amplitude ε, on top of the static model so that all measured spread is attributable to the probe:

Table 14: Absolute scores of the initial-condition-only reference arm (K=32 × M=1: 32 TIGGE-derived perturbed initial conditions advanced by the unmodified frozen backbone), five of the forty scored leads. Every ratio in the other tables of this appendix has this arm as its denominator, so this is the table that makes them dimensional again. n=2 initialisations, 2024-01-05 and 2024-07-03. MSLP in hPa and Q_{700} in g kg^{-1}, other variables in native units. At n=2 no interval is quoted; these are the two pilot initialisations of the store and the table is a pilot-scale readout, not an annual estimate.

	1 d	3 d	5 d	7 d	10 d
RMSE of the ensemble mean					
$T_{2\mathrm{m}}$ [K]	0.74	1.22	1.64	2.25	2.82
$U_{10\mathrm{m}}$ [m s^{-1}]	0.85	1.58	2.47	3.09	3.58
$V_{10\mathrm{m}}$ [m s^{-1}]	0.89	1.66	2.46	3.36	3.75
MSLP [hPa]	0.54	1.49	2.99	4.94	5.79
Z_{500} [m^2 s^{-2}]	52	160	321	536	688
T_{850} [K]	0.68	1.20	1.84	2.57	3.27
U_{850} [m s^{-1}]	1.27	2.32	3.56	4.60	5.19
Q_{700} [g kg^{-1}]	0.656	0.998	1.236	1.467	1.610
fair CRPS					
$T_{2\mathrm{m}}$ [K]	0.41	0.73	0.97	1.22	1.48
$U_{10\mathrm{m}}$ [m s^{-1}]	0.43	0.78	1.18	1.49	1.77
$V_{10\mathrm{m}}$ [m s^{-1}]	0.45	0.81	1.21	1.58	1.85
MSLP [hPa]	0.27	0.70	1.32	2.11	2.65
Z_{500} [m^2 s^{-2}]	33	100	179	273	361
T_{850} [K]	0.37	0.66	1.00	1.35	1.76
U_{850} [m s^{-1}]	0.67	1.22	1.83	2.37	2.71
Q_{700} [g kg^{-1}]	0.306	0.496	0.636	0.754	0.841
spread-skill ratio, per-date mean (ideal 1)					
$T_{2\mathrm{m}}$	0.470	0.481	0.610	0.706	0.749
$U_{10\mathrm{m}}$	0.745	0.779	0.740	0.821	0.856
$V_{10\mathrm{m}}$	0.745	0.779	0.762	0.800	0.860
MSLP	0.861	0.817	0.811	0.807	0.884
Z_{500}	0.778	0.725	0.722	0.741	0.807
T_{850}	0.681	0.648	0.684	0.745	0.779
U_{850}	0.723	0.739	0.734	0.778	0.845
Q_{700}	0.665	0.592	0.623	0.649	0.724

Three results. First, the channel is real and its spread is structurally sound: the mid-band spectral ratio is unchanged from the control at every amplitude (between -0.16 and -0.18 at every lead), so it never degenerates into grid-scale noise. Second, the CRPS minimum falls at $\varepsilon = 0.20$, where day-5 SSR is 0.692 — the same optimum the development-configuration processor ladder (not reproduced here; its campaign-configuration successor is Fig. 8) located from an unrelated perturbation. Third, the last row is the trained processor channel at a similar day-five spread–skill ratio (0.685 against 0.692), not at the same perturbation amplitude: on the four normalised scores shown it is about ten to thirteen percent lower than the random probe at $\varepsilon = 0.20$. The two configurations differ in both channel and training, so the gap cannot be attributed to either alone.

We note one caution the probe also exposes: the 6-hour variogram degrades monotonically with amplitude (0.999, 1.021, 1.069, 1.139), so at large perturbations the short-lead spatial structure goes wrong even while the spectrum stays intact.

G ANTITHETIC SAMPLING

The measurements in this appendix are on the development configuration of App. F and carry its caveats; day five is the last lead of that configuration.

How long does the frozen model's response stay linear? The generated core is close to an odd function of the latent: the cosine between $\delta(z)$ and $\delta(-z)$ is -0.90 and 95–96% of the core's energy

Table 15: Mixture sweep at a fixed budget of 32 members: every arm splits the same 32 members into K perturbed initial conditions $\times$ M latent draws, with the same M latents reused for every IC. Entries are fair CRPS relative to the IC-only ensemble of Table 14; below 1 beats it. The best split migrates with lead: $K{=}8 \times M{=}4$ is the best rung for all eight variables at day 1 in both families, while by day 10 HyperENS is best at $K{=}4 \times M{=}8$ for all eight and HyperENS+ at $K{=}2 \times M{=}16$ for four of them. HyperENS never beats HyperENS+ on Z_{500} at any of the forty leads, but overtakes it over the last two days on $T_{2\mathrm{m}}$, T_{850}, Q_{700}, MSLP and U_{850}. $n{=}2$ initialisations, 2024-01-05 and 2024-07-03. Fair CRPS assumes independent members, which a $K{\times}M$ design violates; it is not adjusted. At $n{=}2$ no interval is quoted; these are the two pilot initialisations of the store and the table is a pilot-scale readout, not an annual estimate.

	$T_{2\mathrm{m}}$	$U_{10\mathrm{m}}$	$V_{10\mathrm{m}}$	MSLP	Z_{500}	T_{850}	U_{850}	Q_{700}
1 d: fair CRPS relative to the IC-only ensemble								
HyperENS $K{=}2 \times M{=}16$	0.863	1.127	1.117	1.201	0.937	1.059	1.146	1.087
HyperENS $K{=}4 \times M{=}8$	0.835	1.047	1.044	1.096	0.860	1.002	1.060	1.019
HyperENS $K{=}8 \times M{=}4$	0.809	1.001	0.997	1.045	0.805	0.962	1.013	0.981
HyperENS+ $K{=}2 \times M{=}16$	0.842	1.102	1.088	1.182	0.864	1.044	1.124	1.066
HyperENS+ $K{=}4 \times M{=}8$	0.828	1.044	1.033	1.096	0.792	1.009	1.059	1.012
HyperENS+ $K{=}8 \times M{=}4$	0.793	1.010	0.983	1.038	0.731	0.953	1.024	0.975
5 d: fair CRPS relative to the IC-only ensemble								
HyperENS $K{=}2 \times M{=}16$	0.712	0.994	0.992	1.033	0.840	0.902	0.976	0.918
HyperENS $K{=}4 \times M{=}8$	0.685	0.935	0.928	0.945	0.761	0.857	0.930	0.899
HyperENS $K{=}8 \times M{=}4$	0.715	0.945	0.931	0.982	0.798	0.880	0.937	0.920
HyperENS+ $K{=}2 \times M{=}16$	0.710	0.978	0.960	1.016	0.767	0.882	0.965	0.917
HyperENS+ $K{=}4 \times M{=}8$	0.689	0.934	0.909	0.950	0.702	0.849	0.930	0.897
HyperENS+ $K{=}8 \times M{=}4$	0.684	0.913	0.884	0.941	0.691	0.829	0.915	0.892
10 d: fair CRPS relative to the IC-only ensemble								
HyperENS $K{=}2 \times M{=}16$	0.720	0.987	0.954	0.958	0.858	0.846	0.998	0.921
HyperENS $K{=}4 \times M{=}8$	0.704	0.976	0.947	0.951	0.835	0.823	0.983	0.908
HyperENS $K{=}8 \times M{=}4$	0.753	0.997	0.971	1.022	0.908	0.871	1.000	0.934
HyperENS+ $K{=}2 \times M{=}16$	0.724	0.990	0.937	0.941	0.802	0.843	1.004	0.928
HyperENS+ $K{=}4 \times M{=}8$	0.728	0.988	0.942	0.963	0.812	0.833	1.004	0.910
HyperENS+ $K{=}8 \times M{=}4$	0.743	0.993	0.958	1.003	0.846	0.853	0.998	0.917

is in its odd part. Whether that antisymmetry survives the frozen trunk measures how far the model's response stays linear in its own weight perturbation. We ran $\pm k$ arms sharing latent seeds against a common static reference and correlated the two deviations (ideal -1):

The decay is strongly variable-dependent. At day five the mass and thermodynamic fields retain roughly half of their antisymmetry (Z_{500} -0.487, MSLP -0.397, $T_{2\mathrm{m}}$ -0.407) while the wind components have lost essentially all of it ($U_{10\mathrm{m}}$ -0.047, $V_{10\mathrm{m}}$ -0.021). The response stays linear longer for the large-scale mass field than for the small-scale wind field, and that is the ordering in which the fields acquire spread (Sec. 4, *The short-lead deficit*).

Pairing does not improve calibration, and the standard diagnostic will say it does. Drawing members in $\pm z$ pairs is a variance reduction on the ensemble mean: with pair correlation $\rho < 0$ the variance of the mean falls from σ^2/M to $\sigma^2(1+\rho)/M$. The finite-ensemble correction $\sqrt{(M+1)/M}$ that makes a calibrated M-member ensemble read $\mathrm{SSR} = 1$ assumes independent members; for paired members the correct factor is $\sqrt{(M+1+\rho)/M}$, smaller when $\rho < 0$ by exactly enough to cancel the improvement in the mean. Properly corrected, the spread–skill ratio is invariant under antithetic sampling, and scored with the standard factor a paired ensemble reports a spurious calibration gain, about four percent at six hours and one percent at day five at the measured ρ. The benefit of pairing is skill, and it should be reported as such.

The training objective inherits the same assumption. Almost-fair CRPS estimates $\mathbb{E}|X - X'|$ by a pairwise average that is unbiased only under independence. For jointly Gaussian partners $\mathbb{E}|X - Y| = 2\sigma\sqrt{(1-\rho)/\pi}$, that is $\sqrt{1-\rho}$ times the independent-pair value. At $M = 8$ arranged as four pairs, four of the 28 unordered pairs are partners, so at the measured $\rho = -0.695$ the

Table 16: Shape of the spread-skill ratio as a function of lead, over all forty scored leads. *min* is the smallest value the profile takes and *at* the lead in hours where it takes it; *depth* is the fall from 6 h to that minimum and *rebound* the recovery from it to day 10. A monotone decay would put *at* at 240 and *rebound* at zero. The IC-only arm is U-shaped in every variable, deepest in the moisture and boundary-layer channels with minima between 42 and 66 h, whereas both retrofits leave at most a shallow dip at 12–18 h and, on Z_{500}, none at all. n=2 initialisations, 2024-01-05 and 2024-07-03. SSR is $\sqrt{(M{+}1)/M}\,\langle\sigma\rangle_{\mathrm{rms}}/\mathrm{RMSE}$, the convention a calibrated ensemble scores 1 on, averaged over initialisations; ratios are ratios of the two arms' means over initialisations. At n=2 no interval is quoted; these are the two pilot initialisations of the store and the table is a pilot-scale readout, not an annual estimate.

	6 h	min	at	10 d	depth	rebound
IC-only (frozen)						
$T_{2\mathrm{m}}$	0.719	0.439	42	0.749	0.280	0.310
$U_{10\mathrm{m}}$	0.943	0.716	30	0.856	0.227	0.140
$V_{10\mathrm{m}}$	0.952	0.716	30	0.860	0.236	0.144
MSLP	0.904	0.769	18	0.884	0.135	0.115
Z_{500}	0.853	0.710	66	0.807	0.143	0.097
T_{850}	0.934	0.609	54	0.779	0.325	0.169
U_{850}	0.906	0.685	30	0.845	0.221	0.159
Q_{700}	0.909	0.577	66	0.724	0.332	0.147
HyperENS $K{=}4 \times M{=}8$						
$T_{2\mathrm{m}}$	0.687	0.631	18	1.079	0.056	0.448
$U_{10\mathrm{m}}$	0.864	0.761	18	0.952	0.103	0.191
$V_{10\mathrm{m}}$	0.883	0.770	18	0.966	0.114	0.197
MSLP	0.793	0.763	18	0.978	0.031	0.216
Z_{500}	0.775	0.775	6	0.938	0.000	0.163
T_{850}	0.833	0.730	18	1.025	0.102	0.295
U_{850}	0.838	0.752	18	0.945	0.086	0.193
Q_{700}	0.812	0.735	18	0.986	0.077	0.251
HyperENS+ $K{=}4 \times M{=}8$						
$T_{2\mathrm{m}}$	0.723	0.716	12	1.110	0.007	0.394
$U_{10\mathrm{m}}$	0.915	0.843	18	0.979	0.072	0.136
$V_{10\mathrm{m}}$	0.905	0.828	18	1.009	0.077	0.181
MSLP	0.831	0.831	6	1.017	0.000	0.185
Z_{500}	0.856	0.856	6	0.995	0.000	0.139
T_{850}	0.854	0.789	18	1.046	0.066	0.257
U_{850}	0.891	0.829	18	0.973	0.061	0.144
Q_{700}	0.850	0.815	18	1.048	0.035	0.234

dispersion term reads about four percent high (six at $\rho = -1$). Since that term rewards spread, overstating it pays the model for spread it does not have, backwards for an under-dispersed model. The correction is exact and assumption-free: exclude the partner pairs from the dispersion sum, leaving the 24 cross-pairs. The antithetic arm below was trained with it.

The screening pair. Table 20 is the one-knob comparison: two trainings from the same cold start, arm A with i.i.d. latents and arm B with paired latents and the corrected estimator, both stopped early at eight to ten thousand gradient steps. Scored in pairs, arm B is at 0.971 of arm A's fair CRPS at six hours and 0.983 at day five, with 0.989 in the RMSE of the mean; scored with i.i.d. draws, the same weights are at 0.994 and 0.996. The third block isolates the sampling: pairs against i.i.d. draws of the same weights are worth 0.974 in RMSE at six hours and 0.991 at day five, most of what the first block shows. Training with paired latents changed the weights little; the pairing at inference is where the gain is; and the spread–skill rows show the spurious calibration gain of the previous paragraph, arm B reading higher in pairs than i.i.d. on identical weights. The arms of the paper draw i.i.d. latents at inference so that every instrument is used as written.

Table 17: Rank-histogram shape, from bin counts pooled over initialisations. $r_i = \sum_b |p_b - 1/(M+1)|$ measures distance from flat but cannot say which way; the other two columns do. *edge* is the outer-bin mass relative to the uniform share, so above 1 is a U (the truth falls outside the ensemble too often) and below 1 a dome; *tilt* is the upper half's mass minus the lower half's, and is positive when the truth ranks above the members, i.e. the forecast is biased low. The separation matters here: the IC-only and HyperENS arms carry a large positive tilt on Z_{500} that survives to day 10, which is a bias rather than a dispersion defect and which the spread-skill ratio cannot see. n=2 initialisations, 2024-01-05 and 2024-07-03. At n=2 no interval is quoted; these are the two pilot initialisations of the store and the table is a pilot-scale readout, not an annual estimate.

	1 d			10 d		
	r_i	edge	tilt	r_i	edge	tilt
IC-only (frozen)						
$T_{2\mathrm{m}}$	0.845	7.12	+0.423	0.811	7.45	+0.386
$U_{10\mathrm{m}}$	0.468	4.22	−0.017	0.340	3.22	−0.025
$V_{10\mathrm{m}}$	0.468	4.22	+0.007	0.314	3.04	+0.007
MSLP	0.295	3.04	+0.000	0.362	2.98	+0.048
Z_{500}	0.763	6.72	+0.495	0.870	7.65	+0.518
T_{850}	0.533	4.64	+0.207	0.617	5.26	+0.363
U_{850}	0.404	3.73	+0.019	0.316	2.89	−0.019
Q_{700}	0.497	4.53	+0.011	0.374	3.64	+0.057
HyperENS K=4 × M=8						
$T_{2\mathrm{m}}$	0.323	3.18	+0.208	0.258	1.24	+0.286
$U_{10\mathrm{m}}$	0.323	3.63	−0.013	0.114	1.60	−0.031
$V_{10\mathrm{m}}$	0.333	3.62	+0.078	0.068	1.42	+0.061
MSLP	0.250	2.78	+0.128	0.108	1.22	+0.123
Z_{500}	0.398	3.34	+0.403	0.481	2.58	+0.462
T_{850}	0.311	3.51	+0.056	0.200	1.29	+0.227
U_{850}	0.288	3.37	+0.009	0.099	1.43	−0.022
Q_{700}	0.297	3.41	+0.040	0.064	1.43	+0.046
HyperENS+ K=4 × M=8						
$T_{2\mathrm{m}}$	0.191	2.10	+0.040	0.228	0.65	+0.139
$U_{10\mathrm{m}}$	0.207	2.67	−0.007	0.077	1.26	−0.039
$V_{10\mathrm{m}}$	0.250	2.91	+0.075	0.059	1.10	+0.079
MSLP	0.152	2.08	−0.011	0.096	0.86	+0.010
Z_{500}	0.172	1.66	−0.034	0.159	1.07	+0.135
T_{850}	0.215	2.64	−0.028	0.107	0.92	+0.045
U_{850}	0.182	2.50	+0.009	0.063	1.12	−0.022
Q_{700}	0.209	2.64	+0.061	0.036	1.12	+0.054

ε	SSR 6 h	SSR d5	RMSE 6 h	RMSE d5	CRPS d5	variogram 6 h	variogram d5	spectral d5
0.00	0.000	0.000	1.000	1.000	1.000	1.000	1.000	−0.173
0.02	0.021	0.066	1.000	0.996	0.944	0.999	0.980	−0.173
0.10	0.334	0.541	1.061	0.966	0.732	1.021	0.814	−0.176
0.20	0.424	0.692	1.125	1.004	**0.723**	1.069	0.809	−0.178
0.40	0.588	0.858	1.266	1.101	0.771	1.139	0.873	−0.180
trained processor channel, at the same reference:								
—	0.483	0.685	1.011	**0.884**	**0.627**	—	**0.716**	—

Table 18: The random-direction perception probe, development configuration (harness paragraph above): the encoder's learned query latents perturbed by a random direction of relative amplitude ε on top of the static model, so that all measured spread is the probe's. Spread–skill ratio at 6 h and day 5 (development-scorecard convention, absolute); RMSE of the ensemble mean, fair CRPS and variogram score as ratios to the static model, nine-variable means of per-variable ratios; the spectral column is the mid-band spectral ratio at day five (development convention, absolute). The last row is the trained processor channel against the same reference, at matched day-5 spread.

lead	6 h	24 h	48 h	120 h
mean pair correlation	-0.695	-0.610	-0.510	-0.196
$\|\delta_+\|/\|\delta_-\|$ (ideal 1)	1.008	0.997	0.987	1.018

Table 19: Antithetic pairs through the frozen trunk, development configuration (App. F): correlation between the deviations of the $+z$ and $-z$ arms from a common static reference, mean over the nine development variables (ideal -1), and the ratio of their deviation norms (ideal 1), against lead.

Table 20: The antithetic screening pair. Two trainings from the same cold start that differ in one knob: arm A draws i.i.d. latents, arm B draws them in $\pm z$ pairs and excludes partner pairs from the dispersion term of the objective; both stopped after eight to ten thousand gradient steps. Scored on 154 daily initialisations of 2019 (1 January–3 June), $M{=}8$, to day five, on the development harness of App. F. Ratios are cross-variable means of per-variable ratios over the eight variables of Sec. 3.3; below one is better. The spread–skill ratio is the mean-of-standard-deviation estimator these artifacts carry, comparable between the rows and not with the RMS-spread ratio of the main text. No interval is given: the artifacts store no per-initialisation values.

	6 h	1 d	3 d	5 d
arm B scored in pairs / arm A				
RMSE of the ensemble mean	0.983	0.986	0.985	0.989
fair CRPS	0.971	0.974	0.977	0.983
spread	1.149	1.139	1.076	1.041
arm B scored i.i.d. / arm A				
RMSE of the ensemble mean	1.009	1.008	1.000	0.998
fair CRPS	0.994	0.993	0.994	0.996
spread	1.103	1.098	1.046	1.021
arm B scored in pairs / arm B scored i.i.d. (same weights)				
RMSE of the ensemble mean	0.974	0.978	0.985	0.991
fair CRPS	0.977	0.981	0.983	0.987
spread	1.042	1.038	1.028	1.020
spread–skill ratio, legacy estimator (absolute)				
arm A	0.528	0.578	0.659	0.658
arm B, i.i.d.	0.570	0.629	0.688	0.672
arm B, paired	0.609	0.664	0.714	0.690