PHYSICS-GUIDED EXPERT SELECTION FOR WEATHER FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Heterogeneous AI weather forecasters excel under different atmospheric states and locations, creating complementary strengths that fusion is expected to exploit. Yet existing fusion often produces a compromise between experts and may dilute the locally better forecast. To better exploit model complementarity, we introduce Physics-guided Expert Selection (PES), a lightweight plug-and-play layer for spatial expert selection. We construct three physically interpretable diagnostics: advection, expert disagreement, and their joint magnitude. The diagnostics reveal where expert selection is most beneficial and turn black-box routing into gray-box selection. Rather than merely interpolating between two experts, PES outperforms the stronger constituent expert across all 40 tasks, reducing RMSE by up to 6.3% with only 110K additional parameters and 50 ms per forecast. The learned weights form weather-dependent expert-preference regions rather than near-equal mixtures. Across additional heterogeneous expert pairs, PES again outperforms every constituent expert, showing that the selection framework transfers across different expert combinations.

1 INTRODUCTION

Recent data-driven and hybrid weather models, including FourCastNet (Pathak et al., 2022), Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), FuXi (Chen et al., 2023), FengWu (Chen et al., 2025), and NeuralGCM (Kochkov et al., 2024), achieve strong medium-range forecasting skill. They differ in architecture, training, and physical formulation, so their relative skill varies across variables, pressure levels, lead times, and weather regimes (Rasp et al., 2024). A model that performs best on average may be less reliable at a particular location or under a particular atmospheric state. This heterogeneity motivates selecting among existing forecasters according to the current forecast state.

Classical forecast combination estimates weights from model diversity (Glahn & Lowry, 1972; Krishnamurti et al., 1999; Raftery et al., 2005), while mixture-of-experts (MoE) methods learn input-dependent allocation (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022). Recent weather methods also use adaptive weighting or specialized experts (Chakraborty et al., 2025; Gan et al., 2025). For a given weather state and location, one expert may be more accurate than the other. Although forecast averaging can benefit from error cancellation, it may dilute the locally better forecast by mixing it with the less reliable one. This motivates weather-state-dependent expert selection: learning which expert is more likely to be reliable and assigning it greater weight at each location.

To address this problem, we propose Physics-guided Expert Selection (PES), a lightweight framework for learning spatial expert preferences. We instantiate PES with Pangu-Weather and NeuralGCM and use physically interpretable forecast-time diagnostics to determine where each expert should receive greater weight. PES operates only on their aligned forecast products while keeping both experts frozen, providing a plug-and-play selection layer without accessing expert internals or rerunning them.

PES introduces three physically interpretable diagnostic fields: signed horizontal advection $\mathcal{A}$, expert disagreement Δ, and joint difficulty Ξ. $\mathcal{A}$ describes local transport, Δ captures the direction and magnitude of inter-expert differences, and Ξ measures their joint standardized magnitude to indicate where expert selection is likely to be most useful. By exposing the routing decision through

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

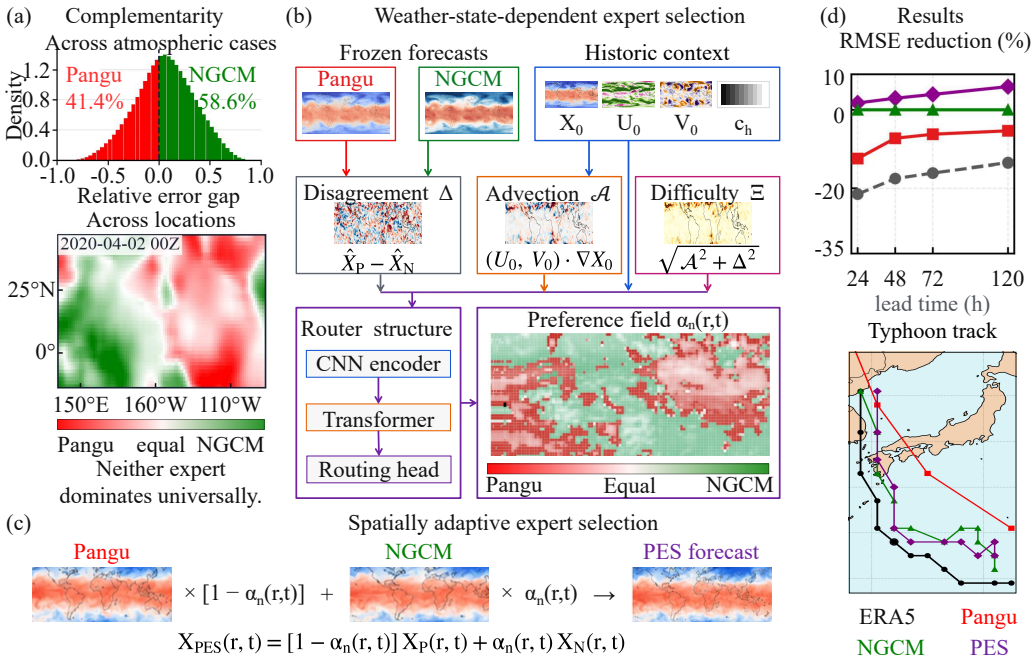

Figure 1: **Framework of Physics-guided Expert Selection.** (a) Pangu-Weather and NeuralGCM show complementary skill across atmospheric cases and locations; neither expert dominates universally. (b) Forecast-time context and physics-guided diagnostics are mapped by a lightweight router to the spatial preference field $\alpha_N(\mathbf{r}, t)$. (c) The learned preference field implements spatially adaptive soft selection between the two frozen expert forecasts. (d) PES outperforms the strongest constituent expert and predicts the typhoon track more accurately.

these explicit diagnostics, PES turns an otherwise black-box mapping from weather states to expert weights into a gray-box selection mechanism. A lightweight CNN-Transformer router maps the forecast-time inputs to a spatially varying expert-preference field, allowing the preferred expert to vary across locations and forecast cases. All routing inputs are available before verification and require no future observations.

Across 40 variable-level-lead-time tasks, PES outperforms the stronger expert in every task, reducing RMSE by 1.2%–6.3% with only 110K routing parameters per task and approximately 50 ms per variable-lead field. Fixed averaging, static preferences, and generic state-based routing do not recover the full benefit of weather-state-dependent selection. A router using only $\mathcal{A}$, Δ, and Ξ nearly matches the full PES router, which additionally receives the raw atmospheric fields, showing that these explicit diagnostics are sufficient to support effective and interpretable expert selection. The learned preference fields form clear weather-dependent expert-preference regions rather than remaining near equal mixing. Comparable gains after replacing either expert with GraphCast further show that PES transfers across heterogeneous weather-model pairs.

Our main contributions are:

- **Problem reframing:** We recast heterogeneous forecast fusion as weather-state-dependent expert selection, learning which frozen forecaster to trust at each location.
- **Physics-guided PES:** We introduce a lightweight, plug-and-play selection layer that improves forecast accuracy across all 40 tasks.
- **Interpretable gray-box mechanism:** We introduce physically meaningful diagnostics that turn black-box routing into gray-box expert selection.

Table 1: Three-day Z500 forecast skill relative to each study's corresponding IFS baseline and model size. The PES count denotes only the 110K task-specific marginal routing parameters.

Model	Reported 3-day Z500 RMSE/IFS (%) ↓	Parameters
FourCastNet (Pathak et al., 2022)	163.0	64M
FengWu (Chen et al., 2025)	100.0	158M
Pangu-Weather (Bi et al., 2023)	98.5	256M
GraphCast (Lam et al., 2023)	91.9	36.7M
FuXi (Chen et al., 2023)	92.6	4.7B
AIFS (Lang et al., 2024)	97.6	253M
NeuralGCM (Kochkov et al., 2024)	85.2	18.3M
Aurora (Bodnar et al., 2025)	96.3	1.3B
GenCast (ensemble mean) (Price et al., 2025)	96.3	57M
PES	**82.4**	**+110K**

2 Related Work

Data-driven weather forecasting. Recent data-driven and hybrid systems, including FourCastNet, Pangu-Weather, GraphCast, FuXi, FengWu, and NeuralGCM, achieve strong forecasting skill through different architectures, training strategies, and physical formulations (Pathak et al., 2022; Bi et al., 2023; Lam et al., 2023; Chen et al., 2023; 2025; Kochkov et al., 2024). PES does not modify these forecasting models; instead, it treats heterogeneous systems as frozen experts and studies how their relative skill varies across weather states and locations.

Forecast combination. Classical methods combine forecasts through estimated weights (Glahn & Lowry, 1972; Krishnamurti et al., 1999; Raftery et al., 2005), while mixture-of-experts methods learn input-dependent expert allocation (Jacobs et al., 1991). Recent weather-specific methods include adaptive weighting over pretrained forecasts, as in MoWE (Chakraborty et al., 2025), and expert specialization within a forecasting architecture, as in EWMoE (Gan et al., 2025). These approaches primarily optimize mixture accuracy or expert specialization. PES instead formulates the task as weather-state-dependent expert selection and learns a spatial expert preference from forecast-time diagnostics.

3 Method

PES operates on aligned frozen forecast products without accessing expert internals. We instantiate the notation with Pangu-Weather and NeuralGCM and learn only their spatial expert preference, as summarized in Figure 1.

3.1 Problem formulation: weather-state-dependent expert selection

For an initial state $\mathbf{x}$ and lead time t, frozen experts Pangu-Weather (f_P) and NeuralGCM (f_N) produce

$$\hat{\mathbf{X}}_P = f_P(\mathbf{x}, t), \qquad \hat{\mathbf{X}}_N = f_N(\mathbf{x}, t). \tag{1}$$

Rather than learning a generic fusion rule, PES predicts a spatially varying expert preference field,

$$\alpha = \sigma(g_\theta(\mathbf{z})), \tag{2}$$

where $\mathbf{z}$ contains forecast-time selection diagnostics and $\alpha \in [0, 1]$ denotes preference toward NeuralGCM. PES implements this preference through soft routing,

$$\hat{\mathbf{X}} = \alpha \hat{\mathbf{X}}_N + (1 - \alpha)\hat{\mathbf{X}}_P. \tag{3}$$

Thus, α is a continuous expert-preference variable.

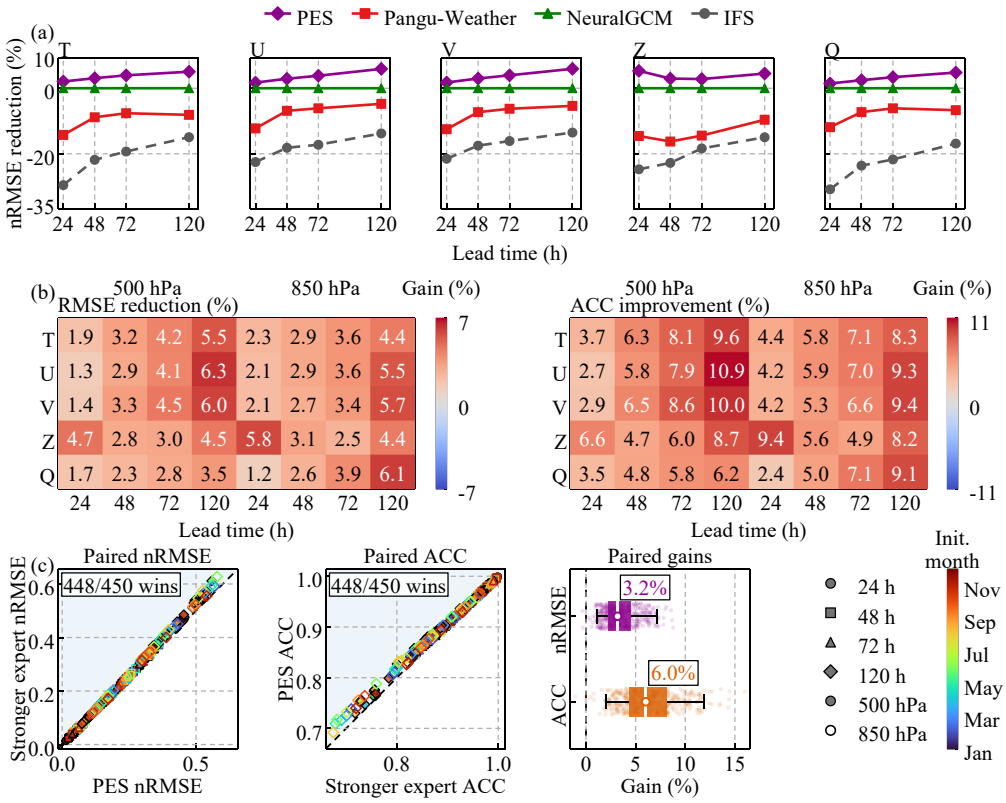

Figure 2: **PES achieves consistent improvements over heterogeneous weather experts.** Performance gains remain robust across variables, pressure levels, lead times, and individual test cases.

3.2 Forecast-time diagnostics for expert selection

PES uses three forecast-time diagnostics constructed without future verification: signed horizontal advection $\mathcal{A}$ for local dynamics, signed expert disagreement Δ for heterogeneous response, and Ξ for their joint magnitude. Let X_0 denote the target field at initialization time t_0, with horizontal winds U_0 and V_0. We define signed horizontal advection as

$$\mathcal{A} = -\left(U_0 \frac{\partial X_0}{\partial x} + V_0 \frac{\partial X_0}{\partial y}\right), \tag{4}$$

where x and y are horizontal directions. $|\mathcal{A}|$ measures transport activity and its sign preserves the advective tendency, characterizing the local dynamical regime in which disagreement occurs.

We define signed expert disagreement at the target lead time as

$$\Delta = \widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}, \tag{5}$$

where $\widehat{X}_{\mathrm{P}}$ and $\widehat{X}_{\mathrm{N}}$ are the two expert forecasts. $|\Delta|$ measures their local separation and its sign preserves its direction. The triangle inequality guarantees that at least one local absolute error is no smaller than $|\Delta|/2$, so larger disagreement makes selection more consequential without revealing the better expert.

For each variable-lead-time task, X_0, U_0, V_0, $\mathcal{A}$, and Δ are standardized using training-set statistics, with tildes denoting standardized fields. We define

$$\Xi = \sqrt{\widetilde{\mathcal{A}}^2 + \widetilde{\Delta}^2 + \epsilon}, \tag{6}$$

where ϵ ensures numerical stability. Ξ is a deterministic, parameter-free indicator of the joint standardized magnitude of transport activity and expert disagreement, not a calibrated uncertainty estimate or expert label.

The initialization hour h_0 is encoded as

$$c_h = \cos\left(\frac{2\pi h_0}{24}\right), \tag{7}$$

with 00 and 12 UTC mapped to $+1$ and -1. The seven-channel input is

$$\mathbf{z} = \left[\widetilde{X}_0,\, c_h,\, \widetilde{U}_0,\, \widetilde{V}_0,\, \widetilde{\mathcal{A}},\, \widetilde{\Delta},\, \Xi\right]. \tag{8}$$

A separate router is trained per variable-lead task, so lead time is encoded by task identity and future verification never appears in $\mathbf{z}$.

Controlled comparisons supporting the signed forms of $\mathcal{A}$ and Δ and the parameter-free definition of Ξ are provided in Appendix I.

3.3 Learning expert preference through soft routing

A lightweight spatial router maps the diagnostic representation to the NeuralGCM preference field,

$$\alpha_{\mathrm{N}} = \sigma(g_\theta(\mathbf{z})), \tag{9}$$

where convolutional encoding captures local spatial structures, self-attention supplies non-local atmospheric context, and the sigmoid function constrains the preference score to $[0, 1]$. The output has the same spatial resolution as the forecast field, allowing PES to learn a mapping from local atmospheric conditions to local expert preference rather than assigning one global trust level to the entire forecast.

The router is optimized by minimizing latitude-weighted mean squared error between the soft-selected forecast and ERA5 verification,

$$\mathcal{L}(\theta) = \frac{\sum_i m_i w_i \left(\widehat{X}_{\mathrm{PES},i} - X_i^*\right)^2}{\sum_i m_i w_i}, \qquad w_i \propto \cos\phi_i, \tag{10}$$

where m_i marks valid points and ϕ_i is latitude. Verification-time ERA5 is used only for supervision and evaluation, never as router input. The weight remains a continuous expert preference, and routing is applied to pre-computed forecasts without feedback into either expert. Implementation details are in Appendix C.

4 Results

Evaluation protocol. We align 2020 ERA5, Pangu-Weather, and NeuralGCM forecasts from WeatherBench2 on a common $1.5°$, 121×240 grid. Five variables at 500 and 850 hPa and four lead times define 40 independently routed tasks. For each month, the third-to-last day is used for checkpoint selection, the final two days for testing, and the remainder for training. We report latitude-weighted RMSE and ACC against ERA5 with training-only normalization; IFS provides an external reference. Full protocol and leakage controls are given in Appendix A.

4.1 Overall Performance

We first evaluate whether a globally fixed expert choice is sufficient, or whether expert preference needs to adapt to individual forecast states and spatial locations. Figure 2 compares PES with both constituent experts and the task-wise stronger expert, which is determined retrospectively by aggregate test-set RMSE for each variable-level-lead-time task and then applied as a global expert choice over all forecast samples and grid points within that task.

Across all 40 variable-level-lead-time tasks, PES consistently outperforms the stronger expert, reducing RMSE by 1.2%-6.3% and improving ACC by 2.4%-10.9%. The improvement is also robust across different months, with PES winning 99.6% of paired monthly comparisons. Paired bootstrap

Table 2: **Routing ablations on Z500 at 120-h lead time.** Forecast improvements and routing-field characteristics are compared across different expert selection strategies.

Method	α_{static}	RMSE gain over Blind (pp)	H_α	V_α	G_α
50/50 simple average	0.500	-0.845	0.6931	0.0000	0.0000
Global static prior	0.668	$+0.004$	0.6357	0.0000	0.0000
Fixed spatial prior	-	-0.365	0.5533	0.0160	0.0310
Blind	-	0.000	0.6100	0.0034	0.0124
Raw forecasts (P, N)	-	-0.189	0.5671	0.0043	0.0105
Δ only	-	$+0.079$	0.5556	0.0129	0.0359
$\mathcal{A}$ only	-	-0.019	0.6105	0.0058	0.0152
w/o Ξ	-	$+0.124$	0.5848	0.0056	0.0223
Full PES	-	$\underline{+0.146}$	0.5475	0.0132	0.0420
Diagnostic only	-	$\mathbf{+0.235}$	0.4782	0.0153	0.0331

evaluation with 20,000 resamples further confirms statistically significant improvements for all 40 tasks, with the 95% confidence intervals of both RMSE reduction and ACC gain remaining above zero.

To examine whether PES generalizes beyond a specific expert pair, we repeat the complete 40-task protocol for two additional combinations: GraphCast–NeuralGCM (GN) and Pangu-Weather–GraphCast (PG). For both pairs, PES improves upon the task-wise stronger constituent expert across all 40 variable-level-lead-time tasks. The complete task-level and paired-comparison results are reported in Appendix E, demonstrating that PES transfers across heterogeneous weather foundation model combinations.

These results show that globally preferred experts cannot capture the complementary strengths of heterogeneous models, and adaptive preference provides additional gains. Whether the learned routing reflects meaningful weather-state-dependent expert selection, rather than only generic forecast combination effects, is examined in the following section.

4.2 What Information Drives Expert Selection?

A key challenge in evaluating expert routing is that improved forecast accuracy does not necessarily indicate that the model has learned when and where each expert is more reliable. Even a fixed combination can reduce forecast error through error compensation. Therefore, we compare PES with several controls that improve forecasts without explicitly modeling weather-dependent expert preference.

Table 2 summarizes the forecast gains and routing-field characteristics of these controls and diagnostic ablations for Z500 at 120 h. A parameter-free 50/50 average already reduces RMSE by 3.433% relative to the stronger expert. Similarly, a global static prior, a fixed spatial prior, and the Blind router achieve gains of 4.282%, 3.913%, and 4.278%, respectively. These results show that a considerable portion of combination benefit can be obtained without learning which expert is more reliable under specific weather conditions. Full PES further improves the RMSE gain to 4.424%.

These gaps are not artifacts of the representative task: across all 40 variable-level-lead-time tasks, Full PES attains a lower RMSE than the capacity-matched Blind router on $40/40$ tasks, with a mean additional reduction of 0.272% relative to that router and paired bootstrap intervals excluding zero on $36/40$ tasks. These results indicate that adaptive expert preference provides additional information beyond fixed combination strategies to make PES more accurate.

To understand what information enables expert selection, we analyze the role of individual routing diagnostics. Expert selection requires answering two different questions: where selection is necessary and how the preference should be determined. Expert disagreement Δ provides a direct measure of the first question. The Δ-only router achieves a 4.357% RMSE gain, approaching the full PES performance of 4.424%, indicating that regions with substantial disagreement between experts are where adaptive selection becomes valuable.

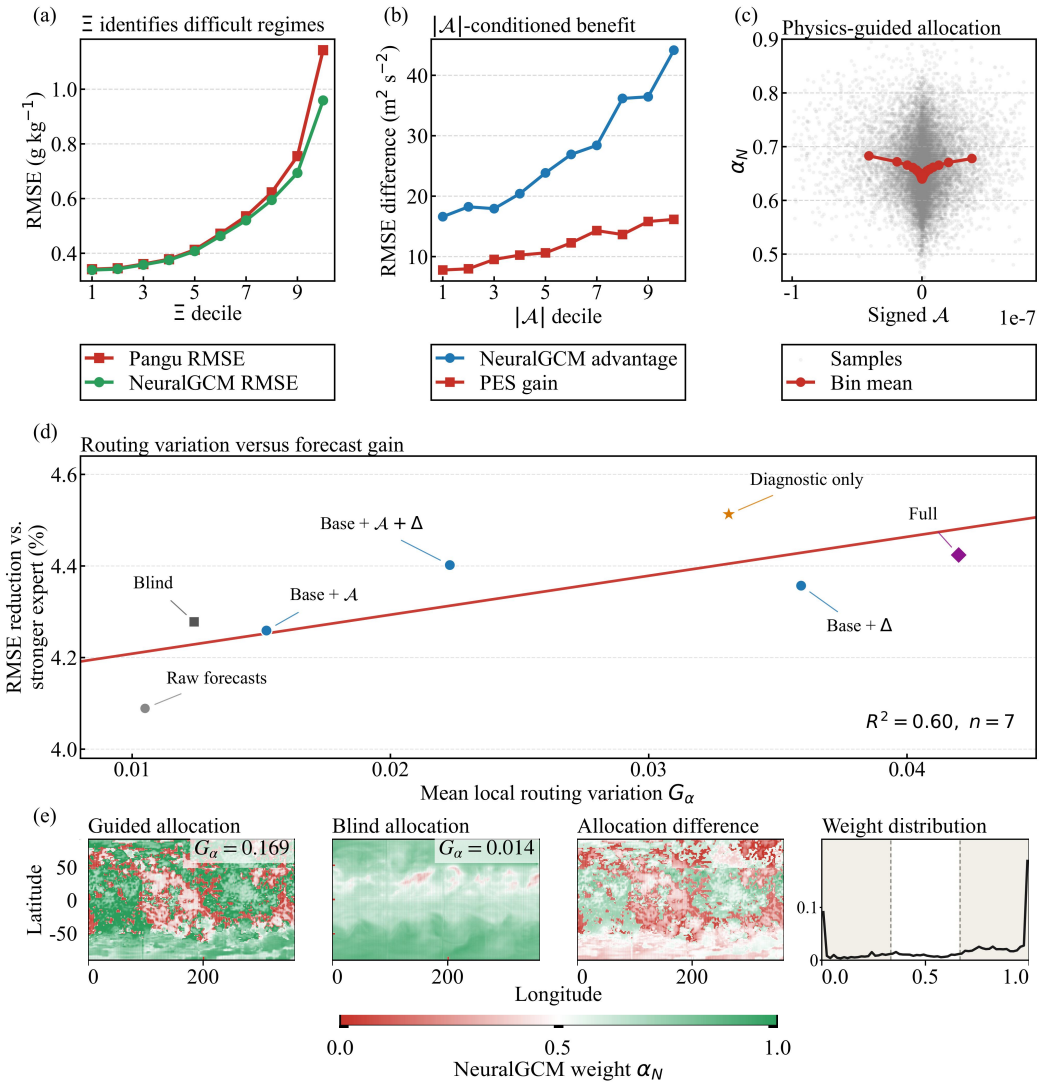

Figure 3: **Interpretation from physics-guided expert selection.** (a) Constituent-expert RMSE increases with the joint difficulty indicator Ξ. (b) The relative advantage of NeuralGCM and the PES gain increase with advection magnitude $|A|$. (c) The learned NeuralGCM preference varies systematically with signed advection A. (d) Across routing variants, greater local routing variation is associated with larger RMSE reductions. (e) Physics-guided routing forms spatially structured preference regions, with most weights lying away from the equal-mixing value of 0.5.

However, disagreement alone only reveals that experts differ and does not determine which expert should be preferred. The remaining diagnostics provide the contextual information required to resolve this ambiguity. The signed horizontal advection $\mathcal{A}$ describes the local atmospheric transport state and determines how expert preference changes across different regimes. As shown in Figure 3, the learned NeuralGCM preference varies systematically with $\mathcal{A}$, and PES obtains larger gains in regions with stronger advective signals. The joint difficulty Ξ summarizes the standardized magnitudes of advection and expert disagreement, highlighting regions where strong transport and forecast differences make expert selection most useful.

Together, $\mathcal{A}$, Δ, and Ξ form a compact diagnostic representation that describes where selection is needed, how preference changes, and how difficult the decision is. Beyond identifying the information used for routing, we further examine whether the learned preference fields exhibit the properties expected from effective expert selection. A useful selection mechanism should not produce arbitrary

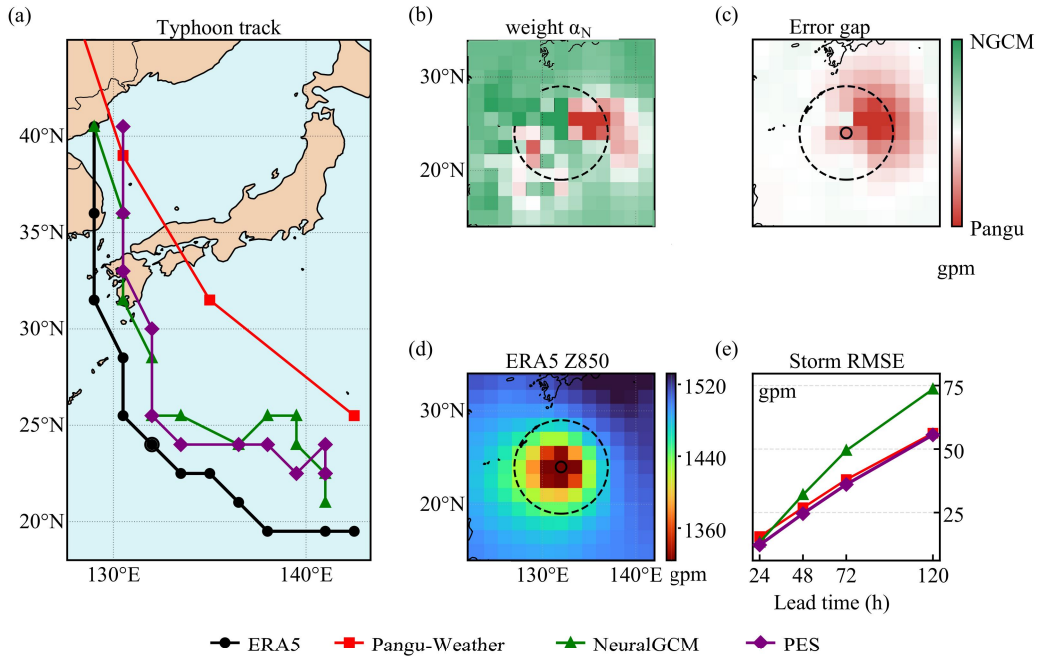

Figure 4: **Spatially adaptive expert selection improves forecasting during Typhoon Haishen.** (a) Forecast tracks from ERA5, Pangu-Weather, NeuralGCM, and PES. (b) Learned NeuralGCM allocation weight α_N. (c) Difference between the absolute errors of the two experts, showing their complementary spatial skills. (d) ERA5 Z850 field. (e) Z850 RMSE within 5° of the ERA5 storm center. All maps show the 120-h forecast.

spatial variation; instead, it should exploit sufficient information to form sharper and more structured expert-preference maps.

We therefore quantify the relationship between the structure of the learned preference field and forecast improvement. Across the seven state-dependent routers, the spatial routing variation G_α is positively associated with PES gain, with $R^2 = 0.60$ and Pearson correlation $r = 0.77$. This indicates that routers producing more structured and spatially differentiated preference maps generally achieve larger RMSE reductions. The spatial examples in Figure 3 further illustrate this behavior: PES forms distinct preference regions corresponding to different expert advantages rather than applying a uniform global weighting. In the representative case shown in Figure 3(e), 82% of latitude-weighted grid points fall within the two outer thirds of the preference range, showing that the learned weights predominantly favor one expert rather than remaining near equal mixing.

Finally, we test whether the explicit diagnostics can themselves support expert selection, rather than serving only as auxiliary features for a router driven by raw atmospheric fields. We compare the full PES router with a capacity-matched diagnostic-only router that removes the raw state variables X_0, U_0, and V_0, while retaining only $[c_h, \mathcal{A}, \Delta, \Xi]$. On the representative Z500@120h task, the diagnostic-only router achieves a 4.513% RMSE gain, exceeding both the Blind router (4.278%) and Full PES (4.424%). Extending the same formulation to all 40 variable-level-lead-time tasks improves upon the stronger expert in all 40/40 tasks, with a mean RMSE reduction of 3.454%. Its overall performance remains comparable to Full PES, and it exceeds Full PES on approximately one-third of the tasks.

These results show that explicit physical diagnostics are sufficient to drive effective expert selection without relying on an opaque encoding of the full atmospheric state. The performance of diagnostic-only routing therefore supports the gray-box interpretation of PES: its expert preferences can be understood through atmospheric transport, expert disagreement, and their joint difficulty.

4.3 Case Study: Typhoon Haishen

Figure 4 illustrates state-dependent expert selection during Typhoon Haishen. The two frozen experts exhibit complementary spatial advantages: their absolute error difference varies across the storm region, indicating that neither model is uniformly superior during the event. PES responds to this heterogeneous skill distribution by producing a spatially varying allocation field α_N.

In particular, regions where Pangu-Weather achieves lower retrospective errors correspond to reduced NeuralGCM preference, despite PES having no access to future verification during inference. This adaptive allocation allows PES to achieve the lowest storm-region Z850 RMSE among the compared forecasts across the evaluated lead times.

4.4 Lightweight Selection over Frozen Forecast Products

PES operates only after the constituent forecasts are available. Each task-specific router contains 110K trainable parameters, while all 40 routers total 4.41M parameters; training requires approximately 10 GPU-hours on one RTX 4090D, and inference adds about 50 ms per global variable-lead field. Table 1 provides literature-level context for three-day Z500 skill and model scale. PES reaches 82.4% of the corresponding IFS RMSE with only 110K task-specific marginal parameters, emphasizing its role as a lightweight selection layer rather than a standalone forecasting model.

5 Limitations

Scope of the expert pool. We evaluate PES with two deterministic experts and pairwise convex routing. This setting is sufficient to study state-dependent complementarity, but it does not establish how PES behaves when the expert pool contains multiple models with overlapping or highly correlated errors. A concrete extension is to replace the binary sigmoid gate with a K-way simplex gate, using a shared spatial encoder together with expert-specific forecast features or pairwise disagreement channels. For larger expert pools, top-k sparse routing and a load-balancing regularizer could control inference cost and reduce the risk of routing collapse toward a single expert. These changes preserve the current forecast-time diagnostic formulation, but require additional analysis of routing calibration and optimization stability.

Uncertainty in expert preference. PES currently produces a deterministic expert-preference field without explicitly quantifying uncertainty in the routing decision. This limitation may be important when the two experts have similar expected skill, during rare weather events, or under distribution shift, where a decisive preference may not be well supported by the available forecast-time information. Future work could introduce Bayesian routing or ensemble-based uncertainty estimation to learn a posterior distribution over expert preference rather than a single weight. The resulting uncertainty could support calibrated selection, conservative mixing when confidence is low, or abstention and fallback strategies in operational forecasting systems.

6 Conclusion and Outlook

This work frames heterogeneous weather forecasting as weather-state-dependent expert selection. PES keeps existing forecasting models frozen and learns when and where different experts should be trusted. Across diverse variables, pressure levels, forecast horizons, and expert pairs, PES consistently improves over individual experts and static combination strategies, demonstrating that adaptive selection can effectively exploit complementary capabilities among heterogeneous weather foundation models. Together with the compact diagnostic representation and cross-model evaluations, these results demonstrate that expert selection can be achieved through a lightweight and physically interpretable gray-box decision space.

More broadly, this work suggests another axis for improving weather forecasting: future progress may come not only from developing increasingly powerful individual models, but also from organizing diverse forecasting capabilities through adaptive selection. As data-driven, physics-hybrid, regional, and event-specialized forecasting models continue to emerge, expert selection provides a potential interface for integrating heterogeneous forecasting systems according to their strengths.

Ethics Statement

All authors have read and agree to abide by the ICLR Code of Ethics. The study uses meteorological fields and model forecasts and involves no human participants, personal data, or private data. PES is a research method rather than an operational warning system; real-world deployment would require expert review and appropriate uncertainty, robustness, and operational validation.

Reproducibility Statement

The main paper and appendices document the PES formulation, routing inputs, data and splits, preprocessing, architecture, hyperparameters, checkpoint selection, evaluation protocol, complete results, and ablations. The anonymized supplement provides the configurations and scripts for preprocessing, training, inference, evaluation, and figure generation.

AI Use Statement

Generative AI tools were used for limited translation, language polishing, and implementation support. The research questions, methodology, experimental design, data processing, result interpretation, and scientific conclusions were developed by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of this work.

References

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(7970):533–538, 2023. https://www.nature.com/articles/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, 2025. https://www.nature.com/articles/s41586-025-09005-y.

Dibyajyoti Chakraborty, Romit Maulik, Peter Harrington, Dallas Foster, Mohammad Amin Nabian, and Sanjay Choudhry. MoWE: A mixture of weather experts. *arXiv preprint arXiv:2509.09052*, 2025. https://arxiv.org/abs/2509.09052.

Kang Chen, Tao Han, Fenghua Ling, Junchao Gong, Lei Bai, Xinyu Wang, Jing-Jia Luo, Ben Fei, Wenlong Zhang, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, Yuanzheng Ci, Bin Li, Xiaokang Yang, and Wanli Ouyang. The operational medium-range deterministic weather forecasting can be extended beyond a 10-day lead time. *Commun. Earth Environ.*, 6:518, 2025. https://www.nature.com/articles/s43247-025-02502-y.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Clim. Atmos. Sci.*, 6(1):190, 2023. https://www.nature.com/articles/s41612-023-00512-1.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *J. Mach. Learn. Res.*, 23(120):1–39, 2022. https://jmlr.org/papers/v23/21-0998.html.

Lihao Gan, Xin Man, Chenghong Zhang, and Jie Shao. EWMoE: An effective model for global weather forecasting with mixture-of-experts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 210–218, 2025. https://ojs.aaai.org/index.php/AAAI/article/view/31997.

Harry R. Glahn and Dale A. Lowry. The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.*, 11(8):1203–1211, 1972. https://journals.ametsoc.org/view/journals/apme/11/8/1520-0450_1972_011_1203_tuomos_2_0_co_2.xml.

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 Abellán, 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, Gábor Radnóti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sébastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.*, 146(730):1999–2049, 2020. https://rmets.onlinelibrary.wiley.com/doi/10.1002/qj.3803.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, 1991. https://direct.mit.edu/neco/article/3/1/79/5560/Adaptive-Mixtures-of-Local-Experts.

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

T. N. Krishnamurti, C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, Sulochana Gadgil, and Sajani Surendran. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550, 1999. https://www.science.org/doi/10.1126/science.285.5433.1548.

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. https://www.science.org/doi/10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Matthew Chantry, Jesper Dramsch, Florian Pinault, Baudouin Raoult, Mariana C. A. Clare, Christian Lessig, Michael Maier-Gerber, Linus Magnusson, Zied Ben Bouallègue, Ana Prieto Nemesio, Peter D. Düben, Andrew Brown, Florian Pappenberger, and Florence Rabier. AIFS - ECMWF's data-driven forecasting system. *arXiv preprint arXiv:2406.01465*, 2024. https://arxiv.org/abs/2406.01465.

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. *arXiv preprint arXiv:2202.11214*, 2022. 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, 2025. https://www.nature.com/articles/s41586-024-08252-9.

Adrian E. Raftery, Tilmann Gneiting, Fadoua Balabdaoui, and Michael Polakowski. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133(5):1155–1174, 2005. https://journals.ametsoc.org/view/journals/mwre/133/5/mwr2906.1.xml.

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 Bouallègue, Peter Düben, 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. *J. Adv. Model. Earth Syst.*, 16(6):e2023MS004019, 2024. https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023MS004019.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. https://openreview.net/forum?id=B1ckMDqlg.

APPENDIX CONTENTS

Notation and Symbols **13**

A Experimental Protocol and Leakage Control **15**

B Evaluation Metrics and Derived Statistics **15**

C Physical Motivation, Diagnostic Construction, and Router Implementation **17**

D Complete Forecast Results **22**

E Additional Expert Pairs for Cross-Model Generalization **28**

F Details of Expert-Selection Baselines and Ablation Variants **31**

G Compact Diagnostic Representation **33**

H Validation of Forecast Difficulty and Difficulty-Dependent Selection Benefit **34**

I Sensitivity to the Definitions of Routing Diagnostics **38**

J Oracle Upper Bound of Convex Expert Selection **40**

K Model Scale and Marginal Cost Scope **41**

NOTATION AND SYMBOLS

Symbol	Definition
Forecast fields, coordinates, and indices	
$X(\lambda, \phi, t)$	Generic prognostic atmospheric target at longitude λ, latitude ϕ, and time t.
$T,\ Q,\ U,\ V,\ Z$	Temperature, specific humidity, zonal wind, meridional wind, and geopotential, respectively; a pressure-level suffix (e.g., $Z500$) denotes the corresponding variable at that pressure level.
$X_0,\ U_0,\ V_0$	Target field and zonal and meridional winds at forecast initialization t_0.
$\mathbf{u}_0 = (U_0, V_0)$	Initialization-time horizontal wind vector.
$X^*,\ X^*_{n,i}$	ERA5 verification field; the indexed form denotes initialization n and grid point i.
$\widehat{X}^{(m)}$	Forecast field produced by model m.
$\widehat{X}_{\mathrm{P}}, \widehat{X}_{\mathrm{N}}, \widehat{X}_{\mathrm{PES}}$	Pangu-Weather, NeuralGCM, and PES forecasts, respectively.
$\mathbf{r},\ t_0,\ \tau$	Spatial location, forecast initialization time, and forecast lead time.
$\lambda,\ \phi,\ R_{\mathrm{E}},\ \nabla_s$	Longitude, latitude, mean Earth radius, and horizontal gradient on the spherical surface.
$n,\ i,\ N,\ G$	Forecast-case index, grid-point index, number of forecast cases, and number of spatial grid points.
$H,\ W$	Meridional and zonal dimensions of a gridded field.
$\mathrm{P}, \mathrm{N}, \mathrm{IFS}$	Pangu-Weather, NeuralGCM, and the operational IFS, respectively.
$m,\ b,\ m^\star$	Generic model, reference model, and task-level stronger constituent expert.
Physics-guided routing and implementation	
$\alpha_{\mathrm{N}},\ \alpha_{n,i}$	NeuralGCM allocation weight in $[0, 1]$; the indexed form denotes case n and grid point i.
$h_0,\ c_h$	UTC initialization hour and its cyclic encoding $c_h = \cos(2\pi h_0/24)$.
$\mathcal{A}$	Signed horizontal-advection tendency, $-\mathbf{u}_0 \cdot \nabla_s X_0$.
Δ	Signed inter-expert disagreement, $\widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}$.
$Y,\ \widetilde{Y}$	A generic continuous routing channel and its training-statistics standardized form.
$\mu_Y^{\mathrm{train}},\ \sigma_Y^{\mathrm{train}}$	Training-subset mean and standard deviation used to standardize channel Y.
$\Xi,\ \epsilon$	Deterministic local forecast-difficulty score and a small numerical stabilizer.
$\mathbf{z}$	Seven-channel routing input $[\widetilde{X}_0, c_h, \widetilde{U}_0, \widetilde{V}_0, \widetilde{\mathcal{A}}, \widetilde{\Delta}, \Xi]$.
$g_\theta,\ \theta,\ \sigma(\cdot)$	Routing network, its trainable parameters, and the sigmoid function.
$\mathcal{L}(\theta)$	Latitude-weighted mean-squared-error training objective.
$m_i,\ w_i$	Valid-grid mask and latitude weight used in training; $w_i \propto \cos\phi_i$.
Evaluation and diagnostic statistics	
$M_{n,i},\ M^{\mathrm{ACC}}_{n,i}$	Common finite-value verification mask and the ACC mask, which additionally requires finite climatology.
$C_{n,i}$	WeatherBench2 climatological field matched to the valid day and UTC hour.
$f'^{(m)}_{n,i},\ o'_{n,i}$	Forecast and observed anomalies relative to $C_{n,i}$.
$\overline{f}'^{(m)}_n,\ \overline{o}'_n$	Latitude-weighted spatial means of the forecast and observed anomalies.
RMSE, ACC, nRMSE	Root-mean-square error, anomaly correlation coefficient, and RMSE normalized by the latitude-weighted standard deviation of the corresponding ERA5 field.
$N_{\mathrm{valid}},\ \mathcal{I}_{\mathrm{valid}}$	Number and index set of forecast cases with valid ACC values.
$I^{(b)}_{\mathrm{RMSE}},\ G^{(b)}_{\mathrm{ACC}}$	RMSE reduction and fraction of the remaining ACC gap to unity closed by PES, relative to reference b.
$W^{(b)}_{\mathrm{RMSE}},\ W^{(b)}_{\mathrm{ACC}}$	Paired case-level RMSE and ACC win rates against reference b.
$\mathbb{I}[\cdot]$	Indicator function.
$\overline{\alpha}_n,\ H_\alpha,\ V_\alpha,\ G_\alpha$	Case-wise mean routing weight, mean binary routing entropy, mean spatial variance, and mean local routing variation.

Symbol	Definition (continued)
$\mathcal{H}(\alpha)$, $\mathrm{Var}_{\mathbf{s}}(\alpha)$, $\lVert\nabla\alpha\rVert$	Routing entropy, spatial variance, and mean local-gradient magnitude used in Table 15; these quantify the decisiveness and spatial structure of the learned allocation field.
δ_λ, δ_ϕ	First-order finite-difference operators in longitude and latitude.
q, $R_{m,q}$, G_q	Forecast-difficulty decile, normalized RMSE of model m in decile q, and PES RMSE gain in that decile.
ρ_s, $\rho_{\Xi,e}$	Spearman correlation and the Spearman correlation between Ξ and the absolute error of the stronger expert.
e_{stronger}	Point-wise forecast error of the task-level stronger constituent expert, used in the forecast-difficulty correlation analysis.
e_{P}, e_{N}	Point-wise forecast errors of Pangu-Weather and NeuralGCM, respectively.
σ_X, d^*	Reference standard deviation of the target field and the normalized verified forecast difficulty $d^* = \min(\lvert e_{\mathrm{P}}\rvert, \lvert e_{\mathrm{N}}\rvert)/\sigma_X$, respectively.
ρ_{align}	Advection-routing-expert-skill alignment, computed from the association between mean NeuralGCM allocation and signed relative expert skill across advection bins.
$\rho(\Xi, d^*)$	Spearman correlation between the routing-difficulty indicator Ξ and the post-hoc verified forecast difficulty d^*.
$\Delta\,\mathrm{RMSE}$, $\Delta\,\mathrm{ACC}$	Case-level improvement of PES over NeuralGCM in RMSE and ACC.
Baselines and oracle	
$\widehat{X}_{\mathrm{static}}$	Forecast produced by the training-fitted global static prior.
α_{static}	Spatially uniform NeuralGCM coefficient fitted on the training subset.
$\widehat{X}_{\mathrm{spatial}}$, $\alpha_{\mathrm{spatial}}(\mathbf{r})$	Forecast and location-specific NeuralGCM coefficient of the fixed spatial prior; the coefficient is fitted on the training subset and held fixed across forecast initializations.
α_{oracle}, $\widehat{X}_{\mathrm{oracle}}$	Verification-informed point-wise convex weight and the corresponding oracle forecast.
η	Fraction of the oracle-available aggregate RMSE improvement captured by PES.

A Experimental Protocol and Leakage Control

Data and grid. All experiments use globally aligned 2020 ERA5 analyses and forecast products from Pangu-Weather, NeuralGCM, and the operational IFS (Hersbach et al., 2020; Bi et al., 2023; Kochkov et al., 2024; Rasp et al., 2024). All fields are represented on, or interpolated to, a common 1.5° latitude-longitude grid with 121×240 spatial points. Forecasts and verification fields are matched by initialization time, valid time, target variable, pressure level, and forecast lead before any diagnostic construction or evaluation is performed.

Targets and routing tasks. The target set contains temperature, specific humidity, zonal wind, meridional wind, and geopotential at 500 and 850 hPa, yielding ten variable-level targets. Forecast lead times are 24, 48, 72, and 120 h. A separate routing network is trained for each variable-lead combination, resulting in 40 routing tasks.

Data splits. The final three calendar days of each month are excluded from gradient-based training and partitioned into non-overlapping validation and test subsets. The third-to-last calendar day of each month forms a seasonally distributed validation subset used exclusively for checkpoint selection. The final two calendar days form the test subset used for all final performance evaluations. The remaining aligned samples constitute the training subset. Consequently, no test sample contributes to gradient updates, normalization statistics, or checkpoint selection. Because all three subsets are drawn from a single calendar year, this protocol evaluates intra-annual generalization and should not be interpreted as a multi-year or climate-shift generalization test.

Leakage control. A central requirement of the experimental design is that the router must not access information from the verification time. For a forecast initialized at t_0 and verified at $t_0 + \tau$, the state channels X_0, U_0, and V_0 are taken from the ERA5 analysis available at initialization. The expert forecasts $\widehat{X}_{\mathrm{P}}$ and $\widehat{X}_{\mathrm{N}}$ are likewise available before the verification time. The future ERA5 field $X^*(t_0+\tau)$ is used only as the optimization target and evaluation reference and is never supplied to the routing network. All normalization statistics are estimated exclusively from the training subset and are fixed before validation and test evaluation.

Operational IFS reference. Operational IFS forecasts are evaluated at matching valid times and on the same spatial grid. Nevertheless, the IFS comparison is not a fully homogeneous reforecast experiment, because the operational forecasts and ERA5 may differ in model cycle, initialization procedure, and data-assimilation configuration. We therefore treat IFS as an external operational reference under a common ERA5-based verification protocol.

B Evaluation Metrics and Derived Statistics

B.1 Common verification protocol

All forecast systems are evaluated using the same valid-time alignment, spatial grid, latitude weights, and finite-value mask. Let $n = 1, \ldots, N$ index forecast initializations and $i = 1, \ldots, G$ index spatial grid points. The verification field is denoted by $X^*_{n,i}$, and the corresponding forecast from model m is denoted by $\widehat{X}^{(m)}_{n,i}$.

For every comparison, we construct a common validity mask $M_{n,i}$ requiring the ERA5 verification and all forecast systems included in that comparison to be finite. Using a common mask prevents apparent skill differences from arising because different models are evaluated over different spatial locations.

For grid point i at latitude ϕ_i, the area weight is

$$w_i = \cos \phi_i. \tag{11}$$

In implementation, the weights are divided by their spatial mean for numerical convenience. This common normalization does not alter the metrics.

B.2 LATITUDE-WEIGHTED RMSE

For one forecast initialization n, the latitude-weighted RMSE of model m is

$$\mathrm{RMSE}_n^{(m)} = \sqrt{\frac{\sum_{i=1}^{G} M_{n,i} w_i \left(\widehat{X}_{n,i}^{(m)} - X_{n,i}^{*}\right)^2}{\sum_{i=1}^{G} M_{n,i} w_i}}. \tag{12}$$

The aggregate RMSE reported in the complete result tables is

$$\mathrm{RMSE}^{(m)} = \sqrt{\frac{\sum_{n=1}^{N}\sum_{i=1}^{G} M_{n,i} w_i \left(\widehat{X}_{n,i}^{(m)} - X_{n,i}^{*}\right)^2}{\sum_{n=1}^{N}\sum_{i=1}^{G} M_{n,i} w_i}}. \tag{13}$$

Case-wise RMSE is used for paired initialization-level analyses, whereas aggregate RMSE is used for the main task-level comparison.

B.3 ANOMALY CORRELATION COEFFICIENT

ACC is computed using the pre-computed climatological fields distributed with WeatherBench2 (Rasp et al., 2024). For each forecast case, the climatology is selected by the day of year and UTC hour of the forecast valid time and is aligned to the same spatial grid as the forecasts and ERA5 verification.

Let $C_{n,i}$ denote the corresponding climatological value. Forecast and observed anomalies are

$$f_{n,i}^{(m)\prime} = \widehat{X}_{n,i}^{(m)} - C_{n,i}, \qquad o'_{n,i} = X_{n,i}^{*} - C_{n,i}. \tag{14}$$

For each initialization, the weighted anomaly means are

$$\overline{f}_n^{(m)\prime} = \frac{\sum_i M_{n,i}^{\mathrm{ACC}} w_i f_{n,i}^{(m)\prime}}{\sum_i M_{n,i}^{\mathrm{ACC}} w_i}, \qquad \overline{o}'_n = \frac{\sum_i M_{n,i}^{\mathrm{ACC}} w_i o'_{n,i}}{\sum_i M_{n,i}^{\mathrm{ACC}} w_i}, \tag{15}$$

where $M_{n,i}^{\mathrm{ACC}}$ additionally requires the climatology to be finite. The case-wise ACC is

$$\mathrm{ACC}_n^{(m)} = \frac{\sum_i M_{n,i}^{\mathrm{ACC}} w_i \left(f_{n,i}^{(m)\prime} - \overline{f}_n^{(m)\prime}\right)\left(o'_{n,i} - \overline{o}'_n\right)}{\sqrt{\left[\sum_i M_{n,i}^{\mathrm{ACC}} w_i \left(f_{n,i}^{(m)\prime} - \overline{f}_n^{(m)\prime}\right)^2\right]\left[\sum_i M_{n,i}^{\mathrm{ACC}} w_i \left(o'_{n,i} - \overline{o}'_n\right)^2\right]}}. \tag{16}$$

The reported ACC is the arithmetic mean over valid initializations,

$$\mathrm{ACC}^{(m)} = \frac{1}{N_{\mathrm{valid}}} \sum_{n \in \mathcal{I}_{\mathrm{valid}}} \mathrm{ACC}_n^{(m)}. \tag{17}$$

B.4 RELATIVE IMPROVEMENTS AND PAIRED WIN RATES

For each variable-lead task, the better constituent expert is defined separately for each metric:

$$\mathrm{RMSE}_{\mathrm{best}} = \min\left(\mathrm{RMSE}_{\mathrm{P}}, \mathrm{RMSE}_{\mathrm{N}}\right), \qquad \mathrm{ACC}_{\mathrm{best}} = \max\left(\mathrm{ACC}_{\mathrm{P}}, \mathrm{ACC}_{\mathrm{N}}\right). \tag{18}$$

The RMSE reduction relative to reference model b is

$$I_{\mathrm{RMSE}}^{(b)} = 100\frac{\mathrm{RMSE}^{(b)} - \mathrm{RMSE}^{(\mathrm{PES})}}{\mathrm{RMSE}^{(b)}}\%. \tag{19}$$

Because ACC is bounded above by one, we report the fraction of the remaining gap to perfect correlation closed by PES:

$$G_{\mathrm{ACC}}^{(b)} = 100\frac{\mathrm{ACC}^{(\mathrm{PES})} - \mathrm{ACC}^{(b)}}{1 - \mathrm{ACC}^{(b)}}\%. \tag{20}$$

For a fixed task and reference b, the paired initialization-level win rates are

$$W_{\mathrm{RMSE}}^{(b)} = \frac{100}{N} \sum_{n=1}^{N} \mathbb{I}\left[\mathrm{RMSE}_n^{(\mathrm{PES})} < \mathrm{RMSE}_n^{(b)}\right] \%, \tag{21}$$

and

$$W_{\mathrm{ACC}}^{(b)} = \frac{100}{N_{\mathrm{valid}}} \sum_{n \in \mathcal{I}_{\mathrm{valid}}} \mathbb{I}\left[\mathrm{ACC}_n^{(\mathrm{PES})} > \mathrm{ACC}_n^{(b)}\right] \%. \tag{22}$$

For the grouped robustness summary in the main text, cases are grouped by calendar month, target variable, pressure level, and lead time. Metrics are aggregated within each group before paired comparison. PES wins 99.6% of the grouped comparisons against NeuralGCM and 100% of the corresponding comparisons against Pangu-Weather and operational IFS. These grouped percentages are distinct from the per-initialization win rates.

Paired-bootstrap uncertainty. We quantify test-sample uncertainty with 20,000 paired bootstrap resamples of held-out forecast initializations. PES and both constituent experts use the same resampled cases, and the task-wise stronger expert is recomputed within each replicate. The 95% confidence intervals for both RMSE improvement and absolute ACC gain remain above zero for all 40 tasks. The smallest RMSE-gain lower bound is 1.069% (Q850 at 24 h), and the smallest absolute ACC-gain lower bound is 5.9×10^{-5} (Z500 at 24 h). These intervals quantify test-sample uncertainty for the trained routers and do not include training-seed variation.

B.5 Routing-field statistics

Let $\alpha_{n,i} \in [0, 1]$ denote the NeuralGCM routing weight. The mean binary routing entropy is

$$H_\alpha = -\frac{1}{NG} \sum_{n=1}^{N} \sum_{i=1}^{G} \left[\alpha_{n,i} \log \alpha_{n,i} + (1 - \alpha_{n,i}) \log(1 - \alpha_{n,i})\right]. \tag{23}$$

Weights are clipped to $[\epsilon, 1 - \epsilon]$ before evaluating the logarithm.

For case n, define the spatial mean

$$\overline{\alpha}_n = \frac{1}{G} \sum_{i=1}^{G} \alpha_{n,i}. \tag{24}$$

The mean spatial variance is

$$V_\alpha = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{G} \sum_{i=1}^{G} \left(\alpha_{n,i} - \overline{\alpha}_n\right)^2. \tag{25}$$

Local routing variation is measured from first-order finite differences between neighboring grid points,

$$G_\alpha = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{G} \sum_{i=1}^{G} \sqrt{\left(\delta_\lambda \alpha_{n,i}\right)^2 + \left(\delta_\phi \alpha_{n,i}\right)^2}. \tag{26}$$

The numerical implementation evaluates spatial derivatives with centered finite differences at interior grid points and one-sided differences at the domain boundaries before taking the Euclidean gradient magnitude. The three quantities characterize routing decisiveness, global spatial differentiation, and local spatial variation, respectively; none is treated as a standalone forecast-skill metric.

C Physical Motivation, Diagnostic Construction, and Router Implementation

C.1 Physical basis of transport-guided routing

PES is designed to learn expert preference from forecast-stage information that has a direct connection to the local atmospheric evolution. Let $X(\lambda, \phi, t)$ denote a prognostic atmospheric variable on

the sphere, where λ and ϕ are longitude and latitude. Near the forecast initialization time t_0, its local evolution may be written schematically as

$$\frac{\partial X}{\partial t} + \frac{U}{R_{\mathrm{E}} \cos\phi} \frac{\partial X}{\partial \lambda} + \frac{V}{R_{\mathrm{E}}} \frac{\partial X}{\partial \phi} = \mathcal{F}, \tag{27}$$

where U and V are the horizontal wind components, R_{E} is the mean Earth radius, and $\mathcal{F}$ collects all tendencies not represented by resolved horizontal advection. Depending on the target variable, $\mathcal{F}$ may include vertical transport, diabatic heating, moisture sources and sinks, pressure-gradient and Coriolis effects, diffusion, subgrid-scale processes, and other unresolved physical terms.

At initialization, the signed horizontal-advection contribution is

$$\mathcal{A} = -\left[\frac{U_0}{R_{\mathrm{E}} \cos\phi} \frac{\partial X_0}{\partial \lambda} + \frac{V_0}{R_{\mathrm{E}}} \frac{\partial X_0}{\partial \phi} \right] = -\mathbf{u}_0 \cdot \nabla_s X_0, \tag{28}$$

where $\mathbf{u}_0 = (U_0, V_0)$, $X_0 = X(t_0)$, and ∇_s denotes the horizontal gradient on the spherical surface. Equation (27) then gives

$$\left. \frac{\partial X}{\partial t} \right|_{t=t_0} = \mathcal{A} + \mathcal{F}_0. \tag{29}$$

A first-order expansion over a short interval τ yields

$$X(t_0 + \tau) = X_0 + \tau \left(\mathcal{A} + \mathcal{F}_0 \right) + \mathcal{O}(\tau^2). \tag{30}$$

Therefore, $\mathcal{A}$ is not an arbitrary handcrafted feature. It is the initialization-time observable contribution of resolved horizontal transport to the immediate evolution of the target field. For temperature, humidity, and geopotential, it represents the horizontal transport of the corresponding scalar-like field. For zonal and meridional wind, it represents component-wise horizontal advection and should not be interpreted as the complete momentum tendency. PES consequently uses $\mathcal{A}$ as a transport diagnostic rather than as a complete governing equation or a hard physical constraint.

The sign of $\mathcal{A}$ indicates the direction of the local transport-induced tendency, whereas $|\mathcal{A}|$ measures the strength of the interaction between the flow and the spatial gradient. Large $|\mathcal{A}|$ can arise from strong winds, sharp gradients, or both, and therefore identifies regions undergoing rapid transport-related evolution.

C.2 Flow-dependent forecast sensitivity

The relevance of advection to expert routing can be understood through local error propagation. Consider expert m, and let δX_m and $\delta \mathbf{u}_m$ denote errors in its represented target field and horizontal flow. Linearizing the horizontal-advection operator around the reference state gives

$$\delta \mathcal{A}_m \approx -\delta \mathbf{u}_m \cdot \nabla_s X_0 - \mathbf{u}_0 \cdot \nabla_s \delta X_m, \tag{31}$$

where second-order products of perturbations are omitted. Consequently,

$$|\delta \mathcal{A}_m| \lesssim \|\delta \mathbf{u}_m\| \, \|\nabla_s X_0\| + \|\mathbf{u}_0\| \, \|\nabla_s \delta X_m\| \, . \tag{32}$$

The first term describes transport error induced by an imperfect wind field, whereas the second describes the advection of errors in the represented target structure. Equation (32) shows that sharp gradients amplify wind errors and strong flow amplifies errors in the spatial structure of the transported field. Thus, dynamically active regions are more sensitive to differences in how forecasting systems represent winds, gradients, phase propagation, and coherent structures.

Pangu-Weather and NeuralGCM have different architectures and physical inductive biases. Their effective errors in flow and transported structure therefore need not grow in the same way. A given atmospheric regime may amplify one expert's transport-related error more strongly than the other's, causing their relative skill to become state dependent.

For additional intuition, under the local proportional-tendency-error model

$$\delta \mathcal{A}_m = \eta_m \mathcal{A}, \qquad \mathbb{E}[\eta_m] = 0, \qquad \mathrm{Var}(\eta_m) = \sigma_{\eta,m}^2, \tag{33}$$

the transport-related error accumulated over an interval τ satisfies

$$\mathrm{Var}\left(\tau\delta\mathcal{A}_m \mid \mathcal{A}\right) = \tau^2\sigma_{\eta,m}^2\mathcal{A}^2. \tag{34}$$

This approximation is not assumed by the router and does not imply that strong advection necessarily produces a large forecast error. It provides a physical motivation for treating $|\mathcal{A}|$ as an observable indicator of flow-dependent forecast sensitivity and of regimes in which expert skill may separate more strongly.

C.3 Expert disagreement and local prediction difficulty

Advection describes the physical regime, but it does not reveal whether the two experts actually produce different outcomes. We therefore introduce the signed expert-disagreement field

$$\Delta = \widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}, \tag{35}$$

where $\widehat{X}_{\mathrm{P}}$ and $\widehat{X}_{\mathrm{N}}$ are the Pangu-Weather and NeuralGCM forecasts at the target lead time.

The sign of Δ identifies which expert predicts the larger local value, while $|\Delta|$ measures the separation of the two forecast products. If their point-wise mean is

$$\overline{X} = \frac{\widehat{X}_{\mathrm{P}} + \widehat{X}_{\mathrm{N}}}{2}, \tag{36}$$

then the variance of this two-member forecast set is

$$\mathrm{Var}_m\left(\widehat{X}^{(m)}\right) = \frac{1}{2}\left[\left(\widehat{X}_{\mathrm{P}} - \overline{X}\right)^2 + \left(\widehat{X}_{\mathrm{N}} - \overline{X}\right)^2\right] = \frac{\Delta^2}{4}. \tag{37}$$

Thus, $|\Delta|$ is directly proportional to the spread of the two-expert forecast set.

Disagreement does not reveal which expert is correct, and two experts may share correlated errors. Nevertheless, for any verification value X^*, the triangle inequality gives

$$\max\left(\left|\widehat{X}_{\mathrm{P}} - X^*\right|, \left|\widehat{X}_{\mathrm{N}} - X^*\right|\right) \geq \frac{|\Delta|}{2}. \tag{38}$$

Large disagreement therefore guarantees that at least one expert has a non-negligible local error. This makes Δ a directly observable model-space indicator of forecast difficulty and of where the routing decision can materially affect the composed forecast.

The two diagnostics describe complementary aspects of the problem. $\mathcal{A}$ identifies atmospheric states whose evolution is sensitive to transport-related representation errors, whereas Δ identifies where the two experts have already produced substantially different predictions. Neither quantity alone is sufficient: large $|\mathcal{A}|$ need not imply low forecast accuracy, and large $|\Delta|$ does not determine which expert is better.

C.4 Forecast-difficulty score and seven-channel representation

Because $\mathcal{A}$ and Δ have different physical units and variable-dependent scales, they are standardized separately for every variable-lead task using statistics estimated exclusively from the training subset. More generally, for

$$Y \in \{X_0,\, U_0,\, V_0,\, \mathcal{A},\, \Delta\}, \tag{39}$$

we define

$$\widetilde{Y} = \frac{Y - \mu_Y^{\mathrm{train}}}{\sigma_Y^{\mathrm{train}} + \epsilon}, \tag{40}$$

where ϵ is a small numerical constant. The same frozen statistics are applied without refitting to both the validation and test subsets.

We combine the dimensionless advection and disagreement fields into the local forecast-difficulty score

$$\Xi = \sqrt{\widetilde{\mathcal{A}}^2 + \widetilde{\Delta}^2 + \epsilon}. \tag{41}$$

Geometrically, Ξ is the distance in the standardized transport-disagreement plane from the comparatively easy regime characterized by weak transport activity and close expert agreement. A large Ξ arises when transport-related dynamical activity is strong, expert disagreement is substantial, or both.

We interpret Ξ as an explicit prediction-difficulty indicator. Empirically, larger Ξ is associated with lower accuracy of the uncomposed expert forecasts, indicating that the original expert predictions become less reliable as the joint transport-disagreement magnitude increases. Therefore, high-Ξ regions are precisely those in which adaptive expert selection has greater potential value.

This interpretation is diagnostic rather than probabilistic. Ξ is not a calibrated predictive variance, does not provide a probability of forecast failure, and does not guarantee a strictly monotonic pointwise relationship with error. It summarizes conditions associated with reduced baseline skill. Moreover, Ξ measures how difficult the local forecast is, but does not specify which expert should receive more weight. For this reason, the router receives $\widetilde{\mathcal{A}}$, $\widetilde{\Delta}$, and Ξ as separate channels. The signed fields preserve the direction of transport and expert difference, while Ξ exposes their nonlinear joint magnitude.

The UTC initialization hour is encoded as

$$c_h = \cos\left(\frac{2\pi h_0}{24}\right), \tag{42}$$

where h_0 denotes the initialization hour. For the 00 and 12 UTC cycles used here, c_h equals $+1$ and -1, respectively.

The complete routing input is

$$\mathbf{z} = \left[\widetilde{X}_0,\, c_h,\, \widetilde{U}_0,\, \widetilde{V}_0,\, \widetilde{\mathcal{A}},\, \widetilde{\Delta},\, \Xi\right] \in \mathbb{R}^{7\times H\times W}. \tag{43}$$

The representation is designed to answer three successive questions: what atmospheric state is present, how difficult its local evolution is to predict, and how differently the two experts respond to it. A separate router is trained for every variable-lead task, so forecast lead time is represented by task identity rather than by an additional input channel.

C.5 Numerical construction of the diagnostic fields

The spherical advection in Equation (28) is evaluated from the raw initialization-time fields before normalization. Longitude and latitude derivatives are taken with respect to radians. Longitude derivatives use centered finite differences with periodic wrapping,

$$\left.\frac{\partial X}{\partial \lambda}\right|_{i,j} \approx \frac{X_{i,j+1} - X_{i,j-1}}{2\Delta\lambda}, \tag{44}$$

where the longitude index is interpreted periodically. Latitude derivatives use centered differences at interior points,

$$\left.\frac{\partial X}{\partial \phi}\right|_{i,j} \approx \frac{X_{i+1,j} - X_{i-1,j}}{2\Delta\phi}, \tag{45}$$

and one-sided differences at the northern and southern boundaries. At exact polar grid points, where $\cos\phi$ vanishes, the zonal contribution is set to zero for numerical stability. The meridional contribution is retained.

The expert difference is constructed from forecasts already available at the target lead time, after all fields have been aligned to the common grid and valid time. The difficulty score Ξ is computed only after $\mathcal{A}$ and Δ have been standardized using the variable-lead-specific training statistics. No verification-time ERA5 field enters the construction of any routing channel.

C.6 Lightweight spatial routing network

Both expert models remain frozen throughout training. Given $\mathbf{z}$, the router first applies a local convolutional encoder,

$$\mathbf{F}_{\mathrm{loc}} = f_{\mathrm{CNN}}(\mathbf{z}) \in \mathbb{R}^{64\times H\times W}. \tag{46}$$

The encoder contains two consecutive 3×3 convolutions. The first maps the seven input channels to 64 hidden channels, and the second preserves the 64-channel width. Each convolution is followed by Group Normalization with eight groups and a GELU activation. This stage extracts local gradients, front-like structures, transport signatures, and spatial variations in expert disagreement.

The feature map is flattened into HW spatial tokens with dimension 64 and processed by a two-layer Transformer encoder,

$$\mathbf{T}_{\mathrm{ctx}} = f_{\mathrm{Tr}}\left(\mathrm{Flatten}\left(\mathbf{F}_{\mathrm{loc}}\right)\right). \tag{47}$$

Each Transformer layer uses four self-attention heads, hidden dimension 64, feed-forward dimension 128, pre-normalization, ReLU activation in the feed-forward block, and dropout 0.1. No additional positional embedding is used; the convolutional encoder provides local spatial context before tokenization.

The contextual features are reshaped to the spatial grid and combined with the convolutional features through an external residual connection,

$$\mathbf{F}_{\mathrm{fused}} = \mathbf{F}_{\mathrm{loc}} + \mathbf{F}_{\mathrm{ctx}}. \tag{48}$$

A routing head composed of 1×1 convolutions with channel dimensions $64 \to 32 \to 1$, with a GELU activation between the two layers, produces the NeuralGCM routing logit. The final weight is

$$\alpha_{\mathrm{N}} = \sigma\left(g_\theta(\mathbf{z})\right) \in [0, 1]. \tag{49}$$

The composed forecast is

$$\widehat{X}_{\mathrm{PES}} = \alpha_{\mathrm{N}} \widehat{X}_{\mathrm{N}} + \left(1 - \alpha_{\mathrm{N}}\right) \widehat{X}_{\mathrm{P}}. \tag{50}$$

Thus, PES predicts neither a new atmospheric field nor an additive bias correction. Its only learned output is a spatial allocation field governing the local convex selection of two pre-computed forecasts. Each router contains 110,337 trainable parameters.

C.7 Optimization and global-field inference

The routing parameters are optimized with the latitude-weighted mean squared error

$$\mathcal{L}(\theta) = \frac{\sum_i m_i w_i \left[\widehat{X}_{\mathrm{PES},i} - X_i^*\right]^2}{\sum_i m_i w_i}, \qquad w_i = \frac{\cos\phi_i}{\langle \cos\phi \rangle}, \tag{51}$$

where m_i is a common validity mask requiring the verification field and both expert forecasts to be finite. The normalization of w_i does not change the objective but keeps its numerical scale stable.

Training is performed on randomly sampled 64×64 spatial patches. For every valid initialization, 64 patches are sampled during each epoch, exposing the router to diverse spatial regions and atmospheric regimes while limiting memory consumption. The batch size is 128. Each variable-lead-specific router is trained for 12 epochs with AdamW, a maximum learning rate of 3×10^{-4}, and a OneCycle learning-rate schedule. Automatic mixed precision with bfloat16 is used on compatible hardware, and the gradient norm is clipped to 1.0. After every epoch, latitude-weighted RMSE is evaluated exclusively on the validation subset, and the checkpoint with the lowest validation RMSE is retained. The selected checkpoint is then evaluated on the test subset without further parameter updates or model selection.

During full-field validation and test inference, the 121×240 field is processed using shifted 64×64 spatial tiles. Tiles near terminal latitude and longitude boundaries are shifted to remain inside the global domain, leading to limited overlap. For a location $\mathbf{r}$ covered by the tile set $\mathcal{K}(\mathbf{r})$, overlapping predictions are averaged:

$$\alpha_{\mathrm{N}}(\mathbf{r}) = \frac{\sum_{k \in \mathcal{K}(\mathbf{r})} \alpha_{\mathrm{N}}^{(k)}(\mathbf{r})}{|\mathcal{K}(\mathbf{r})|}. \tag{52}$$

The resulting global routing field is then applied to the two complete expert forecast fields.

Forty routers are trained in total, corresponding to ten variable-level targets and four forecast lead times. All routing experiments are conducted on a single NVIDIA GeForce RTX 4090D. The reported PES cost includes diagnostic construction, routing inference, and convex selection, but does not include the cost of generating the underlying Pangu-Weather and NeuralGCM forecasts.

D COMPLETE FORECAST RESULTS

D.1 COMPLETE RMSE AND ACC RESULTS

Tables 4-5 and 6-7 report the complete numerical results for all ten variable-level targets and four forecast lead times. Pangu-Weather, NeuralGCM, operational IFS, and PES are evaluated using the same valid-time alignment, latitude weighting, climatological reference, and common finite-value masks defined in Appendix B. RMSE values retain the native units of each target variable. Lower RMSE and higher ACC indicate better forecast skill.

Across all 40 variable-lead tasks, PES achieves both lower RMSE and higher ACC than Pangu-Weather, NeuralGCM, and operational IFS. Relative to the better of the two constituent experts for each task, the RMSE reduction ranges from 1.17% to 6.27%, with a mean reduction of 3.52%. The absolute ACC gain ranges from 6.72×10^{-5} to 2.69×10^{-2}, with a mean gain of 5.20×10^{-3}. Expressed as the fraction of the remaining ACC gap to unity closed by PES, the improvement ranges from approximately 2.4% to 11.0%. These gains span thermodynamic, moisture, wind, and geopotential variables at both 500 and 850 hPa, and remain positive from 24 to 120 h.

Relative to operational IFS, the RMSE reduction ranges from 16.26% to 26.19%, with a mean reduction of 19.51%. As discussed in Appendix A, IFS is used as an external operational reference under the same ERA5-based verification protocol rather than as a fully homogeneous reforecast baseline.

Table 4: **Complete latitude-weighted RMSE results at 24-h and 48-h lead times.** Lower is better; the best value in each row is bold.

(a) 24-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.4314	0.3748	0.4752	**0.3679**
T850	0.6209	0.5483	0.7261	**0.5359**
Q500	2.8993×10^{-4}	2.6717×10^{-4}	3.4140×10^{-4}	2.6251×10^{-4}
Q850	6.6995×10^{-4}	5.8211×10^{-4}	7.7936×10^{-4}	5.7530×10^{-4}
U500	1.4833	1.3074	1.5801	**1.2897**
U850	1.1626	1.0479	1.3016	**1.0260**
V500	1.5316	1.3452	1.6174	**1.3262**
V850	1.1949	1.0751	1.3218	**1.0527**
Z500	44.4506	38.1147	48.0320	**36.3141**
Z850	39.7563	35.3474	43.5715	**33.2825**

(b) 48-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.6213	0.5715	0.6855	**0.5530**
T850	0.8217	0.7534	0.9327	**0.7317**
Q500	3.8547×10^{-4}	3.6111×10^{-4}	4.4417×10^{-4}	3.5279×10^{-4}
Q850	8.7011×10^{-4}	8.0737×10^{-4}	1.0024×10^{-3}	7.8640×10^{-4}
U500	2.0897	1.9514	2.2653	**1.8953**
U850	1.5922	1.4933	1.7949	**1.4495**
V500	2.1591	2.0243	2.3380	**1.9578**
V850	1.6387	1.5185	1.8158	**1.4781**
Z500	81.3946	69.8695	85.6980	**67.9168**
Z850	66.6673	57.4525	70.6371	**55.6627**

Table 5: **Complete latitude-weighted RMSE results at 72-h and 120-h lead times.** Lower is better; the best value in each row is bold.

(a) 72-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.8715	0.8118	0.9526	**0.7773**
T850	1.0540	0.9738	1.1821	**0.9390**
Q500	4.6383×10^{-4}	4.3291×10^{-4}	5.2916×10^{-4}	4.2062×10^{-4}
Q850	1.0319×10^{-3}	9.8209×10^{-4}	1.1896×10^{-3}	9.4385×10^{-4}
U500	2.8204	2.6715	3.1006	**2.5626**
U850	2.1198	1.9884	2.3543	**1.9172**
V500	2.9199	2.7861	3.1936	**2.6620**
V850	2.1730	2.0168	2.3758	**1.9478**
Z500	135.0482	118.9493	140.0454	**115.3461**
Z850	105.6136	91.3547	108.7208	**89.0263**

(b) 120-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	1.5681	1.4577	1.6480	**1.3768**
T850	1.7376	1.5911	1.8592	**1.5204**
Q500	6.3169×10^{-4}	5.7398×10^{-4}	6.8157×10^{-4}	5.5382×10^{-4}
Q850	1.3895×10^{-3}	1.3447×10^{-3}	1.5454×10^{-3}	1.2628×10^{-3}
U500	4.5696	4.3855	4.9907	**4.1105**
U850	3.4328	3.2572	3.7106	**3.0768**
V500	4.8118	4.5622	5.1730	**4.2864**
V850	3.4762	3.2908	3.7394	**3.1028**
Z500	295.7312	268.3292	309.4698	**256.2954**
Z850	220.4196	201.2571	231.4862	**192.4705**

Table 6: **Complete ACC results at 24-h and 48-h lead times.** Higher is better; the best value in each row is bold.

(a) 24-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.9906	0.9929	0.9887	**0.9931**
T850	0.9834	0.9871	0.9774	**0.9877**
Q500	0.9373	0.9467	0.9133	**0.9486**
Q850	0.9283	0.9460	0.9070	**0.9473**
U500	0.9842	0.9878	0.9822	**0.9881**
U850	0.9790	0.9830	0.9737	**0.9837**
V500	0.9834	0.9872	0.9815	**0.9876**
V850	0.9772	0.9816	0.9721	**0.9824**
Z500	0.9986	0.9990	0.9984	**0.9991**
Z850	0.9977	0.9982	0.9971	**0.9983**

(b) 48-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.9806	0.9835	0.9766	**0.9846**
T850	0.9715	0.9758	0.9630	**0.9772**
Q500	0.8892	0.9008	0.8513	**0.9056**
Q850	0.8776	0.8935	0.8443	**0.8989**
U500	0.9689	0.9728	0.9635	**0.9744**
U850	0.9604	0.9652	0.9499	**0.9672**
V500	0.9674	0.9714	0.9618	**0.9732**
V850	0.9573	0.9633	0.9477	**0.9653**
Z500	0.9953	0.9965	0.9948	**0.9967**
Z850	0.9935	0.9951	0.9926	**0.9954**

Table 7: **Complete ACC results at 72-h and 120-h lead times.** Higher is better; the best value in each row is bold.

(a) 72-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.9623	0.9669	0.9552	**0.9696**
T850	0.9536	0.9597	0.9411	**0.9626**
Q500	0.8424	0.8580	0.7913	**0.8662**
Q850	0.8276	0.8414	0.7800	**0.8526**
U500	0.9432	0.9490	0.9318	**0.9531**
U850	0.9291	0.9375	0.9132	**0.9419**
V500	0.9400	0.9453	0.9286	**0.9500**
V850	0.9226	0.9333	0.9084	**0.9377**
Z500	0.9871	0.9898	0.9859	**0.9904**
Z850	0.9833	0.9874	0.9821	**0.9880**

(b) 120-h lead time

Target	Pangu-Weather	NeuralGCM	IFS	PES
T500	0.8745	0.8890	0.8621	**0.8996**
T850	0.8674	0.8851	0.8477	**0.8946**
Q500	0.7112	0.7486	0.6569	**0.7643**
Q850	0.6886	0.7043	0.6283	**0.7312**
U500	0.8459	0.8576	0.8181	**0.8732**
U850	0.8090	0.8275	0.7799	**0.8435**
V500	0.8293	0.8465	0.8065	**0.8619**
V850	0.7937	0.8143	0.7672	**0.8318**
Z500	0.9341	0.9443	0.9265	**0.9492**
Z850	0.9205	0.9331	0.9115	**0.9385**

D.2 CASE-LEVEL ROBUSTNESS FOR Z500

The aggregate results above establish task-level improvement, but do not show whether the gain is broadly distributed across forecast initializations. We therefore additionally compare PES and NeuralGCM for each test-set Z500 case. NeuralGCM is used as the reference because it is the stronger constituent expert for Z500 at all four evaluated lead times.

Table 8 shows that PES improves both RMSE and ACC for every 24-h case. The paired win rates remain 95.45% for both metrics at 48 h, and remain above 86% at 72 and 120 h. The mean RMSE advantage also increases with forecast range, from 1.7614 at 24 h to 12.3449 at 120 h. Thus, the aggregate improvement is not produced by only a small number of extreme cases; it persists across most individual initializations, including at the longest evaluated lead time.

Table 8: **Case-level robustness of PES against NeuralGCM for Z500.** Win rates are computed over individual test-set forecast initializations. Positive mean differences indicate improvement by PES.

Lead	Cases	RMSE win (%)	ACC win (%)	Mean ΔRMSE	Mean ΔACC
24 h	46	100.00	100.00	1.7614	0.0001
48 h	44	95.45	95.45	1.9711	0.0002
72 h	44	88.64	86.36	3.6944	0.0006
120 h	44	88.64	86.36	12.3449	0.0049

Here,

$$\Delta\,\mathrm{RMSE} = \mathrm{RMSE}_{\mathrm{NGCM}} - \mathrm{RMSE}_{\mathrm{PES}}, \tag{53}$$

and

$$\Delta\,\mathrm{ACC} = \mathrm{ACC}_{\mathrm{PES}} - \mathrm{ACC}_{\mathrm{NGCM}}, \tag{54}$$

so positive values consistently indicate improvement by PES.

E Additional Expert Pairs for Cross-Model Generalization

To test whether the reported benefit depends on the particular choice of experts, we repeat the complete 40-task protocol with GraphCast substituted for each of the two experts used in the main comparison, yielding two additional heterogeneous pairs: GraphCast with NeuralGCM (GN) and Pangu-Weather with GraphCast (PG). For each pair the router is retrained from scratch under the identical data split, routing architecture, optimizer, training budget, and latitude-weighted evaluation protocol; only the two expert forecast fields change, while all normalization statistics and static coefficients are fitted exclusively on the router-training subset. Task-level uncertainty is estimated by $20{,}000$ paired bootstrap resamples of held-out forecast initializations, using identical resampled cases for Full PES and each baseline. GraphCast is the stronger constituent expert in both pairs. Full PES improves upon it across all 40 tasks for both GN and PG, and wins $448/450$ and $450/450$ paired task–month comparisons, respectively. Figures 5 and 6 report the complete results.

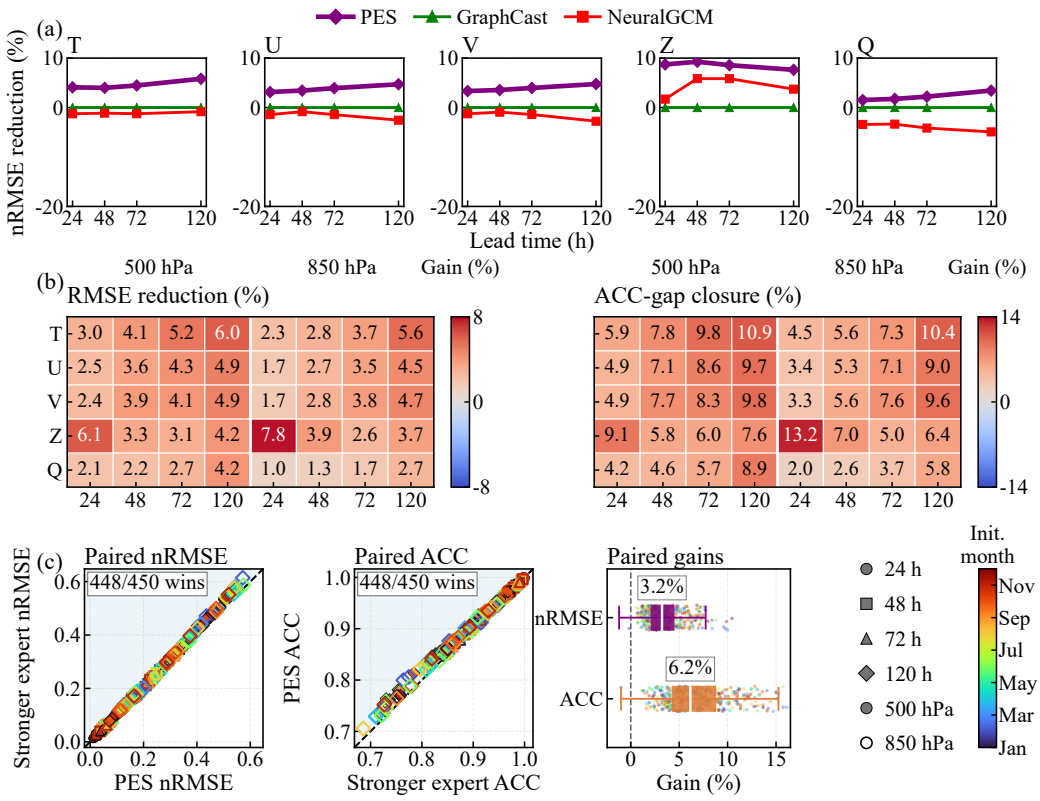

Figure 5: **Cross-model generalization for GraphCast–NeuralGCM (GN).** (a) Mean nRMSE reduction relative to GraphCast across atmospheric variables (T, U, V, Z, and Q) and four forecast lead times, averaged over pressure levels and initialization months; GraphCast is fixed at zero and positive values indicate lower nRMSE. PES, GraphCast, and NeuralGCM are shown. (b) RMSE reduction and ACC-gap closure relative to the task-wise stronger expert; each cell is one variable–level–lead-time task. (c) Paired comparisons against the stronger expert over all task–month aggregates: nRMSE, ACC, and the distribution of paired gains. Color, marker, and fill denote initialization month, lead time, and pressure level, respectively; points in the shaded region indicate that PES outperforms the stronger expert.

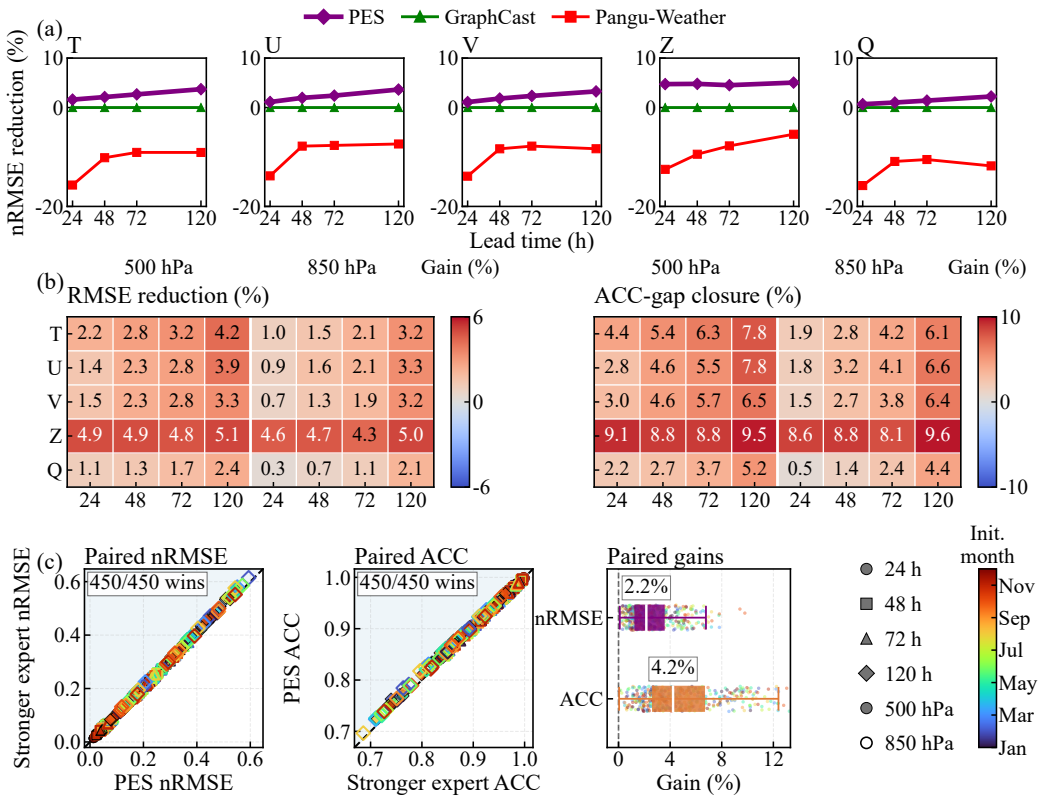

Figure 6: **Cross-model generalization for Pangu-Weather–GraphCast (PG).** Panels follow Figure 5, with Pangu-Weather replacing NeuralGCM as the non-GraphCast constituent expert.

F Details of Expert-Selection Baselines and Ablation Variants

This section specifies the comparison variants used in the controlled Z500 experiment at 120 h. Their interpretation is discussed in the main text; here we provide the exact construction and numerical results needed for reproducibility. The controlled diagnostic ablations use the same data split, routing architecture, optimizer, patch construction, checkpoint-selection rule, and latitude-weighted evaluation protocol as the matched Full reference. All learned variants in this controlled comparison use the same six-epoch training budget. Checkpoints are selected on the validation subset, and the reported comparison metrics are computed on the test subset.

F.1 Expert-selection and routing baselines

The parameter-free 50/50 baseline applies the uniform average

$$\widehat{X}_{50/50} = \tfrac{1}{2}\widehat{X}_{\mathrm{N}} + \tfrac{1}{2}\widehat{X}_{\mathrm{P}}. \tag{55}$$

It uses neither fitted parameters nor spatially varying weights.

The global static prior uses

$$\widehat{X}_{\mathrm{static}} = \alpha_{\mathrm{static}}\widehat{X}_{\mathrm{N}} + (1-\alpha_{\mathrm{static}})\,\widehat{X}_{\mathrm{P}}, \tag{56}$$

where one scalar coefficient is fitted on the training subset by minimizing the same latitude-weighted mean squared error used for PES. For Z500 at 120 h, the fitted NeuralGCM coefficient is

$$\alpha_{\mathrm{static}} = 0.668. \tag{57}$$

The coefficient is then fixed and applied uniformly to every spatial location and forecast initialization in the test subset.

The fixed spatial prior generalizes this construction by fitting one location-specific coefficient on the training subset,

$$\widehat{X}_{\mathrm{spatial}}(\mathbf{r}) = \alpha_{\mathrm{spatial}}(\mathbf{r})\widehat{X}_{\mathrm{N}}(\mathbf{r}) + [1-\alpha_{\mathrm{spatial}}(\mathbf{r})]\,\widehat{X}_{\mathrm{P}}(\mathbf{r}), \tag{58}$$

with $\alpha_{\mathrm{spatial}}(\mathbf{r}) \in [0,1]$. Once fitted, this spatial map is held fixed for every validation and test initialization. It therefore captures persistent geographical expert preference while excluding weather-state-dependent adaptation.

The raw-forecast selector receives only the two standardized expert forecast fields,

$$\mathbf{z}_{\mathrm{raw}} = \left[\widetilde{\widehat{X}}_{\mathrm{P}}, \widetilde{\widehat{X}}_{\mathrm{N}}\right], \tag{59}$$

and predicts a spatial preference field that is applied through the same convex expert composition. It uses the same convolution-Transformer routing backbone with a two-channel input and contains 107,457 trainable parameters. This baseline tests whether direct access to the two forecast fields is sufficient without the explicit atmospheric and diagnostic representation used by PES.

The Blind router retains the base state channels $[\widetilde{X}_0, c_h, \widetilde{U}_0, \widetilde{V}_0]$ but removes $\mathcal{A}$, Δ, and Ξ. The routing network and optimization protocol are otherwise matched to Full PES. Removed inputs are replaced by zero-valued maps, so the controlled learned ablations expose seven model channels and have identical parameter counts.

F.2 Controlled diagnostic ablations

Starting from the base representation

$$\mathbf{z}_{\mathrm{base}} = \left[\widetilde{X}_0,\, c_h,\, \widetilde{U}_0,\, \widetilde{V}_0\right], \tag{60}$$

we construct the Δ-only, $\mathcal{A}$-only, and w/o Ξ variants by adding $\widetilde{\Delta}$, adding $\widetilde{\mathcal{A}}$, or adding both signed diagnostics without the explicit difficulty score, respectively. Full PES adds $\widetilde{\mathcal{A}}$, $\widetilde{\Delta}$, and Ξ. All learned variants, including the matched Full model reported in this controlled comparison, use the same six-epoch budget, random seed, optimizer, routing architecture, and evaluation protocol.

Table 9: **Expert-selection baselines and controlled routing ablations for test-set Z500 forecasts at 120 h.** Panel (a) reports forecast performance; panel (b) reports routing-field structure. RMSE gain is measured relative to NeuralGCM, the stronger individual expert for this task, and Δ_{Blind} reports the RMSE-gain difference from the Blind router in percentage points.

(a) Forecast performance

Method	RMSE gain (%)	Δ_{Blind} (pp)
50/50 simple average	3.433	-0.845
Global static prior	4.282	$+0.004$
Fixed spatial prior	3.913	-0.365
Blind router	4.278	0.000
Raw forecasts (P, N)	4.089	-0.189
Δ only	4.357	$+0.079$
$\mathcal{A}$ only	4.259	-0.019
w/o Ξ	4.402	$+0.124$
Full PES	4.424	$+0.146$
Diagnostic only	**4.513**	$\mathbf{+0.235}$

(b) Routing-field diagnostics

Method	H_α	V_α	G_α
50/50 simple average	0.6931	0.0000	0.0000
Global static prior	0.6357	0.0000	0.0000
Fixed spatial prior	0.5533	**0.0160**	0.0310
Blind router	0.6100	0.0034	0.0124
Raw forecasts (P, N)	0.5671	0.0043	0.0105
Δ only	0.5556	0.0129	0.0359
$\mathcal{A}$ only	0.6105	0.0058	0.0152
w/o Ξ	0.5848	0.0056	0.0223
Full PES	0.5475	0.0132	**0.0420**
Diagnostic only	**0.4782**	0.0153	0.0331

For each method, the RMSE gain relative to NeuralGCM is

$$I_{\mathrm{RMSE}} = 100\frac{\mathrm{RMSE}_{\mathrm{N}} - \mathrm{RMSE}_{\mathrm{method}}}{\mathrm{RMSE}_{\mathrm{N}}}\%. \tag{61}$$

To isolate the contribution beyond generic learned spatial routing, we additionally report

$$\Delta_{\mathrm{Blind}} = I^{(\mathrm{method})}_{\mathrm{RMSE}} - I^{(\mathrm{Blind})}_{\mathrm{RMSE}}, \tag{62}$$

in percentage points. Blind is therefore the zero point, and positive values indicate additional improvement beyond generic learned spatial routing. Routing entropy H_α, spatial variance V_α, and local variation G_α are defined in Appendix B.

The 50/50 average establishes that uniform fusion already recovers part of the experts' complementarity, while Blind provides the direct learned-routing reference. Relative to Blind, the global static prior is essentially tied ($+0.004$ percentage points), whereas the fixed spatial and raw-forecast selectors are lower by 0.365 and 0.189 points. Full PES adds 0.146 points over Blind, and the diagnostic-only selector adds 0.235 points. The Δ-only selector also exceeds Blind, supporting expert disagreement as the primary discriminative signal. Full PES nevertheless produces the largest local routing variation. Routing-field statistics characterize the allocation patterns and are not interpreted as standalone forecast-skill metrics.

F.3 Expert Selection versus Forecast Averaging

Although PES produces the final forecast through a convex combination of two expert forecasts, its objective differs from conventional forecast averaging. The predicted preference map α is not a fixed coefficient or static ensemble weight, but a spatially and weather-state-dependent selection

rule. The fixed spatial prior experiment further demonstrates that location-dependent preference alone is insufficient when it cannot adapt to individual forecast instances. PES should therefore be interpreted as a learned expert-selection mechanism rather than a generic forecast averaging scheme.

G Compact Diagnostic Representation

Motivation. We further examine whether the information required for expert selection is better represented by the raw atmospheric state or by the explicit routing diagnostics used in PES.

Capacity-matched variants. All learned variants use the same seven-channel CNN-Transformer router. Channels excluded from a variant are replaced by zero maps, so the network architecture and parameter count remain unchanged. The state-only Blind selector receives

$$\mathbf{z}_{\mathrm{blind}} = [\widetilde{X}_0, c_h, \widetilde{U}_0, \widetilde{V}_0, 0, 0, 0], \tag{63}$$

whereas the diagnostic-only selector receives

$$\mathbf{z}_{\mathrm{diag}} = [0, c_h, 0, 0, \widetilde{\mathcal{A}}, \widetilde{\Delta}, \Xi]. \tag{64}$$

The comparison therefore contrasts raw atmospheric-state information with the compact diagnostic representation while keeping model capacity fixed.

Representative task. On the controlled Z500@120h experiment, the Blind selector obtains a 4.278% RMSE gain relative to the stronger constituent expert, whereas the diagnostic-only selector reaches 4.513%. This result indicates that the explicit diagnostic coordinates contain information that is more directly relevant to expert allocation than the raw state fields alone.

All 40 tasks. To test whether this observation generalizes beyond the representative task, we train the same capacity-matched diagnostic-only selector independently on all 40 variable-level-lead-time tasks. Table 10 reports the complete results. The diagnostic-only selector improves upon the stronger constituent expert on all $40/40$ tasks, with a mean gain of 3.454% and a median gain of 3.253%. The mean gain also increases with forecast lead time, from 2.294% at 24 h to 5.152% at 120 h. Notably, the corresponding Z500@120h gain in the 40-task extension is 4.510%, closely reproducing the controlled 4.513% result.

Interpretation. Together, the controlled Blind comparison and the complete 40-task diagnostic-only evaluation support a consistent interpretation: the direct raw-state channels are not the primary source of routing skill. Instead, most selection-relevant structure can be expressed through the compact diagnostic coordinates $\mathcal{A}$, Δ, and Ξ.

Table 10: RMSE reduction (%) of the capacity-matched diagnostic-only selector relative to the stronger constituent expert. Positive values indicate improvement.

Target	24 h	48 h	72 h	120 h
T500	1.738	3.215	4.210	5.593
T850	1.903	2.770	3.610	4.450
Q500	1.698	2.272	2.807	3.392
Q850	1.117	2.533	3.754	5.870
U500	1.328	2.865	4.090	6.200
U850	1.978	2.900	3.590	5.465
V500	1.402	3.291	4.483	6.046
V850	2.025	2.619	3.436	5.597
Z500	4.438	2.803	3.050	4.510
Z850	5.315	2.787	2.595	4.396
Mean	2.294	2.806	3.562	5.152

H Validation of Forecast Difficulty and Difficulty-Dependent Selection Benefit

The forecast-difficulty score Ξ is designed to identify regions in which local atmospheric transport and disagreement between the two frozen experts jointly indicate a more challenging forecast state. We therefore evaluate two questions. First, does a larger Ξ correspond to larger forecast errors of the constituent experts? Second, is adaptive expert selection especially beneficial in the regimes identified as difficult by Ξ?

For each variable-lead task, all valid grid points in the test subset are sorted by Ξ and divided into ten approximately equal-count bins. The bin index therefore represents relative forecast difficulty within each task rather than an absolute threshold shared across variables. This within-task construction avoids direct comparisons between variables with different physical units and numerical ranges.

We define the task-level stronger expert as the constituent model with the lower aggregate latitude-weighted RMSE on the test subset,

$$m^\star = \arg\min_{m\in\{\mathrm{P},\mathrm{N}\}} \mathrm{RMSE}_m, \tag{65}$$

where P denotes Pangu-Weather and N denotes NeuralGCM. For difficulty bin q, the normalized RMSE of model m is

$$R_{m,q} = \frac{\mathrm{RMSE}_{m,q}}{\mathrm{RMSE}_{m^\star}}, \tag{66}$$

where the denominator is the aggregate test-set RMSE of the stronger expert for the corresponding variable-lead task. This normalization preserves within-task variation across Ξ while allowing curves from different lead times to be compared on a common scale.

The composition benefit in bin q is defined as

$$G_q = 100\frac{\mathrm{RMSE}_{m^\star,q} - \mathrm{RMSE}_{\mathrm{PES},q}}{\mathrm{RMSE}_{m^\star,q}}. \tag{67}$$

A positive G_q therefore indicates that PES improves upon the stronger constituent expert within that forecast-difficulty regime.

H.1 Forecast Difficulty Identified by Ξ

To quantify the relationship between Ξ and realized forecast error, we compute

$$\rho_{\Xi,e} = \rho_s\left(\Xi, \left|\widehat{X}_{m^\star} - X^*\right|\right) \tag{68}$$

for every variable-lead task, where X^* denotes the ERA5 verification field. Verification error is used only for this post hoc diagnostic and is never supplied to the routing model. For computational tractability, the correlation is evaluated on a deterministic subsample of at most 200,000 valid test-set grid points per task.

Table 11: **Association between forecast difficulty and expert error.** Entries are Spearman correlations $\rho_s(\Xi, |e_{\text{stronger}}|)$ for all 40 variable-lead tasks. All correlations are computed on the test subset.

Target	24 h	48 h	72 h	120 h
T500	0.127	0.182	0.249	0.350
T850	0.233	0.261	0.303	0.379
Q500	0.524	0.524	0.514	0.495
Q850	0.455	0.477	0.485	0.490
U500	0.110	0.162	0.216	0.311
U850	0.143	0.182	0.236	0.326
V500	0.120	0.173	0.231	0.332
V850	0.152	0.198	0.248	0.353
Z500	0.204	0.308	0.419	0.515
Z850	0.213	0.289	0.368	0.481

Table 12: **PES benefit in difficult regimes.** Entries report the mean RMSE reduction relative to the stronger task-level expert over the upper three within-task Ξ deciles.

Target	24 h	48 h	72 h	120 h
T500	3.12	4.94	5.83	7.10
T850	3.56	4.37	5.27	6.23
Q500	1.63	2.21	2.81	3.75
Q850	1.45	2.92	4.42	7.28
U500	2.33	4.73	6.15	8.96
U850	3.39	4.66	5.29	7.35
V500	2.59	5.31	6.49	8.35
V850	3.33	4.10	4.82	7.39
Z500	4.89	3.27	3.99	5.89
Z850	7.37	3.66	3.53	6.01

Table 11 shows that the correlation is positive for all 40 variable-lead tasks, ranging from 0.110 to 0.524, with an overall mean of 0.308. Averaged across the ten atmospheric targets, the correlation increases from 0.228 at 24 h to 0.276, 0.327, and 0.403 at 48, 72, and 120 h, respectively. Thus, larger values of Ξ consistently identify regions with larger constituent-model errors, and this relationship generally becomes stronger as forecast lead time increases.

H.2 Selection Benefit in Difficult Regimes

We next examine whether the regimes identified by Ξ also contain greater opportunity for adaptive selection. Table 12 reports the mean bin-wise RMSE reduction over the three highest difficulty deciles, $q \in \{8, 9, 10\}$. PES improves upon the stronger constituent expert in the high-Ξ regime for every one of the 40 variable-lead tasks. The reductions range from 1.45% to 8.96%, with an overall mean of 4.77%.

The mean difficult-regime improvement also increases with lead time: 3.37% at 24 h, 4.02% at 48 h, 4.86% at 72 h, and 6.83% at 120 h. These results indicate that Ξ does not merely correlate with large forecast errors; it also identifies regions in which heterogeneous expert forecasts provide greater value for adaptive selection.

H.3 Representative Difficulty Structure for Z850

For the Appendix difficulty-structure visualization, we use Z850 as a representative target. Z850 is selected through a transparent composite ranking based on three quantities averaged over the four lead times: the Ξ-error correlation, the contrast between the highest and lowest Ξ deciles, and the PES gain in the upper three difficulty deciles. Let $P_v(\cdot)$ denote the percentile rank of target v among the ten evaluated targets. The selection score is

$$S_v = 0.4P_v\left(\overline{\rho_{\Xi,e}}\right) + 0.3P_v\left(\overline{R_{m^\star,10} - R_{m^\star,1}}\right) + 0.3P_v\left(\overline{G_{8:10}}\right), \tag{69}$$

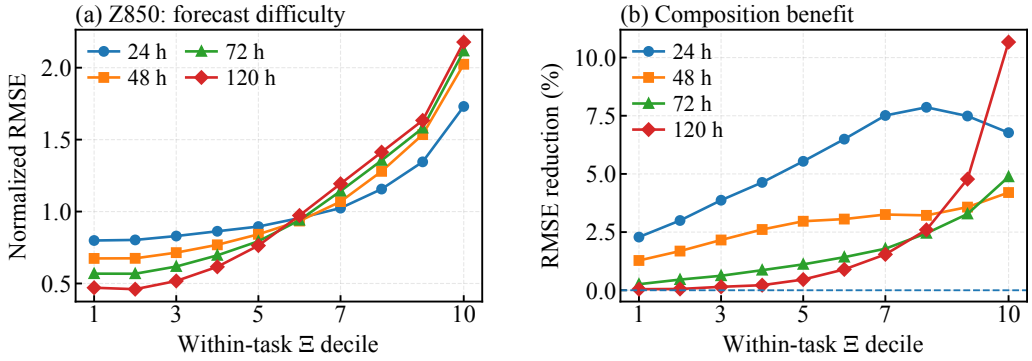

Figure 7: **Forecast difficulty and composition benefit for Z850 across four lead times.** (a) Normalized RMSE of the stronger task-level expert as a function of the within-task forecast-difficulty decile. Increasing Ξ consistently corresponds to larger constituent-expert error. (b) RMSE reduction of PES relative to the stronger task-level expert. The largest improvements are primarily concentrated in intermediate- and high-difficulty regimes, particularly at longer lead times.

where the overbar denotes averaging over the four lead times and $\overline{G_{8:10}}$ is the mean gain over deciles 8-10. The complete 40-task results in Tables 11 and 12 ensure that the conclusions do not depend on this representative example.

Figure 7 shows that the normalized RMSE of the stronger constituent expert increases substantially with Ξ at every lead time. At 120 h, for example, the normalized RMSE increases from 0.471 in the lowest difficulty decile to 2.178 in the highest difficulty decile. Similar increases occur at 24, 48, and 72 h.

The PES benefit is also concentrated toward the difficult end of the distribution. At 120 h, the RMSE reduction is only 0.05% in the lowest Ξ decile but reaches 10.67% in the highest decile. At the other lead times, the largest gains likewise occur primarily in intermediate- to high-difficulty regimes, although the detailed relationship is not strictly monotonic in every case.

Table 13 provides the complete numerical values underlying Figure 7. Although both the stronger expert and PES experience increasing RMSE as the forecast regime becomes more difficult, PES increasingly separates from the stronger expert in the upper difficulty deciles. This indicates that the gain does not arise from selecting only intrinsically easy samples. Instead, adaptive selection is most useful where the constituent forecasts are less reliable and their complementary information has greater potential value.

Taken together, the results establish a consistent two-stage relationship. First, Ξ identifies regions in which the constituent forecasts are more difficult and exhibit larger errors. Second, these high-difficulty regimes also contain greater opportunity for adaptive expert selection. The forecast-difficulty score therefore provides the router with information that is directly relevant to where heterogeneous expert forecasts are most valuable, rather than serving only as an additional unconstrained input feature.

Table 13: **Decile-level statistics for Z850.** Panels report normalized RMSE for the stronger constituent expert and PES, followed by PES's relative RMSE reduction. Splitting the quantities into aligned panels preserves the same decile and lead-time ordering without a compressed 13-column layout.

(a) Stronger-expert normalized RMSE

Ξ decile	24 h	48 h	72 h	120 h
1	0.799	0.674	0.568	0.471
2	0.803	0.675	0.568	0.460
3	0.830	0.714	0.618	0.517
4	0.863	0.769	0.696	0.616
5	0.896	0.842	0.793	0.764
6	0.955	0.934	0.940	0.973
7	1.024	1.069	1.142	1.193
8	1.156	1.279	1.355	1.413
9	1.345	1.535	1.580	1.634
10	1.730	2.024	2.116	2.178

(b) PES normalized RMSE

Ξ decile	24 h	48 h	72 h	120 h
1	0.781	0.666	0.567	0.470
2	0.779	0.664	0.565	0.460
3	0.798	0.699	0.614	0.517
4	0.823	0.749	0.690	0.615
5	0.846	0.817	0.785	0.760
6	0.893	0.906	0.926	0.964
7	0.947	1.034	1.122	1.175
8	1.065	1.238	1.322	1.376
9	1.245	1.480	1.528	1.556
10	1.613	1.939	2.013	1.946

(c) PES RMSE reduction (%)

Ξ decile	24 h	48 h	72 h	120 h
1	2.29	1.28	0.26	0.05
2	3.00	1.68	0.46	0.06
3	3.87	2.16	0.62	0.15
4	4.63	2.61	0.86	0.22
5	5.54	2.97	1.11	0.46
6	6.49	3.06	1.42	0.90
7	7.51	3.25	1.79	1.54
8	7.86	3.22	2.45	2.60
9	7.49	3.57	3.28	4.78
10	6.77	4.20	4.88	10.67

I SENSITIVITY TO THE DEFINITIONS OF ROUTING DIAGNOSTICS

The main method represents routing difficulty using three forecast-time diagnostics: signed horizontal advection $\mathcal{A}$, signed expert disagreement Δ, and their composite magnitude Ξ. These diagnostics are intended to characterize two complementary challenges: the dynamical complexity of the atmospheric state and the selection difficulty induced by heterogeneous expert responses. We conduct a targeted sensitivity analysis to examine whether the routing behavior depends strongly on their particular definitions.

Alternative diagnostic definitions. The reference configuration uses

$$\mathcal{A}_{\mathrm{adv}} = -\left(U_0 \frac{\partial X_0}{\partial x} + V_0 \frac{\partial X_0}{\partial y}\right), \qquad \Delta_{\mathrm{signed}} = \widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}, \tag{70}$$

together with

$$\Xi_{\mathrm{L2}} = \sqrt{\widetilde{\mathcal{A}}^2_{\mathrm{adv}} + \widetilde{\Delta}^2_{\mathrm{signed}} + \epsilon}. \tag{71}$$

We compare this formulation with representative alternatives from each diagnostic family. For atmospheric dynamics, the signed advection tendency is replaced by the magnitude-only transport diagnostic

$$\mathcal{A}_{\mathrm{mag}} = \sqrt{U_0^2 + V_0^2}\sqrt{\left(\frac{\partial X_0}{\partial x}\right)^2 + \left(\frac{\partial X_0}{\partial y}\right)^2}, \tag{72}$$

which retains transport intensity but removes the directional interaction between the flow and the target-field gradient. For expert disagreement, we consider

$$\Delta_{\mathrm{abs}} = \left|\widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}\right|, \qquad \Delta_{\mathrm{grad}} = \left\|\nabla\left(\widehat{X}_{\mathrm{P}} - \widehat{X}_{\mathrm{N}}\right)\right\|_2. \tag{73}$$

The former tests the effect of removing the sign of the expert difference, whereas the latter retains only its spatial structural variation.

We additionally construct a state-based difficulty indicator from the mean expert forecast

$$\overline{X} = \frac{1}{2}\left(\widehat{X}_{\mathrm{P}} + \widehat{X}_{\mathrm{N}}\right). \tag{74}$$

Specifically, let q_{anom}, q_{grad}, and q_{lap} denote the normalized climatological anomaly, gradient magnitude, and absolute Laplacian of $\overline{X}$, respectively. The alternative state-based indicator is

$$\Xi_{\mathrm{state}} = \sqrt{\widetilde{q}^2_{\mathrm{anom}} + \widetilde{q}^2_{\mathrm{grad}} + \widetilde{q}^2_{\mathrm{lap}}}, \tag{75}$$

where all normalization statistics are estimated exclusively from the training subset. We also include a no-Ξ control in which the difficulty channel is replaced by a zero map while the network architecture remains unchanged. Whenever $\mathcal{A}$ or Δ is replaced, Ξ_{L2} is recomputed using the corresponding alternative so that each configuration remains internally consistent. None of the evaluated diagnostics uses future observations or error-calibrated information.

Evaluation protocol. This design study is conducted on Z850 at lead times of 24, 48, 72, and 120 h. All variants use the same seven-channel router architecture, training samples, optimization settings, six-epoch budget, and random seed 20260717. The reported checkpoint is the best validation checkpoint within this fixed budget. The test subset is not accessed in this screening experiment.

We report the relative RMSE improvement over the stronger individual expert,

Table 14: **Sensitivity of Z850 forecast performance to diagnostic definitions.** Panel (a) reports relative RMSE improvements (%) over the stronger expert on the internal validation subset. Panel (b) gives the four-lead mean and its difference from the reference configuration in percentage points.

(a) Lead-wise RMSE improvement (%)

Variant	24 h	48 h	72 h	120 h
Current Full	5.274	2.683	2.300	4.417
$\mathcal{A}_{\mathrm{mag}}$	**5.368**	**2.683**	2.305	4.463
Δ_{abs}	4.270	2.316	2.201	4.170
Δ_{grad}	4.217	2.243	2.160	4.099
Ξ_{state}	5.263	2.643	**2.334**	**4.559**
w/o Ξ	5.283	2.659	2.320	4.472

(b) Four-lead summary

Variant	Mean	Δ_{Full} (pp)
Current Full	3.668	-
$\mathcal{A}_{\mathrm{mag}}$	**3.705**	$+0.037$
Δ_{abs}	3.239	-0.429
Δ_{grad}	3.180	-0.489
Ξ_{state}	3.700	$+0.032$
w/o Ξ	3.684	$+0.015$

$$G_{\mathrm{RMSE}} = 100\frac{\mathrm{RMSE}_{\mathrm{best}} - \mathrm{RMSE}_{\mathrm{router}}}{\mathrm{RMSE}_{\mathrm{best}}}, \tag{76}$$

where a larger value indicates better forecast performance.

Forecast accuracy alone does not describe how the diagnostic definitions affect spatial routing. We therefore additionally report the routing entropy $\mathcal{H}(\alpha)$, spatial variance $\mathrm{Var}_{\mathbf{s}}(\alpha)$, and mean local gradient $\|\nabla\alpha\|$. We also measure the alignment between the diagnostic and learned expert preference. The advection-skill alignment ρ_{align} is the correlation between the mean NeuralGCM weight and the signed relative expert skill $e_{\mathrm{P}}^2 - e_{\mathrm{N}}^2$ across advection bins. The difficulty correlation is the Spearman correlation between Ξ and

$$d^* = \frac{\min\left(|e_{\mathrm{P}}|, |e_{\mathrm{N}}|\right)}{\sigma_X}, \tag{77}$$

which is used only as a post-hoc evaluation quantity and is never provided to the router.

Table 15: **Routing and mechanism diagnostics averaged over the four Z850 lead times.** Panel (a) summarizes routing-field structure; panel (b) reports diagnostic alignment. Lower entropy indicates more decisive allocation, while larger spatial variance and gradient indicate richer spatial structure.

(a) Routing-field structure

Variant	$\mathcal{H}(\alpha)$	$\mathrm{Var}_{\mathbf{s}}(\alpha)$	$\|\nabla\alpha\|$
Current Full	0.477	0.0424	0.0687
$\mathcal{A}_{\mathrm{mag}}$	0.500	0.0405	0.0638
Δ_{abs}	0.589	0.0064	0.0274
Δ_{grad}	0.581	0.0059	0.0172
Ξ_{state}	0.497	0.0373	0.0628
w/o Ξ	0.480	0.0433	0.0680

(b) Diagnostic alignment

Variant	ρ_{align}	$\rho(\Xi, d^*)$
Current Full	0.426	0.250
$\mathcal{A}_{\mathrm{mag}}$	0.197	0.170
Δ_{abs}	0.285	0.223
Δ_{grad}	0.451	0.113
Ξ_{state}	0.298	0.248
w/o Ξ	0.470	-

Results and interpretation. The clearest result concerns the expert-disagreement diagnostic. Replacing the signed difference with its absolute value or spatial-gradient magnitude reduces the mean RMSE gain by 0.429 and 0.489 percentage points, respectively. Both alternatives also produce substantially higher routing entropy and much lower spatial variance and gradient magnitude. Thus,

the signed expert difference provides more than the magnitude of forecast separation: its direction contains important information for learning spatially selective expert preference.

The magnitude-only transport diagnostic yields a slightly larger RMSE gain than signed advection at all four lead times, but the mean difference is only 0.037 percentage points. Meanwhile, its average advection-skill alignment decreases from 0.426 to 0.197, and its routing field becomes moderately smoother. This comparison therefore does not establish a clear and mechanism-consistent advantage over signed advection. We retain $\mathcal{A}_{\mathrm{adv}}$ because it preserves the directional flow-gradient interaction and provides a stronger connection between atmospheric dynamics and learned expert allocation.

The more elaborate Ξ_{state} also does not consistently improve upon Ξ_{L2}: it is weaker at 24 and 48 h but stronger at 72 and 120 h, while their mean difficulty correlations are nearly identical (0.248 and 0.250, respectively). The no-Ξ control is also close to the Full configuration on this particular task. Consequently, this experiment does not imply that Ξ independently improves RMSE for every variable and lead time. Instead, Ξ_{L2} is retained as a compact, training-free summary of dynamical activity and expert disagreement, without introducing additional state descriptors or truth-dependent calibration.

Overall, no alternative produces a substantial improvement that is simultaneously consistent in forecast skill, routing discrimination, and physical interpretation. We therefore retain $\mathcal{A}_{\mathrm{adv}}$, Δ_{signed}, and Ξ_{L2} as a parsimonious and deployable representation of routing difficulty. Because this analysis is restricted to Z850, four lead times, and one random seed, it should be interpreted as a targeted design sensitivity study rather than an exhaustive demonstration of statistical optimality.

J Oracle Upper Bound of Convex Expert Selection

Motivation. PES is restricted to a point-wise convex combination of two frozen expert forecasts. This constraint prevents extrapolation outside their local convex hull, but it also places an upper bound on the improvement available to the router. We quantify this remaining opportunity using a verification-informed oracle.

Oracle weight and projection. For a grid point, initialization, target variable, and lead time, the verification-informed NeuralGCM weight is

$$\alpha_{\mathrm{oracle}} = \begin{cases} \mathrm{clip}\left(\dfrac{X^* - \widehat{X}_{\mathrm{P}}}{\widehat{X}_{\mathrm{N}} - \widehat{X}_{\mathrm{P}}}, 0, 1\right), & \widehat{X}_{\mathrm{N}} \neq \widehat{X}_{\mathrm{P}}, \\ \frac{1}{2}, & \widehat{X}_{\mathrm{N}} = \widehat{X}_{\mathrm{P}}, \end{cases} \tag{78}$$

where X^* is the ERA5 verification value. When the expert forecasts coincide, $1/2$ is used as a neutral convention; the oracle forecast is unchanged. The corresponding oracle forecast is

$$\widehat{X}_{\mathrm{oracle}} = \alpha_{\mathrm{oracle}} \widehat{X}_{\mathrm{N}} + (1 - \alpha_{\mathrm{oracle}})\, \widehat{X}_{\mathrm{P}}. \tag{79}$$

Equivalently, the implementation projects X^* onto the point-wise interval spanned by the two expert forecasts,

$$\widehat{X}_{\mathrm{oracle}} = \min\left[\max\left(X^*, \min(\widehat{X}_{\mathrm{P}}, \widehat{X}_{\mathrm{N}})\right), \max(\widehat{X}_{\mathrm{P}}, \widehat{X}_{\mathrm{N}})\right]. \tag{80}$$

This projection avoids introducing a fixed denominator regularizer whose effect would depend on the physical scale of the target variable.

Scope of the oracle. The oracle is not a deployable forecasting method because it explicitly uses verification-time ERA5 to determine the local weight. It is reported only as a diagnostic upper bound on the forecast skill available from unconstrained point-wise convex selection, not as a structural reference for the learned routing fields.

Captured fraction. For each task, let

$$\mathrm{RMSE}_{\mathrm{best}} = \min\left(\mathrm{RMSE}_{\mathrm{P}}, \mathrm{RMSE}_{\mathrm{N}}\right). \tag{81}$$

The fraction of the oracle-available RMSE improvement captured by PES is defined as

$$\eta = \frac{\mathrm{RMSE}_{\mathrm{best}} - \mathrm{RMSE}_{\mathrm{PES}}}{\mathrm{RMSE}_{\mathrm{best}} - \mathrm{RMSE}_{\mathrm{oracle}}}. \tag{82}$$

Table 16: **Oracle upper bound of convex expert selection at 120 h.** The oracle uses ERA5 verification fields to determine the locally optimal convex weight and is therefore reported only as a diagnostic upper bound. The first three numeric columns are RMSE; humidity rows use the multiplier shown beside the target. Captured fractions are calculated from unrounded evaluation values.

Target	Best expert	PES	Oracle	Captured
T500	1.4577	1.3768	1.0934	22.20%
T850	1.5911	1.5204	1.2077	18.44%
Q500 ($\times 10^{-4}$)	5.739787	5.538236	4.489051	16.11%
Q850 ($\times 10^{-3}$)	1.344675	1.262782	1.019043	25.15%
U500	4.3855	4.1105	3.2478	24.17%
U850	3.2572	3.0768	2.4454	22.22%
V500	4.5622	4.2864	3.3905	23.54%
V850	3.2908	3.1028	2.4616	22.67%
Z500	268.3292	256.2954	201.4654	18.00%
Z850	201.2571	192.4705	151.6420	17.71%

Thus, $\eta = 0$ corresponds to no improvement over the better individual expert, whereas $\eta = 1$ corresponds to matching the point-wise convex oracle in aggregate RMSE.

Interpretation. Across the ten evaluated targets at 120 h, PES captures between 16.11% and 25.15% of the RMSE reduction available within the point-wise convex hull, with a mean captured fraction of 21.02%. The remaining gap is expected: the learned router must infer expert preference from forecast-stage diagnostics, whereas the oracle has direct access to the verification field. The oracle analysis therefore quantifies both the improvement already recovered by PES and the remaining opportunity for more informative forecast-stage diagnostics or more expressive constrained routing mechanisms.

K Model Scale and Marginal Cost Scope

K.1 Marginal cost of PES

PES is applied after both constituent expert forecasts are available. Its reported cost therefore excludes the generation of Pangu-Weather and NeuralGCM forecasts and includes only

$$T_{\text{marginal}} = T_{\text{diagnostic}} + T_{\text{router}} + T_{\text{composition}}, \tag{83}$$

where the three terms denote diagnostic construction, routing-network inference, and point-wise convex selection, respectively.

Each variable-lead-specific router contains 110,337 trainable parameters. The ten variable-level targets and four forecast lead times require 40 independently trained routers, giving

$$110{,}337 \times 40 = 4{,}413{,}480 \tag{84}$$

trainable parameters for the complete evaluated suite, or approximately 4.41M in total. At inference, only the router corresponding to the requested variable-lead task is loaded and executed.

All routing experiments are conducted on one NVIDIA GeForce RTX 4090D. Training all 40 routers requires approximately 10 GPU-hours. Once both expert forecasts are available, PES adds approximately 50 ms per global variable-lead field.

K.2 Scope of the reported comparison

Table 1 compiles published three-day Z500 RMSE values rather than rerunning every model under a single evaluation protocol. For each model, the reported RMSE is divided by the corresponding IFS baseline. This removes the constant conversion between geopotential ($\mathrm{m^2\,s^{-2}}$) and geopotential height (m), but it does not remove differences in evaluation period, IFS cycle, initialization, verification source, regridding, or deterministic-versus-ensemble-mean evaluation. The resulting percentages provide literature-level context and should not be interpreted as a strict benchmark ranking.

Parameter counts refer to the complete forecasting system when multiple lead-time-specific or cascaded networks are required. In contrast, the PES entry reports 110K task-specific marginal routing parameters; the complete evaluated suite of 40 routers contains 4.41M additional routing parameters in total. The Pangu-Weather and NeuralGCM experts are frozen and their parameters are neither retrained nor duplicated by PES. Consequently, the table supports the positioning of PES as a lightweight composition layer, whereas the paper's comparative forecast claims rely on the internally aligned Pangu-Weather, NeuralGCM, IFS, and PES evaluations reported in Figure 2 and Appendix D.

K.3 Scope of the PES timing

The reported PES latency includes preparation of the seven routing channels, construction of $\mathcal{A}$, Δ, and Ξ, 64×64 tiled inference over the 121×240 global grid, overlap averaging, and application of the resulting routing field to the two expert forecasts. It excludes data download, preprocessing performed before an experiment, and generation of the frozen expert forecasts.

The intended setting is therefore one in which heterogeneous forecast products have already been generated, archived, or distributed. PES acts as a lightweight decision layer over these products; the reported values quantify its additional routing cost rather than the end-to-end cost of running both constituent forecasting systems.