LOCAL CALIBRATION IS NOT ENOUGH: COHERENT SPATIAL DEPENDENCE FOR QUERY-LEVEL UNCERTAINTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Probabilistic models for spatial fields can appear well calibrated locally while giving unreliable uncertainty for downstream spatial queries. We study what is lost when local predictive uncertainty is preserved but cross-region dependence is discarded. For linear queries, we derive the exact factorization $R_w = N_{\mathrm{eff}}(w)\kappa(w)$, separating effective spatial participation from cross-region coherence. On Western Australian sea-surface temperature, removing cross-patch sample correspondence reduces nominal-90% domain-mean coverage from 89.5% to 20.4% and underestimates its variance by a factor of 46.65. A leading coherent mode explains only 9.99% of total field variance but 99.7% of domain-mean variance; removing it changes R from 46.65 to 0.14. The effect is also present in stochastic PDE reference distributions and is reproduced by two distinct generative architectures. These results show that query-level uncertainty depends critically on coherent spatial structure, motivating scalable uncertainty representations that preserve local uncertainty alongside coherent large-scale structure.

1 INTRODUCTION

Generative models are increasingly used to represent uncertainty in high-dimensional spatial systems, from global weather forecasting and regional downscaling to stochastic surrogate models for partial differential equations (Li et al., 2024; Andrae et al., 2025; Price et al., 2025; Mardani et al., 2025; Park et al., 2026; Kovachki et al., 2023). Their value is not simply that they produce multiple plausible fields. A predictive distribution can be propagated to quantities used downstream, including regional averages, spatial contrasts, threshold-exceedance areas, maxima, and integrated physical quantities. For spatial aggregates and, more generally, multivariate functionals, uncertainty depends not only on the distribution at each location but also on how fluctuations across the field occur together. Here we focus on linear spatial functionals, for which this dependence can be characterized exactly.

This distinction is well recognized in probabilistic forecasting. Proper scoring rules provide a principled basis for assessing predictive distributions (Gneiting & Raftery, 2007), while multivariate scores and spatial diagnostics were developed precisely because univariate verification does not characterize dependence (Scheuerer & Hamill, 2015; Jacobson et al., 2020; Allen et al., 2024). Aggregation and transformation can further target application-specific properties of multivariate forecasts (Pic et al., 2025), while the Schaake shuffle and ensemble copula coupling explicitly reconstruct spatial or temporal dependence after marginal post-processing (Clark et al., 2004; Schefzik et al., 2013). Classical change-of-support and design-effect theory likewise link aggregate variance to spatial dependence and support (Gotway & Young, 2002; Chilès & Delfiner, 2012; Kish, 1965). Here we hold the query and block marginals fixed and isolate the distortion from deleting cross-block posterior dependence.

Modern generative forecasting systems confront the same issue: GenCast, for example, evaluates spatially aggregated quantities, cyclone tracks, and regional wind power because these depend on the joint forecast distribution (Price et al., 2025). Probabilistic neural operators also target uncertainty over function-valued outputs (Bülte et al., 2025). Our question is narrower and mechanistic: when

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.

cross-region dependence is approximated, which query-aligned structures determine the resulting downstream uncertainty error?

The less understood problem is what information is lost when that dependence is represented only approximately. For scalability, parallelism, storage, or domain decomposition, uncertainty may instead be generated or retained locally over tiles or spatial blocks, preserving local distributions while representing cross-region dependence separately or not at all. Local samples or marginal summaries may also be retained without global realization identity. These representations need not destroy dependence if an explicit coupling mechanism restores it, but local uncertainty alone does not encode which fluctuations occur together across the domain. A spatial ensemble is therefore more than a collection of local distributions: correspondence between its members across space is itself part of the predictive information.

Our block-independent construction isolates this information. We use block independence as a controlled dependence intervention rather than as a proposed inference procedure. We preserve the empirical distribution within every block and remove only cross-block sample correspondence. It is therefore the sample-based analogue of replacing the joint distribution by the product of its block marginals. In a structured patchwise generative model, this corresponds to the distinction between drawing one globally paired latent realization whose patch coefficients remain associated across the domain and drawing each patch independently from its learned marginal coefficient distribution. The intervention asks a controlled question: what downstream uncertainty is lost when local uncertainty is retained but global coupling is not?

For a linear spatial query $F_w(u) = w^\top u$, its variance depends on the joint covariance through $\mathrm{Var}[F_w(u)] = w^\top \Sigma_u w$. Two representations can therefore agree on every marginal variance while assigning substantially different uncertainty to F_w. More importantly, the magnitude and even the direction of that disagreement depend on the geometry of w. Existing diagnostics can reveal misspecified dependence, and post-processing can attempt to reconstruct it, but they do not by themselves explain why one downstream query is severely affected while another, involving much of the same domain, is not. This is the problem we study.

We first examine a structured coefficient-space generative model for sea-surface temperature over Western Australia. For the domain mean, the retained joint ensemble gives 89.5% coverage for a nominal 90% interval, whereas the block-independent representation gives only 20.4% coverage and underestimates functional variance by a factor of $R = 46.65$. Yet spatial extent alone does not explain the error. Broad positive averages and broad sign-changing contrasts involve many of the same regions but exhibit very different amplification. We show that the full-to-block variance ratio admits the exact factorization $R_w = N_{\mathrm{eff}}(w)\kappa(w)$, where N_{eff} measures the effective participation of uncertain block contributions and κ measures their coherence in the joint ensemble. Participation bounds the potential amplification; coherence determines how much is realized.

For SST, the mechanism is strikingly low dimensional. The leading coherent mode explains only 9.99% of total field variance, yet accounts for 99.7% of the variance of the domain mean. Removing this mode changes R from 46.65 to 0.14, reversing the direction of the block-independence error. Different coherent modes dominate north–south and east–west contrasts. These interventions expose the mechanism directly: query geometry selects coherent modes, which determine coherence and hence the resulting functional uncertainty error. Field-space variance explained is therefore a poor proxy for functional importance.

Simple corrections do not remove this query dependence. Post-hoc rescaling transfers poorly across queries and years, while local Matérn covariance misses persistent long-range correlation. Adding a common component helps broad positive aggregates but not sign-changing contrasts, motivating local dependence augmented by coherent large-scale components when retaining the full joint ensemble is impractical (Cressie & Johannesson, 2008; Banerjee et al., 2008).

Finally, the effect is not specific to SST or PatchPODiff. It appears directly in a stochastic advection-diffusion reference distribution and is reproduced by PatchPODiff. In stochastic Darcy flow, Diffusion Last Layer (Park et al., 2026) similarly tracks the query-dependent amplification of an empirical PDE reference. The magnitude varies across systems, but the phenomenon is not architecture specific.

Our contributions are: (i) we formalize block independence as a controlled intervention that preserves local uncertainty while removing cross-region sample correspondence, and derive the exact decomposition $R_w = N_{\text{eff}}(w)\kappa(w)$ with $R_w \leq N_{\text{eff}}(w)$; (ii) we show across designed and random spatial queries that participation alone does not determine functional uncertainty error, because query geometry controls coherence; (iii) we identify and intervene on the coherent modes underlying this behavior, showing that modes explaining modest field-space variance can dominate particular downstream quantities; and (iv) we demonstrate the effect directly in stochastic PDE reference distributions and across two generative architectures, while calibration and covariance baselines motivate local-plus-low-rank coherent dependence as a scalable alternative to block independence.

2 From Joint Fields to Query-Level Uncertainty

2.1 Predictive distribution and downstream queries

Let $u \in \mathbb{R}^d$ denote a spatial field with predictive samples $\{u^{(m)}\}_{m=1}^M$. We consider downstream linear queries $F_w(u) = w^\top u$, where $w \in \mathbb{R}^d$ specifies how locations contribute to the quantity of interest. Positive weights represent regional or domain averages, sign-changing weights define spatial contrasts, and compact weights describe local averages. Thus, w is not merely a readout vector. It defines the geometry through which the predictive distribution is observed.

To isolate the role of cross-region dependence, we partition the field into P non-overlapping spatial blocks. The joint ensemble retains the original correspondence between blocks within each predictive sample. The block-independent ensemble preserves the empirical distribution within every block while removing cross-block sample correspondence. This construction is not proposed as an inference procedure. It is the sample-based analogue of replacing the joint distribution by the product of its block marginals, and isolates the information lost whenever local uncertainty is retained without an explicit mechanism for global coupling. The intervention therefore changes dependence only; the predictive model and all blockwise samples remain unchanged.

We quantify its effect for query w by $R_w = V_{\text{full}}(w)/V_{\text{block}}(w)$. Here, $R_w > 1$ indicates that block independence underestimates query variance, $R_w < 1$ indicates overestimation, and $R_w = 1$ indicates no change.

2.2 Sample-space and structured latent evaluation

The dependence analysis is model independent. Let $U_c = [U_{c,1}, \ldots, U_{c,P}]$ contain centered predictive samples, partitioned consistently with w, and define $v_p = U_{c,p} w_p$. For the product of the empirical block marginals,

$$V_{\text{full}}(w) = \frac{1}{M-1} \left\| \sum_{p=1}^{P} v_p \right\|_2^2, \qquad V_{\text{block}}(w) = \frac{1}{M-1} \sum_{p=1}^{P} \|v_p\|_2^2. \tag{1}$$

Thus, block independence preserves each block's contribution magnitude but removes whether contributions reinforce or cancel across samples.

For an affine structured model $u = b + Da$, the same quantities can be evaluated directly in latent space. With $q = D^\top w$ and centered latent samples A_c,

$$F_w(u) = w^\top b + q^\top a, \qquad V_{\text{full}}(w) = \frac{\|A_c q\|_2^2}{M-1},$$

with $v_p = A_{c,p} q_p$ under the corresponding latent partition. This avoids dense covariance construction and full-resolution decoding; repeated-query timing is reported in Appendix J.

2.3 Participation and coherence

Equation 1 separates two properties of the query. Defining $s_p = \|v_p\|_2/\sqrt{M-1}$, we introduce effective participation and coherence as

$$N_{\text{eff}}(w) = \frac{\left(\sum_p s_p\right)^2}{\sum_p s_p^2}, \qquad \kappa(w) = \frac{\left\|\sum_p v_p\right\|_2^2}{\left(\sum_p \|v_p\|_2\right)^2}. \tag{2}$$

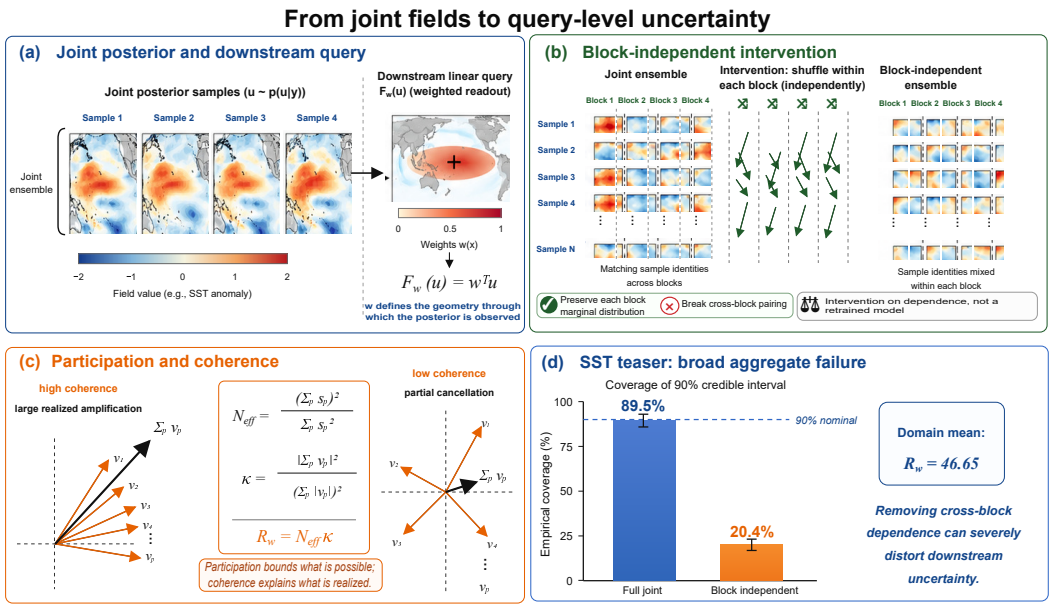

Figure 1: **From joint fields to query-level uncertainty.** **(a)** Joint posterior and spatial query F_w. **(b)** Block independence preserves local uncertainty while removing cross-block correspondence. **(c)** Alignment or cancellation of block contributions motivates N_{eff} and κ. **(d)** SST domain-mean example: $R_w = 46.65$ and nominal-90% coverage drops from 89.5% to 20.4%.

N_{eff} measures how broadly uncertainty is distributed across the spatial partition, while κ measures whether those contributions reinforce or cancel in the retained joint ensemble.

Combining these quantities gives the exact factorization

$$\boxed{R_w = N_{\text{eff}}(w)\kappa(w)}, \qquad 0 \le \kappa(w) \le 1, \qquad R_w \le N_{\text{eff}}(w). \tag{3}$$

Thus, R_w can be viewed as a query-specific design-effect analogue, with the query fixed and only cross-block posterior dependence removed. Participation therefore bounds the potential amplification caused by removing cross-block dependence, but does not determine the amplification that is realized. The latter depends on coherence, which requires the joint ensemble.

This distinction makes the role of query geometry explicit. For a regional mean, same-sign weights allow coherent large-scale fluctuations to reinforce across the queried region. For a spatial contrast, sign-changing weights can make the same fluctuation cancel after projection onto the query. Two queries may therefore have similar N_{eff} but very different κ, and consequently very different R_w. Figure 1 summarizes the dependence intervention and its effect on query-level uncertainty. Section 4.2 tests this prediction across designed and randomly generated spatial queries.

3 Experimental Setting

We study three systems. Our primary SST experiment uses PatchPODiff (Jadhav et al., 2026) over the Western Australian coastal domain and evaluates a held-out 2011 predictive ensemble using domain, north, south, N–S, and E–W queries. A controlled stochastic advection–diffusion system provides direct conditional PDE samples against which PatchPODiff can be compared. Finally, stochastic Darcy flow provides a second reference distribution and Diffusion Last Layer (DLL) supplies an independent probabilistic architecture. Dataset, architecture, training, partition, sampling, and query details are provided in Appendix A.

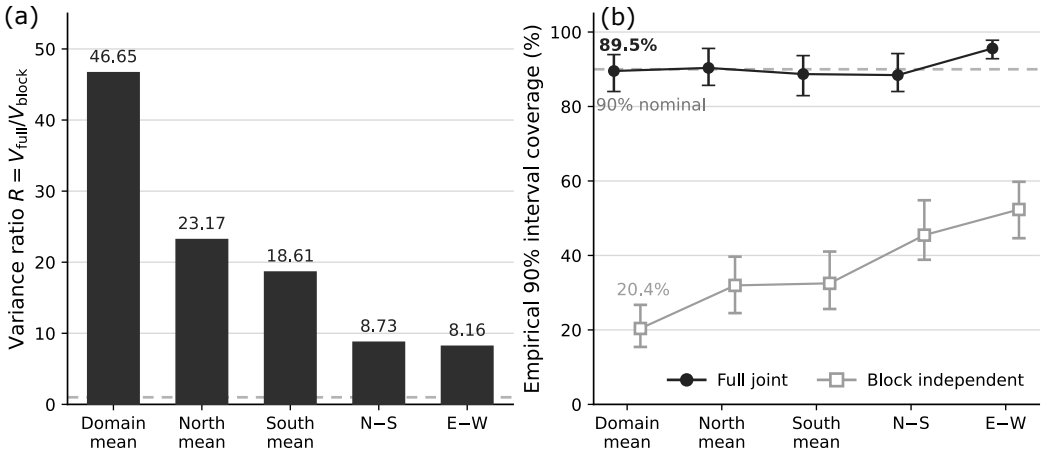

Figure 2: **Effect of removing cross-block dependence on canonical SST queries. (a)** Functional variance ratio $R = V_{\text{full}}/V_{\text{block}}$; the dashed line marks $R = 1$. **(b)** Empirical coverage of nominal 90% intervals. Error bars show 95% confidence intervals from a 14-day moving-block bootstrap with 5000 replicates.

4 Query Geometry Controls Functional Uncertainty Error

4.1 Local uncertainty does not determine uncertainty of spatial aggregates

The SST predictive ensemble is reasonably well behaved marginally: PatchPODiff attains an RMSE of 0.414, a CRPS of 0.223, and 86.8% coverage for nominal 90% intervals, increasing to 87.7% after marginal calibration. Additional marginal metrics are reported in Appendix C.

This local calibration does not carry over to spatial aggregates. For the domain mean, the retained joint ensemble gives 89.5% coverage, while removing cross-block dependence reduces it to 20.4% and yields $R = 46.65$ (Figure 2). The effect remains large but varies across queries: R ranges from 18.61–23.17 for the north and south means and from 8.16–8.73 for the two spatial contrasts. The joint ensemble remains near nominal across all five queries, whereas block independence substantially undercovers. The moving-block bootstrap intervals in Figure 2 and Table 7 show that this separation is robust to temporal dependence. A 14-day moving-block bootstrap gives non-overlapping 95% intervals for joint and block-independent coverage, including 84.02–93.94% versus 15.43–26.73% for the domain mean. The effect also persists under season-matched evaluation and ensemble-size reduction, with domain $R = 44.41$ already at $M = 25$ versus 46.65 at $M = 100$; full robustness results are in Appendix E.

4.2 Participation is not enough

The decomposition in Section 2.3 explains why broad queries need not fail equally. For the domain mean, the mean daily values are $N_{\text{eff}} = 266.98$, $\kappa = 0.174$, and $R = 46.65$. By comparison, the north-minus-south and east-minus-west contrasts still have broad participation, with mean $N_{\text{eff}} = 186.44$ and 187.04, but much lower coherence, $\kappa = 0.047$ and 0.043, with mean $R = 8.73$ and 8.16. Thus, spatial participation sets the opportunity for amplification, but coherence determines how strongly it is realized.

A support-size sweep shows the complementary trend. Positive averages over 3×3 patches give $R \approx 4.0$–4.9, increasing to $R \approx 7.5$–9.5 for 5×5 supports and $R = 46.65$ for the full-domain mean. Broader support increases the potential effect of discarded dependence, but does not determine its magnitude.

4.3 Random queries confirm the geometry effect

We next test 500 smooth random spatial queries, split equally between non-negative aggregates and zero-mean contrasts. Both families participate broadly, with median $N_{\text{eff}} = 218.8$ and 187.8, respectively, but their coherence differs sharply: median $\kappa = 0.180$ for positive queries and 0.0395 for contrasts. Consequently, median R falls from 38.81 to 7.44. The coverage results show the same separation. Joint-ensemble coverage remains close to nominal for both families, at 88.9% and 91.9%, while block-independent coverage falls to 22.2% for positive queries and 55.2% for contrasts. Across correlation length scales, contrast queries increase from roughly $R = 4.0$ to 8.6, yet remain far less coherent than positive aggregates. Query sign structure therefore changes coherence much more strongly than participation.

5 Coherent Predictive Covariance Modes Explain the Error

5.1 A mode can be small globally and dominant for one query

To identify the structures behind the query-dependent amplification, we compute eigenmodes $g_j \in \mathbb{R}^{8700}$ of the empirical latent covariance from the 2010 predictive ensemble, freeze them, and evaluate them on 2011. Figure 3a–c visualizes the corresponding decoded field patterns Dg_j. Their importance depends strongly on the query. Mode 1 explains only 9.99% of total field variance, but 99.70% of the variance of the domain mean. Its contribution is also large for the north and south means, at 78.36% and 69.53%, but falls to only 1.34% for north-minus-south and 0.97% for east-minus-west. Field-space variance explained is therefore a poor proxy for query importance.

5.2 Different query geometries select different modes

The spatial structure of the leading modes explains this difference. Mode 1 is coherent and predominantly same-sign across the domain, making it naturally aligned with positive spatial averages. Modes 2 and 3 contain large-scale sign changes and contribute much more strongly to spatial contrasts. Together, modes 2 and 3 explain 55.1% and 33.5% of the north-minus-south variance, and 53.5% and 31.6% of the east-minus-west variance. Different queries therefore extract different coherent structures from the same predictive ensemble. This gives the mechanistic chain

$$\text{query geometry} \rightarrow \text{mode alignment} \rightarrow \kappa \rightarrow R. \tag{4}$$

The participation of a query determines how broadly it draws uncertainty from the domain, while its alignment with these coherent modes determines whether those contributions reinforce or cancel.

5.3 Mode removal reveals the mechanism

We test this interpretation directly by removing the frozen modes before recomputing R on 2011. For the domain mean, removing Mode 1 changes R from 46.65 to 0.14 (Figure 3e). Block independence initially underestimates the domain-mean variance by a factor of 46.65, but after the dominant coherent mode is removed it overestimates the remaining variance by approximately $1/0.14 \approx 7.1$. The omitted cross-region contribution is therefore neither universally positive nor recoverable by a one-sided variance inflation. The contrast queries respond differently. Removing Mode 1 increases north-minus-south from $R = 8.73$ to 10.74 and east-minus-west from 8.16 to 9.73, confirming that this mode is not their main source of amplification. After removing Modes 1–3, the corresponding ratios fall to 1.08 and 1.47; the south and north means reach 1.13 and 1.89, respectively. The ratios do not all return exactly to one, so residual dependence remains. Nevertheless, most of the large query-dependent amplification is associated with only a few coherent modes. Exact mode contributions and mode-removal ratios are reported in Appendix B.5.

6 Recovery and generality of query-level uncertainty

The preceding analysis identifies coherent cross-region structure as the main source of query-level error. We next compare post-hoc and dependence-aware corrections, then test the phenomenon in independent stochastic PDE settings.

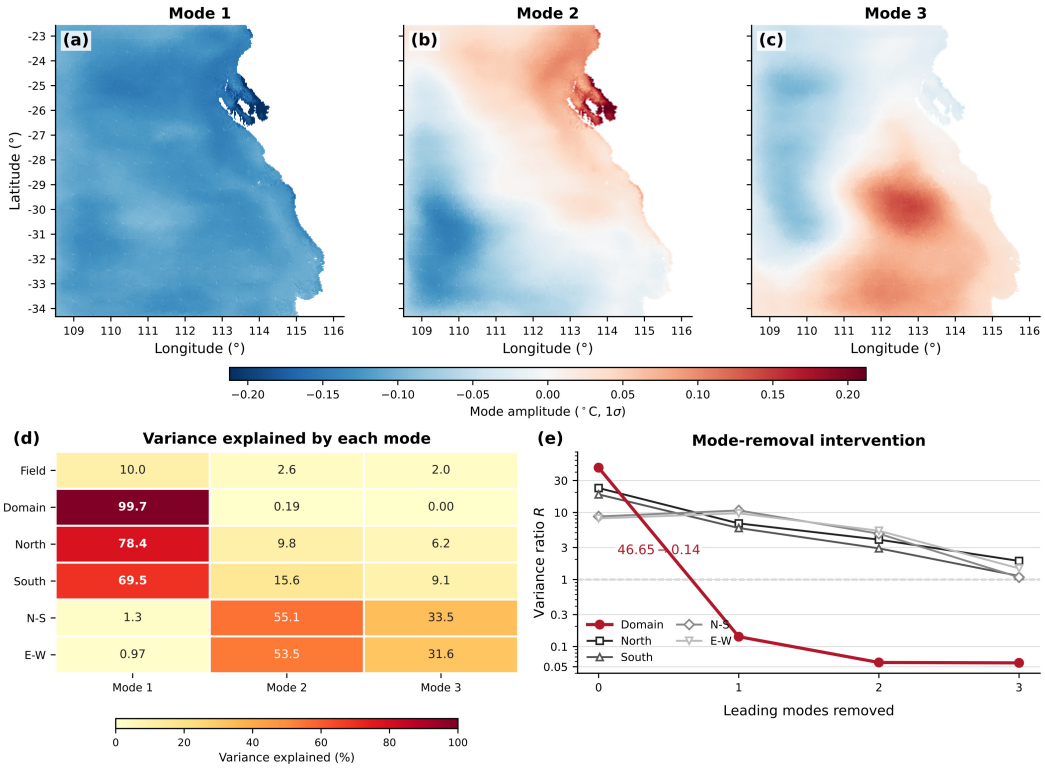

Figure 3: **Coherent modes explain query-dependent amplification.** **(a–c)** leading predictive covariance modes learned on 2010. **(d)** Their field- and query-level variance contributions. **(e)** Variance ratio after successive mode removal on 2011.

Table 1: **Nominal-90% functional coverage on 2011.** Scaling and conformal corrections are fitted on 2010; ECC uses a 2010 dependence template.

Query	Block	Global	Q-spec.	Conf.	ECC	Joint
Domain	20.4	55.4	66.7	73.6	73.0	89.5
North	32.0	75.5	79.9	85.4	81.0	90.4
South	32.5	77.4	74.9	78.8	70.0	88.7
N-S	45.5	90.9	82.1	82.9	68.0	88.4
E-W	52.3	97.8	90.4	88.2	77.0	95.6

6.1 Post-hoc corrections remain query dependent

A natural alternative is to retain independent blocks and recalibrate the resulting functional uncertainty. We fit a global scale and separate query-specific scales on 2010, then evaluate them without refitting on 2011. Table 1 shows that both improve coverage, but neither recovers the joint predictive distribution. For the domain mean, coverage rises from 20.4% under block independence to 55.4% with a global scale and 66.7% with a query-specific scale, compared with 89.5% for the retained joint ensemble.

The fitted scales also transfer poorly across queries. The east-minus-west scale gives 90.4% coverage for that contrast but only 41.6% for the domain mean, while the domain-mean scale raises east-minus-west coverage to 99.7%. The domain scale also drifts from 4.44 in 2010 to an oracle value of approximately 7.25 in 2011.

Functional conformal calibration improves coverage relative to block independence but remains query specific. A 2010-template ensemble copula coupling (ECC) baseline restores sample dependence and improves coverage, but remains at 68.0%–81.0% across the five queries. Its mean absolute coverage gap from the retained joint ensemble is 16.7 percentage points, versus 2.3 points

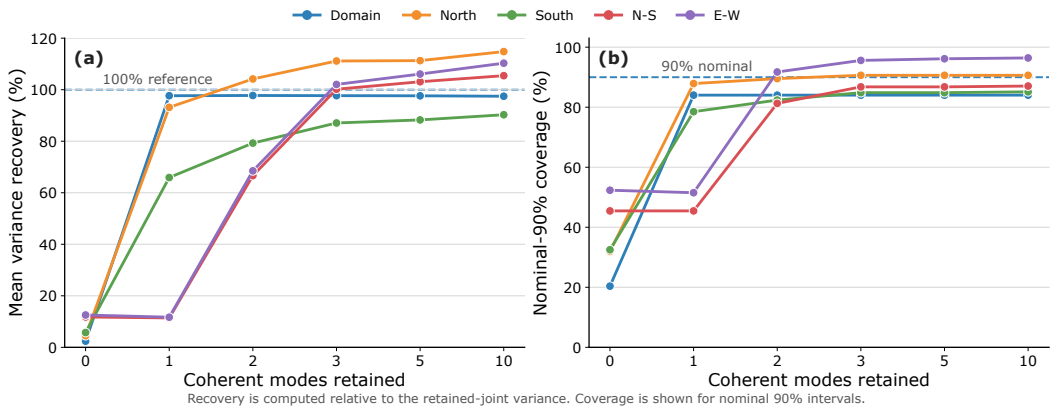

Figure 4: **A few coherent modes recover query-level uncertainty. (a)** Mean functional variance recovery after restoring the cross-block covariance carried by the first r frozen modes. **(b)** Corresponding nominal-90% coverage on 2011. A single mode restores the domain mean, while three modes recover the five query variances to 87.1%–111.1% of the retained-joint values.

for the rank-3 coherent correction in Section 6.2. Thus, dependence reconstruction helps, but a fixed historical template does not fully recover query-level uncertainty.

6.2 A FEW COHERENT MODES RECOVER QUERY-LEVEL UNCERTAINTY

We next test whether the missing dependence can be recovered with a small number of coherent modes estimated on 2010. Starting from the block-independent variance, we add only the cross-block contribution of the first r frozen modes,

$$V_r(w) = V_{\text{block}}(w) + \sum_{j=1}^{r} \lambda_j \left[(g_j^\top q_w)^2 - \sum_{p=1}^{P} (g_{j,p}^\top q_{w,p})^2 \right], \tag{5}$$

where $q_w = D^\top w$. The subtraction avoids double-counting within-block variance already contained in V_{block}. For coverage, V_{block} is replaced by V_r in the same day-wise interval construction used in Appendix A.1. Recovery is strongly query dependent (Figure 4). One mode raises domain-mean variance recovery from 2.4% to 97.7%, but leaves north–south and east–west at only 11.4% and 11.7%. With three modes, recovery reaches 97.7%, 111.1%, 87.1%, 100.2%, and 102.0% for the domain, north, south, north–south, and east–west queries, respectively. Higher ranks provide little additional benefit. Nominal-90% coverage at rank three is 84.0%, 90.6%, 84.8%, 86.8%, and 95.6%, close to the retained-joint values of 89.5%, 90.4%, 88.7%, 88.4%, and 95.6%.

The rank dependence matches the mode geometry in Section 5: Mode 1 restores broad positive aggregates, while Modes 2 and 3 supply the dominant covariance for spatial contrasts. With three modes, the compact representation requires approximately 0.626 MB in float32, compared with 3.48 MB for the retained 100-member latent ensemble, a 5.6× reduction. Stationary Matérn baselines are reported in Appendix G.

6.3 THE EFFECT IS ALREADY PRESENT IN A STOCHASTIC PDE REFERENCE

For controlled stochastic advection–diffusion, repeated numerical solves allow R to be measured directly from the reference conditional distribution. At $\mathrm{Pe} = 80$, the reference exhibits substantial query-dependent amplification, ranging from $R = 4.27$ for the central-box average to 10.58 for east-minus-west. PatchPODiff closely reproduces the same structure, with corresponding ratios ranging from 4.11 to 10.94 (Figure 5a). An independent-patch control instead gives mean $R \approx 0.992$, confirming that the much larger reference values are not explained by finite-ensemble variation. Full results and controls are reported in Appendix H.

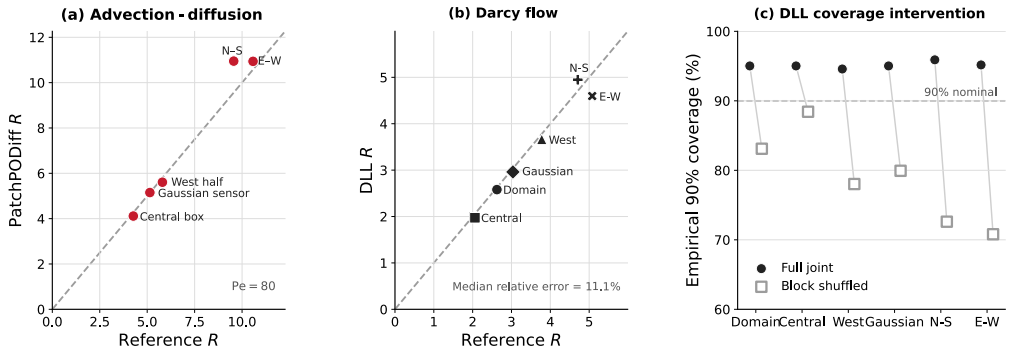

Figure 5: **Generality across stochastic PDEs and architectures.** **(a)** Advection–diffusion reference versus PatchPODiff R at $\mathrm{Pe} = 80$. **(b)** Darcy reference versus DLL R. **(c)** DLL nominal-90% coverage before and after removing cross-block correspondence.

6.4 A DIFFERENT PDE AND ARCHITECTURE RECOVER THE SAME STRUCTURE

We next change both the stochastic system and the probabilistic architecture. For stochastic Darcy flow, the empirical PDE reference gives query-dependent ratios ranging from $R = 2.06$ to 5.08, and the independently developed Diffusion Last Layer (DLL) model closely reproduces their ordering and magnitude (Figure 5b), with 11.1% median relative error across test cases and queries. The smaller amplification than in SST and advection–diffusion further indicates that R is a property of the stochastic system and query geometry, rather than an architecture-specific constant. DLL is mildly overdispersed, which partially cushions the effect on coverage. Nevertheless, removing cross-block correspondence reduces mean nominal-90% coverage from 95.1% to 78.8% (Figure 5c), including $95.9\% \to 72.6\%$ for north-minus-south and $95.2\% \to 70.8\%$ for east-minus-west. The stronger generality result is therefore the agreement between the reference and learned R structure. Full numerical results and DLL diagnostics are reported in Appendix I. Together, the two PDE experiments show that query-dependent amplification is present in the reference stochastic distributions and recovered by distinct generative architectures.

6.5 BEYOND LINEAR FUNCTIONALS

Although Eq. (3) is specific to linear queries, the practical effect of discarded dependence also appears for nonlinear quantities. For the fraction of valid ocean area exceeding a fixed SST-anomaly threshold estimated from 2010, nominal-90% coverage on held-out 2011 data drops from 87.1% for the joint ensemble to 58.4% after independently permuting ensemble identities across patches. The mean interval width decreases from 0.090 to 0.056, while CRPS increases from 0.0213 to 0.0341 (60.1%); the mean prediction remains unchanged at 0.184. Thus, the dependence failure is not restricted to linear spatial functionals.

7 DISCUSSION AND CONCLUSION

Local calibration does not determine uncertainty for spatial functionals because aggregation exposes cross-region dependence. The factorization $R_w = N_{\mathrm{eff}}\kappa$ separates spatial participation from coherence, while mode interventions identify the structures driving the largest errors. Reference-PDE experiments show that this behavior is present in the underlying stochastic distributions and reproduced across distinct generative architectures. When full joint ensembles are impractical, preserving coherent large-scale structure is therefore critical for reliable query-level uncertainty.

Limitations. Our exact decomposition is restricted to linear queries, although the nonlinear threshold experiment shows that dependence loss extends beyond them. Block independence is a controlled intervention, the low-dimensional SST mechanism need not be universal, and the PDE references are finite ensembles. Developing general nonlinear diagnostics and scalable representations of query-relevant coherent dependence remains future work.

Reproducibility Statement

We provide detailed experimental settings in the appendix, including dataset construction and splits, model architecture and training settings, spatial partitions, sampling procedures, query definitions, calibration and evaluation protocols, and additional robustness analyses. The stochastic PDE experiments likewise include the numerical setup, dataset sizes, model configurations, and evaluation procedures required to reproduce the reported results. We also report the relevant random seeds, ensemble sizes, and implementation details where applicable.

AI Use Statement

Generative AI tools were used to assist with language polishing, improving clarity and readability, and checking manuscript formatting and presentation. The authors remained responsible for the scientific content, methodology, experiments, analyses, interpretation of results, and final manuscript, and reviewed and verified all AI-assisted edits.

References

Sam Allen, Johanna Ziegel, and David Ginsbourger. Assessing the calibration of multivariate probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 150(760):1315–1335, 2024. doi: 10.1002/qj.4647.

Martin Andrae, Tomas Landelius, Joel Oskarsson, and Fredrik Lindsten. Continuous ensemble weather forecasting with diffusion models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ePEZvQNFDW.

Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008. doi: 10.1111/j.1467-9868.2008.00663.x.

Christopher Bülte, Philipp Scholl, and Gitta Kutyniok. Probabilistic neural operators for functional uncertainty quantification. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=gangoPXSRw.

Jean-Paul Chilès and Pierre Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, 2 edition, 2012.

Martyn P. Clark, Subhrendu Gangopadhyay, Lauren E. Hay, Balaji Rajagopalan, and Robert L. Wilby. The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, 5(1):243–262, 2004. doi: 10.1175/1525-7541(2004)005⟨0243:TSSAMF⟩2.0.CO;2.

Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008. doi: 10.1111/j.1467-9868.2007.00633.x.

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

Carol A. Gotway and Linda J. Young. Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648, 2002.

Josh Jacobson, William Kleiber, Michael Scheuerer, and Joseph Bellier. Beyond univariate calibration: Verifying spatial structure in ensembles of forecast fields. *Nonlinear Processes in Geophysics*, 27:411–427, 2020. doi: 10.5194/npg-27-411-2020.

Onkar Jadhav, Tim French, Matthew Rayson, and Nicole L. Jones. Patch-podiff-vit: Structured latent diffusion with patchwise pod for super-resolution and uncertainty quantification, 2026. URL https://arxiv.org/abs/2606.31290.

Leslie Kish. *Survey Sampling*. Wiley, 1965.

Nikola B. Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://www.jmlr.org/papers/v24/21-1524.html.

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

Morteza Mardani, Noah Brenowitz, Yair Cohen, Jaideep Pathak, Chieh-Yu Chen, Cheng-Chin Liu, Arash Vahdat, Mohammad Amin Nabian, Tao Ge, Akshay Subramaniam, Karthik Kashinath, Jan Kautz, and Mike Pritchard. Residual corrective diffusion modeling for km-scale atmospheric downscaling. *Communications Earth & Environment*, 6:124, 2025. doi: 10.1038/s43247-025-02042-5.

Sungwon Park, Anthony Zhou, Hongjoong Kim, and Amir Barati Farimani. Generative neural operators through diffusion last layer, 2026. URL https://arxiv.org/abs/2602.04139. Accepted at the International Conference on Machine Learning (ICML), 2026.

Romain Pic, Clément Dombry, Philippe Naveau, and Maxime Taillardat. Proper scoring rules for multivariate probabilistic forecasts based on aggregation and transformation. *Advances in Statistical Climatology, Meteorology and Oceanography*, 11:23–58, 2025. doi: 10.5194/ascmo-11-23-2025.

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

Roman Schefzik, Thordis L. Thorarinsdottir, and Tilmann Gneiting. Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4):616–640, 2013. doi: 10.1214/13-STS443.

Michael Scheuerer and Thomas M. Hamill. Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, 143(4):1321–1334, 2015. doi: 10.1175/MWR-D-14-00269.1.

A Experimental Details

A.1 Western Australian SST and PatchPODiff

The primary experiment uses daily sea-surface-temperature anomaly fields over the Western Australian coastal domain on a 640×480 ROMS grid. Land points are masked throughout the analysis and all spatial queries are evaluated only over valid ocean pixels. The anomaly climatology excludes both 2010 and 2011, so neither the calibration nor test targets contribute to the climatological reference.

The structured generative model represents the high-resolution field using non-overlapping 32×32 spatial patches arranged on a 20×15 grid, giving $P = 300$ patches. Each patch is represented by $K = 29$ standardized POD coefficients, giving a joint latent dimension of $PK = 8700$. The fitted POD artifact retains 99.5% of patchwise training energy. Before operational clipping or other nonlinear post-processing, reconstruction and spatial assembly are affine and can be written as

$$u = b + Da, \tag{6}$$

where a stacks the coefficients from all patches.

The transformer denoiser uses model width 256, six transformer blocks, eight attention heads, and a cosine DDPM schedule with 1000 diffusion steps. Functional uncertainty experiments use $M = 100$ independently seeded posterior samples with the 1000-step reverse sampler. The joint generative model samples all patch coefficients together, so cross-patch dependence can be represented even though the spatial representation itself is patchwise.

The model is trained on pre-2010 data. The available 2010 period is reserved for calibration, with 257 days spanning 17 April to 29 December, and 2011 is held out for evaluation, with 363 available days spanning 1 January to 29 December. The global-mode analysis follows a stricter separation: posterior modes are estimated using 2010 only, frozen, and then evaluated on 2011.

For SST covariance quantities, R_w, N_{eff}, and κ are computed independently for each evaluation day and then aggregated across the test period using the mean. Interval coverage is evaluated day by day.

We use five canonical linear queries. The domain query is a normalized average over all valid ocean points. The north and south queries are normalized means over their fixed regional masks. The north-minus-south query is the difference of these separately normalized regional means, with the analogous construction used for east-minus-west. Thus, the contrast queries have positive and negative weights but remain comparable in physical units to the regional averages.

For each SST evaluation day, V_{full} and V_{block} are computed exactly from the sample-space decomposition in Appendix B. Nominal-90% functional intervals are formed from the corresponding day-wise predictive means and variances, and coverage is then evaluated across days.

A.2 Controlled stochastic advection–diffusion

The controlled experiment uses a two-dimensional stochastic periodic advection–diffusion system for which the conditional solution distribution can be sampled directly. This permits the dependence intervention to be applied to the data-generating distribution itself, independently of any learned model.

High-resolution solution fields are represented on a 128×128 grid using 16×16 patches. We consider Péclet regimes

$$\mathrm{Pe} \in \{20, 80, 320\},$$

with $\mathrm{Pe} = 80$ used for the primary comparison between the reference conditional distribution and PatchPODiff. The data split contains 6000 training cases, 120 calibration cases, and 240 test cases. Patchwise POD representations retain 99.9% of training energy, and learned posterior evaluations use $M = 100$ samples.

Five linear queries are evaluated: a central-box average, west-half average, smooth Gaussian-weighted average, north-minus-south contrast, and east-minus-west contrast. The same query definitions are applied to the numerical reference samples and to the learned posterior.

To provide a dependence null, we additionally construct a reference ensemble in which patches are generated independently by construction. In this setting the expected cross-block contribution vanishes, so R should remain close to one up to finite-ensemble variability.

A.3 Stochastic Darcy flow and Diffusion Last Layer

The second independent system uses stochastic Darcy flow on a 64×64 grid together with Diffusion Last Layer (DLL), a probabilistic operator architecture distinct from PatchPODiff. The spatial output is evaluated using a 4×4 block partition.

The dataset contains 4000 training inputs, 16 validation inputs, and 32 test inputs. For each test input, the empirical conditional PDE reference contains 64 repeated conditional numerical solutions, while DLL produces 100 posterior samples. The random seed is 42. The deterministic operator encoder is trained for 60 epochs and the DLL probabilistic component is trained for a further 60 epochs.

The six downstream queries are the domain mean, central-box average, west-half average, Gaussian-weighted average, north-minus-south contrast, and east-minus-west contrast. Reference and DLL variance ratios are computed using identical query masks and block partitions. We refer to the repeated PDE solutions as an *empirical conditional PDE reference*, rather than an analytic oracle.

B Additional Theory and Derivations

B.1 Sample-space evaluation without a dense covariance

Let $A_c \in \mathbb{R}^{M\times K}$ contain centered latent samples and let $q = D^\top w$ be the latent representation of a linear spatial query. The empirical latent covariance is

$$\widehat{\Sigma}_a = \frac{A_c^\top A_c}{M-1}.$$

Therefore

$$V_{\text{full}}(w) = q^\top \widehat{\Sigma}_a q = \frac{\|A_c q\|_2^2}{M-1}. \tag{7}$$

No dense field covariance or latent covariance needs to be explicitly formed.

B.2 Block decomposition and omitted cross terms

Partition A_c and q according to the P spatial blocks and define

$$v_p = A_{c,p} q_p.$$

Then

$$V_{\text{full}} = \frac{1}{M-1}\left\|\sum_{p=1}^{P} v_p\right\|_2^2, \tag{8}$$

$$V_{\text{block}} = \frac{1}{M-1}\sum_{p=1}^{P}\|v_p\|_2^2. \tag{9}$$

Expanding the squared norm gives

$$V_{\text{full}} = V_{\text{block}} + \frac{2}{M-1}\sum_{p<r}\langle v_p, v_r\rangle. \tag{10}$$

Thus, removing cross-block dependence deletes the sample-space inner products between regional contributions. These terms may be positive or negative depending on the query geometry and joint posterior.

The variance ratio can therefore also be written as

$$R_w = 1 + \frac{2\sum_{p<r}\langle v_p, v_r\rangle}{\sum_p \|v_p\|_2^2}. \tag{11}$$

Equation 11 makes explicit why R_w can be either above or below one.

B.3 Participation–coherence factorization

Define

$$s_p = \frac{\|v_p\|_2}{\sqrt{M-1}}.$$

The effective participation and coherence are

$$N_{\text{eff}} = \frac{(\sum_p s_p)^2}{\sum_p s_p^2}, \qquad \kappa = \frac{\|\sum_p v_p\|_2^2}{(\sum_p \|v_p\|_2)^2}. \tag{12}$$

Their product is

$$N_{\text{eff}}\kappa = \frac{(\sum_p \|v_p\|_2)^2}{\sum_p \|v_p\|_2^2}\frac{\|\sum_p v_p\|_2^2}{(\sum_p \|v_p\|_2)^2} \tag{13}$$

$$= \frac{\|\sum_p v_p\|_2^2}{\sum_p \|v_p\|_2^2} = R_w. \tag{14}$$

Table 2: Fraction of field or functional variance associated with the first three frozen predictive covariance modes.

Quantity	Mode 1	Mode 2	Mode 3
Field	9.99	2.63	2.01
Domain	99.70	0.19	0.00
North	78.36	9.8	6.2
South	69.53	15.6	9.1
N-S	1.34	55.14	33.51
E-W	0.97	53.5	31.6

The triangle inequality gives

$$\left\| \sum_p v_p \right\|_2 \leq \sum_p \|v_p\|_2,$$

and therefore

$$0 \leq \kappa \leq 1, \qquad R_w \leq N_{\text{eff}}. \tag{15}$$

For a query supported entirely within one block, $N_{\text{eff}} = 1$ and the full and block-independent representations coincide, giving $R_w = 1$. This serves as an algebraic consistency check for the implementation.

Both R_w and the participation–coherence decomposition are invariant to multiplying w by a non-zero scalar. Consequently, the normalization of a query affects its physical units but not the dependence ratios studied here.

B.4 Functional importance of predictive covariance modes

Let the latent covariance admit eigenpairs (λ_j, g_j),

$$\Sigma_a = \sum_j \lambda_j g_j g_j^\top, \qquad h_j = D g_j.$$

The corresponding field covariance is

$$\Sigma_u = D \Sigma_a D^\top = \sum_j \lambda_j h_j h_j^\top.$$

The fraction of total field variance associated with mode j is

$$\frac{\lambda_j \|h_j\|_2^2}{\sum_k \lambda_k \|h_k\|_2^2},$$

whereas its contribution to the variance of query w, with $q_w = D^\top w$, is

$$\frac{\lambda_j (g_j^\top q_w)^2}{\sum_k \lambda_k (g_k^\top q_w)^2}. \tag{16}$$

These two quantities need not be similar. A mode containing a modest fraction of field variance can dominate a downstream functional when it is strongly aligned with that query.

B.5 Frozen-mode contributions and intervention

Tables 2 and 3 report the frozen-mode variance contributions and successive mode-removal ratios used in Figure 3.

Table 3: Variance ratio after removing successive leading predictive covariance modes.

Query	Original	Remove 1	Remove 1–2	Remove 1–3
Domain	46.65	0.14	0.058	0.057
North	23.17	6.88	3.94	1.89
South	18.61	5.86	2.92	1.13
N-S	8.73	10.74	4.83	1.08
E-W	8.16	9.73	5.34	1.47

Table 4: SST reconstruction and marginal uncertainty metrics on 2011. The calibration correction is fitted using 2010 and does not change the predictive mean.

Method	RMSE	MAE	CRPS	MACE	90% cov.
PatchPODiff	0.414	0.314	0.223	0.028	86.8
PatchPODiff + calibration	0.414	0.314	0.223	0.018	87.7
5-member U-Net ensemble	0.585	0.456	0.391	0.500	32.2

C Marginal SST Calibration

The functional analysis is performed on a posterior that is reasonably accurate and calibrated at individual locations. Table 4 reports reconstruction and marginal probabilistic metrics on the 2011 SST test period.

The mild marginal calibration correction improves pointwise coverage while leaving RMSE, MAE, and CRPS unchanged. The severe functional undercoverage observed after removal of cross-block dependence therefore cannot be explained by unusable local predictive uncertainty.

D Query Geometry Robustness

D.1 Support-size sweep

To isolate the effect of spatial support, we construct positive averaging queries over progressively larger collections of patches while preserving their same-sign structure. Table 5 summarizes the resulting trend.

Table 5: Functional variance amplification for positive queries with increasing spatial support.

Support	R
Single block	1
3×3 patches	approximately 4.0–4.9
5×5 patches	approximately 7.5–9.5
Full domain	46.65

Increasing support raises the opportunity for many cross-block contributions to accumulate. The support experiment alone, however, does not explain the difference between broad positive averages and broad sign-changing contrasts, which motivates the coherence analysis.

D.2 Random smooth queries

We generate 500 smooth random spatial queries from Gaussian-process weight fields over several spatial correlation scales. Half are constrained to be non-negative and therefore behave as smooth spatial aggregates. The remaining half are centered to have zero mean and therefore contain both positive and negative regions. Table 6 summarizes their dependence geometry.

The change in query sign structure has a much larger effect on coherence than on participation. This reproduces the designed-query result without relying on a small hand-selected set of functionals.

Table 6: Median participation, coherence, variance ratio, and nominal-90% coverage across 500 smooth random queries.

Family	N_{eff}	κ	R	Full cov.	Block cov.
Positive	218.8	0.1800	38.81	88.9	22.2
Zero-mean	187.8	0.0395	7.44	91.9	55.2

D.3 Length-scale sweep

The same distinction persists as the smoothness scale of the random query is varied. For zero-mean contrast queries, median R increases from approximately 4.0 at the shortest tested correlation scale to approximately 8.6 at the largest scale. Larger support therefore increases amplification, but the contrast family remains substantially less coherent than positive queries over the same range. This confirms that spatial scale and sign geometry contribute distinct information.

E Temporal and Ensemble Robustness

E.1 Moving-block bootstrap

Daily SST errors are temporally correlated, so uncertainty intervals for coverage are computed using a 14-day moving-block bootstrap with 5000 replicates. Each replicate resamples temporal blocks rather than treating individual days as independent.

Table 7: Nominal-90% coverage and 95% moving-block-bootstrap intervals for representative SST queries.

Query	Joint	Block independent
Domain	89.53 [84.02, 93.94]	20.39 [15.43, 26.73]
E-W	95.59 [92.83, 97.80]	52.34 [44.63, 59.78]

The intervals remain well separated. The north, south, and north-minus-south queries show the same qualitative behavior.

E.2 Season-matched evaluation

To test whether the result is caused by different seasonal composition between calibration and test periods, we repeat the evaluation on a season-matched 257-day subset. The result is nearly unchanged (Table 8).

Table 8: Nominal-90% coverage in the 257-day season-matched SST evaluation.

Query	Joint	Block independent
Domain	91.05	21.40
North	94.55	35.02
South	89.49	35.02
N-S	91.05	50.58
E-W	96.89	54.86

E.3 Posterior ensemble size

We recompute the SST domain-mean ratio after subsampling the posterior ensemble (Table 9).

The headline amplification is already present at $M = 25$ and changes only modestly as the ensemble size increases.

Table 9: Domain-mean amplification as a function of posterior ensemble size.

M	Domain R
25	44.41
50	45.50
75	46.15
100	46.65

F CALIBRATION BASELINES

F.1 FULL FIVE-QUERY COMPARISON

Table 10 gives the complete nominal-90% coverage comparison used to assess post-hoc calibration strategies. All calibration parameters are fitted using 2010 and frozen before evaluation on 2011.

Table 10: Nominal-90% functional coverage on 2011.

Query	Block	Global	Query-spec.	Conformal	Joint
Domain	20.4	55.4	66.7	73.6	89.5
North	32.0	75.5	79.9	85.4	90.4
South	32.5	77.4	74.9	78.8	88.7
N-S	45.5	90.9	82.1	82.9	88.4
E-W	52.3	97.8	90.4	88.2	95.6

F.2 VARIANCE INFLATION SCALES

A single standard-deviation multiplier fitted jointly across all five functionals gives

$$s_{\text{global}} = 3.257.$$

The separately fitted 2010 multipliers are listed in Table 11.

Table 11: Query-specific standard-deviation multipliers fitted on 2010.

Query	Scale
Domain	4.44
North	3.52
South	3.15
N-S	2.39
E-W	2.17

The large variation between queries is expected because each query projects a different subset and sign pattern of the omitted covariance.

F.3 CROSS-QUERY TRANSFER

To test whether a scale fitted for one query is reusable for another, each of the five query-specific scales is applied to all five evaluation functionals. Table 12 reports rounded 2011 coverage.

The two most asymmetric transfers are especially informative. The east-minus-west scale gives only 41.6% coverage for the domain mean, while the domain scale gives approximately 99.7% coverage for east-minus-west.

F.4 TEMPORAL DRIFT

Calibration also changes across time. The domain-mean scale fitted on 2010 is 4.44, whereas the diagnostic scale required to attain nominal coverage on 2011 is approximately 7.25. The correction therefore depends not only on the functional but also on the calibration distribution.

Table 12: Cross-query transfer of 2010-fitted scales. Rows identify the query used to fit the scale; columns identify the functional evaluated on 2011. Values are coverage percentages rounded to the nearest percentage point.

Scale fitted for	Domain	North	South	N-S	E-W
Domain	67	88	90	97	100
North	59	80	80	93	99
South	54	72	75	90	98
N-S	44	61	62	82	93
E-W	42	56	58	78	90

F.5 Functional conformal calibration

Functional conformal calibration is performed after projection to the scalar downstream query and separately for each functional using the 2010 calibration period. The calibrated scalar interval is then evaluated without refitting on 2011. This improves coverage relative to the raw block representation, but does not reconstruct a reusable joint spatial posterior. A new downstream query requires a new scalar calibration problem.

F.6 Ensemble copula coupling

We apply ensemble copula coupling to the block-independent 2011 predictive ensemble using a dependence template estimated exclusively from 2010. Within-block samples and their marginal distributions are unchanged; ECC only reorders ensemble members to impose the historical rank dependence. The resulting coupled ensemble is evaluated using the same five functionals and nominal-90% coverage calculation as the retained-joint and block-independent ensembles.

G Covariance Baselines

G.1 Stationary Matérn models

We test whether the discarded dependence can be approximated by a stationary local covariance model. Let h denote spatial separation and ℓ a fitted length scale. The two Matérn forms used in the principal diagnostics include

$$\rho_{1/2}(h;\ell) = \exp(-h/\ell), \tag{17}$$

$$\rho_{3/2}(h;\ell) = \left(1 + \frac{\sqrt{3}h}{\ell}\right)\exp\left(-\frac{\sqrt{3}h}{\ell}\right). \tag{18}$$

The practical residual-fit baseline estimates the correlation scale using 2010 residual structure. The best reported $\nu = 1/2$ model recovers only 9.3% of domain-mean variance and gives 42.7% nominal-90% coverage.

The limitation is visible directly in the posterior correlation. Empirical correlation does not decay to zero over the domain but plateaus near

$$\rho(h) \approx 0.087$$

for separations beyond approximately 400 km.

G.2 Matérn plus a common component

We augment the stationary correlation with a constant coherent term,

$$\rho_{\mathrm{CM}}(h) = c + (1-c)\rho_\nu(h;\ell). \tag{19}$$

The constant term represents a domain-wide component that remains correlated at long distance.

The common component almost restores the domain-mean variance, but it is not uniformly beneficial. Recovery for the two sign-changing contrasts decreases when the same global component is

Table 13: Representative domain-mean covariance-recovery diagnostics.

Fit	ν	Domain variance recovery	Coverage
Practical residual Matérn	$1/2$	9.3%	42.7%
Posterior-informed Matérn	$3/2$	13.8%	–
Posterior-informed Matérn + common	$3/2$	94.2%	–

added. Representative recovery results are reported in Table 13. This is consistent with the mode analysis: a same-sign coherent component aligns strongly with broad positive averages but not with all possible query geometries.

These results motivate the approximation

$$\Sigma_u \approx \Sigma_{\text{local}} + \sum_{j=1}^{r} \lambda_j h_j h_j^\top, \qquad h_j = Dg_j. \tag{20}$$

while leaving the optimal rank, basis, and fitting strategy as open questions.

H Controlled PDE Checks

H.1 Péclet-number sweep

The dependence intervention is evaluated directly on reference conditional distributions at

$$\text{Pe} \in \{20, 80, 320\}.$$

Across all five queries and three regimes, reference full-to-block variance ratios range from approximately 4.27 to 11.82. Averaged across the 15 query–regime combinations, nominal-90% coverage is approximately 89.8% under the full reference dependence and 46.7% under block independence.

At $\text{Pe} = 80$, the learned posterior reproduces the reference query structure closely (Table 14):

Table 14: Reference and learned functional variance ratios at $\text{Pe} = 80$.

Query	PDE reference R	PatchPODiff R
Central box	4.27	4.11
Gaussian	5.14	5.15
West half	5.80	5.61
N-S	9.56	10.95
E-W	10.58	10.94

For example, the east-minus-west reference query has $R = 10.58$, with approximately 88.9% full coverage and 38.4% block-independent coverage.

H.2 Independent-patch null

The independent-patch reference is constructed so that no cross-block dependence exists in the data-generating distribution. Across the stored finite ensembles, the mean ratio is

$$\overline{R} = 0.992,$$

with observed values spanning approximately 0.90–1.12.

This range is used only as an empirical indication of finite-ensemble variability under genuine independence. It is not interpreted as a formal confidence interval or universal noise floor.

H.3 PDE ensemble-size sensitivity

The learned PDE ratios are stable under posterior subsampling (Table 15).

The changes from $M = 25$ to $M = 100$ are small relative to the magnitude of the effect.

Table 15: PDE variance-ratio sensitivity to posterior ensemble size.

M	E-W R	N-S R
25	10.69	10.83
50	10.83	10.93
75	10.87	10.95
100	10.94	10.95

H.4 POD COMPRESSION ACROSS PÉCLET REGIMES

The POD rank required to retain 99.9% patchwise energy increases as the system becomes more advection dominated (Table 16).

Table 16: Patchwise POD rank and approximate compression across Péclet regimes.

Pe	K	Approx. compression
20	6	$42.7\times$
80	9	$28.4\times$
320	16	$16.0\times$

I DARCY AND DLL VERIFICATION

I.1 REFERENCE VERSUS DLL VARIANCE RATIOS

Table 17 reports the complete comparison between the empirical conditional PDE reference and DLL.

Table 17: Median functional variance ratios for stochastic Darcy flow.

Query	PDE reference	DLL
Domain	2.625	2.580
Central box	2.058	1.973
West half	3.779	3.665
Gaussian	3.037	2.961
N-S	4.707	4.948
E-W	5.078	4.597

Across individual test-case/query pairs, the median absolute relative error in R is 11.10%.

I.2 SPREAD AND COVERAGE DIAGNOSTICS

DLL is somewhat overdispersed relative to the empirical conditional PDE reference. The global DLL-to-reference spread ratio is

$$1.23946.$$

This overdispersion raises raw nominal-90% coverage and therefore partially cushions the coverage loss caused by removing cross-block dependence (Table 18).

The absolute coverage drop should therefore be interpreted conservatively. The stronger evidence for architecture independence is the close match between the reference and DLL query-dependent R structure.

I.3 TRAINING AND SAMPLING CONFIGURATION

The verified Darcy and DLL configuration is summarized in Table 19.

Table 18: Selected DLL nominal-90% functional coverage before and after removing cross-block dependence.

Query	Full DLL	Block independent
Mean across queries	95.12	78.82
Domain	95.10	78.80
N-S	95.90	72.61
E-W	95.17	70.80

Table 19: Verified stochastic-Darcy and DLL configuration.

Setting	Value
Spatial grid	64×64
Spatial partition	4×4 blocks
Training inputs	4000
Validation inputs	16
Test inputs	32
PDE reference draws per test input	64
DLL samples per test input	100
Random seed	42
Operator encoder training	60 epochs
DLL training	60 epochs

J Repeated Query Timing

For an affine decoder, a linear functional can be evaluated directly from the retained latent sample factor after computing the projected query $q = D^\top w$. We benchmark this latent-space evaluation against querying a cached high-resolution ensemble and against decoding all high-resolution posterior members before evaluating the query.

Each timing configuration uses 25 warm-up iterations followed by 120 timed repetitions. The $Q = 1$ configuration is excluded from headline comparisons because the timing artifact flags this regime as a small-kernel launch anomaly and the strict monotonicity sanity check does not pass at that point.

Table 20: Corrected post-hoc functional-query benchmark.

Q	Latent	Cached HR	Decode + query	Speedup
10	0.081 ms	0.335 ms	1.365 ms	$16.86\times$
100	0.083 ms	1.128 ms	2.164 ms	$26.14\times$
1000	0.154 ms	8.264 ms	9.313 ms	$60.40\times$

For $Q = 10$–1000, direct latent-space evaluation is approximately 16.9–$60.4\times$ faster than decode-and-query evaluation (Table 20). The benchmark concerns repeated post-hoc evaluation once the structured posterior has already been generated; it does not include model training or posterior sampling time.